BenchMetrics: A systematic benchmarking method for binary-classification performance metrics

Gürol Canbek[1], Tugba Taskaya Temizel[2], and Seref Sagiroglu[3]

1 Corresponding author: gcanbek@aselsan.com.tr, ASELSAN and Middle East Technical University, Ankara, Turkey, 0000-0002-9337-097X

2 Middle East Technical University, Informatics Institute, Ankara, Turkey, 0000-0001-7387-8621

3 Gazi University, Computer Engineering Department, Ankara, Turkey, 0000-0003-0805-5818

Abstract

This paper proposes a systematic benchmarking method called BenchMetrics to analyze and compare the robustness of binary-classification performance metrics based on the confusion matrix for a crisp classifier. BenchMetrics, introducing new concepts such as meta-metrics (metrics about metrics) and metric-space, has been tested on fifteen well-known metrics including Balanced Accuracy, Normalized Mutual Information, Cohen's Kappa, and Matthews Correlation Coefficient (MCC), along with two recently proposed metrics, Optimized Precision and Index of Balanced Accuracy in the literature. The method formally presents a pseudo universal metric-space where all the permutations of confusion matrix elements yielding the same sample size are calculated. It evaluates the metrics and metric-spaces in a two-staged benchmark based on our proposed eighteen new criteria and finally ranks the metrics by aggregating the criteria results. The mathematical evaluation stage analyzes metrics' equations, specific confusion matrix variations, and corresponding metric-spaces. The second stage, including seven novel meta-metrics, evaluates the robustness aspects of metric-spaces. We interpreted each benchmarking result and comparatively assessed the effectiveness of BenchMetrics with the limited comparison studies in the literature. The results of BenchMetrics have demonstrated that widely used metrics have significant robustness issues, and MCC is the most robust and recommended metric for binary-classification performance evaluation.

Keywords binary classification, performance metric, performance evaluation, benchmarking, meta-metric

Declarations

Funding: The authors did not receive support from any organization for the submitted work.
Conflicts of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.
Availability of data and material: The authors confirm that all data and materials support their published claims and comply with field standards (*see* Appendix A).
Code availability: The authors confirm that all software application or custom code support their published claims and comply with field standards (*see* Appendix A).
Authors' contributions: *Gürol Canbek*: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review and editing, Visualization. *Tugba Taskaya Temizel*: Validation, Writing - review and editing, Supervision. *Seref Sagiroglu*: Conceptualization, Validation, Writing - review & editing, Supervision.

1 Introduction

Binary-classification performance metrics are widely used instruments for numerous classification application areas to evaluate and report the performance of classifiers. The choice of a metric is often dependent on the classification problem domain and the previous practices in the related literature. In recent years, however, the limitations of the commonly used metrics such as accuracy and F1 have been noted. For instance, accuracy (ACC) was reported not to be robust in class imbalanced problems because it produces an over-optimistic prediction performance towards the majority class [1, 2]. Likewise, F1, despite its reputation in many fields such as information retrieval, has been criticized as invariant to class swapping and independent from the number of true negative samples [3, 4]. Researchers proposed metrics to improve the existing ones such as accuracy, G, and F1 (*e.g.*, Optimized Precision (OACC) to improve Accuracy [5] or IBA$_{\alpha}$(G) to improve G [6]). However, they often had difficulties demonstrating the superiority of the proposed metrics systematically over the existing ones. The evaluation of performance metrics has been carried out in the literature by (1) manual examination of known machine-learning (ML) algorithms trained and tested on balanced and imbalanced synthetic or real-world datasets and comparing the metric outputs in tabular form or on the plots in graphical form, (2) examining each metric's behavior when the confusion matrix elements are exchanged to see whether they are mathematically correct,

(3) introducing comparison criteria or (4) proposing evaluation requirements to see the differences among the metric outputs.

- In the first approach, the studies in the literature used simulated baseline classifiers on synthetic datasets [7] or real-world datasets [8, 9] with different class ratios. Then, the results expressed in terms of various metrics or their correlations were examined individually [10]. These comparisons were generally carried out by visual inspection of related graphs or manual interpretation, which are limited to show marginal differences between the metrics [11, 12]. Several factors provided as is, including the ML model and parameter selection, feature selection, and noise in the datasets, could not be taken into account in those evaluations. Such approaches covered a limited metric-space.
- In the second approach, the variations of metrics upon simply changing the confusion matrix elements are examined. Sokolova and Lapalme [13] examined invariance behaviors such as changing one element value while the remaining ones are the same and swapping TP with TN and FP with FN.
- In the third approach, the studies proposed criteria or constraints to evaluate metrics. Huang and Ling [14] suggest consistency and discriminancy degrees for comparing two performance metrics tested with only ACC and Area-Under-ROC-Curve (AUCROC, ROC: receiver operating characteristic) metrics in balanced/imbalanced dataset examples. Consistency refers to the degree of agreement between two metrics where the corresponding pair of metric output values are in the same order of precedence (*i.e.* the first values of both metrics are higher than the second values). Discriminancy indicates the degree of the cases in which a metric yields a different pair of output values where the other returns the same value.
- For the fourth approach, Forbes [15] poses six coarse constraints for measures of agreements, such as being statistically principled, readily interpretable, and generalizable to more than one class. From a classification perspective, Pereira et al. [16] address uninformed decisions about using a metric for multi-label classifications in the literature, analyze the correlations among metric pairs, and suggest alternative sets of metrics to use. Straube and Krell [17] propose the following criteria for choosing a metric: i) performance-oriented (not data-oriented), ii) intuitive (interpretable), and iii) comparable (accepted in the literature). Some essential criteria are also suggested for assessing performance metrics. For example, Hossin and Sulaiman [18] discuss factors for binary and multi-class classification performance metrics, such as multi-class compatibility, less complexity/computational cost, distinctive and discriminable metric outputs, informativeness (discrimination of equivalent or different cases), and minority class sensitivity.

However, these approaches exhibit several drawbacks: (i) they cover only limited aspects of the metrics, (ii) the results based on a small number of synthetic or real-world datasets cannot be generalized, (iii) partial coverage of the error surface, (iv) they do not produce a scaled mathematical score for facilitating objective comparison, and (v) visual examination of graphics or manual interpretation of tabular values requires expertise and provides little insights. Moreover, the evaluation scope was limited. Tharwat [19] reviews performance metrics where CK is not included. Likewise, Brown finds MCC (as the best) and F1 (conflicting with our assessments) more *realistic* comparing with only TPR, TNR, PPV, and ACC by examining the TPR *versus* TNR in balanced and imbalanced datasets [3]. Chicco and Jurman [20] claim that MCC is more *informative* than ACC and F1 by examining the classifiers yielding random and largest class binary-labels and a set of class-imbalance cases. In brief, two key problems with much of the literature on reviewing performance metrics are observed:

- the assessments are limited, as elaborated in Section 3 and specifically reviewed in Table **11**, and
- the comparisons are conducted with only a few metrics.

In this paper, a novel systematic benchmarking method for evaluating binary-classification performance metrics, which overcome the limitations of the previous studies in the literature is proposed. The method is tested on fifteen performance metrics. The results reveal that CK and MCC are undifferentiated from many perspectives (also interpreted in [21]), but clearly distinguishes one of them as the most robust metric.

In the literature, binary-classification performance metrics evaluated only limited aspects of metrics. On the other hand, our method proposes fifteen new criteria that scrutinize different aspects of metrics and three criteria that have been studied before in the literature. This study differentiates itself from the literature by producing scaled scores, which facilitate comparison and ranking. Moreover, the proposed metric space includes all possible cases, enabling us to carry out comparisons systematically.

2 Research questions and objectives

This study's main research question is “Whether we can systematically analyze and comprehensively compare the binary-classification metrics?”, and the objective of the study is to identify which instruments are robust to use in binary classification. This study is distinctive for the following contributions to the literature (refer to Appendix B for the abbreviations of the performance measures and metrics addressed in this study.[1]):

- Proposing a systematic and comprehensive benchmarking method for binary-classification performance metrics,
- Suggesting meta-metrics to assess the desired measurable capacity of any given performance metric, and
- Evaluating and comparing thirteen metrics, namely TPR, TNR, PPV, NPV, ACC, INFORM, MARK, BACC, G, nMI, F1, CK, and MCC, along with two recently proposed metrics, via the proposed benchmark.

Note that zero-one "loss" metrics are not included in benchmarking to avoid redundancy because they are just the complement of the evaluated performance metrics (*i.e.* $MCR = 1 - ACC$, $FPR = 1 - TNR$, $FNR = 1 - TPR$, FDR $= 1 - PPV$, and FOR $= 1 - NPV$). Some of our benchmark findings align with the literature for some metrics such as TPR, ACC, F1, CK, and MCC. However, numerous new defects have been identified for widely-used metrics such as BACC, G, nMI, and F1. Besides, to the best of our knowledge, BACC, nMI, CK, and MCC have not been explicitly compared in the literature. nMI is an entropy-based metric that normalizes the mutual information (MI) indicating the strength of association in the contingency table is used for binary-classification, namely *prior* ("ground truth": P or N) and *posterior* ("prediction": OP or ON) distributions. For the entropy-based instruments, which are the subtype of confusion-matrix derived instruments, the following equation is valid: $MI = HC + HO - HOC$ (*see* Appendix B) where HC: class entropy, HO: outcome entropy, and HOC: Joint Entropy [22, 23].

The rest of the paper is structured as follows. Section 3 proposes a benchmarking method to assess performance metrics. Section 4 summarizes benchmarking findings. Section 5 reviews benchmarking methods used in the literature and compares them with our results. Section 6 demonstrates another application of meta-metrics for evaluating the class-imbalance effect on synthetic classifiers. Section 7 describes BenchMetrics usage in graphical-based performance metrics. Section 8 describes the exclusion of probabilistic error/loss performance instruments. The final section summarizes the methods and highlights the findings and significance of this study. Appendix A provides complementary materials online for the proposed method, such as the BenchMetrics library, an interactive benchmarking platform, and other materials and datasets. Appendix B lists the abbreviations of the performance measures and metrics addressed in this study.

3 BenchMetrics: A proposed method for performance-metrics benchmarking

The literature has addressed performance metrics' weakness in a limited scope, such as the prevalence effect (also known as class-skew sensitivity or class imbalance problem). Most of the performance metrics are sensitive to class skewness [24]. Without any change in a classification model, its performance in terms of those metrics can increase upon changing the positive/negative class samples' distribution. Straube and Krell [17] conclude that ACC, F1, MCC, and nMI are sensitive to class skew and DPR, BACC, WACC, and G are class-insensitive based on a single example via a single hypothetical classifier having $TPR = 0.9$ and $TNR = 0.7$. Brzezinski et al. [21] conduct a manual analysis of eight metrics' outcomes for five class-ratio categories (1:15, 1:4, 1:1, 4:1, 15:1) via histogram graphics.

The literature touches upon problematic performance metrics, especially TPR, PPV, ACC, and F1. Valverde-Albacete and Peláez-Moreno [25] report that higher Accuracy values could be misleading. Shepperd [26] also indicates that F1 yields significantly high values (about 0.7) on highly skewed datasets and exhibits a misrepresenting of high performance in low prevalence datasets. Straube and Krell [17] recommend DPR, BACC, WACC, and G instead of ACC, F1, MCC, and nMI considering prevalence effect. Schröder, Thiele, and Lehner suggest using INFORM, MARK, and MCC instead of PPV, TPR, and F1 [27]. Forbes [15] recommends nMI as a nontraditional metric. Delgado and Tibau [28] examine CK and MCC in unbalanced datasets via the specific confusion matrix cases and show the pitfalls in CK. Chicco and Jurman [20] review ACC, F1, and MCC. They compare those metrics based on (a) metric values in terms of class sizes corresponding to the perfect classification/misclassification (FP and $FN = 0$ / TP and $TN = 0$) and random classification (*i.e.* expected confusion matrix elements: $TP = FP = P/2$ and $TN = FN = N/2$), (b) metric values corresponding to six class-ratio categories for

[1] Note that 'performance metrics' that are in [0, 1] or [-1, 1] directly represents the success of a classifier (*e.g.*, Accuracy or True Positive Rate). Those metrics are the instruments published in the literature to report, evaluate, and compare classifiers. Whereas, 'performance measures' that are usually not published represent other aspects such as dataset or classifier's output characteristics (*e.g.*, PREV is the ratio of positive examples in a dataset and BIAS is the ratio of positive outcomes of a classifier). Some instruments indicating the performance in an unbounded interval $[0, \infty)$ or $(-\infty, \infty)$ are also 'measures' that are not applicable to publish and compare classification performances in the literature (*e.g.*, Odds Ratio or Discriminant Power) because of limitations in interpretability.

Sn = 100 (high/middle positive/negative class imbalance and balanced class sizes), and (c) the linear relationship among the three metrics.

To evaluate all the metrics from a comprehensive perspective in a methodological manner, we proposed a benchmarking method comprising two stages described in the following subsections and depicted in Fig. 1:

- Stage-1: *Mathematical evaluation*: The equations of each metric, confusion matrix variance behaviors, and the metric-spaces are evaluated according to eleven different criteria.
- Stage-2: *Meta-metrics*: The robustness of each metric is evaluated by seven novel meta-metrics (*i.e.* metrics about (performance) metrics) defined formally in metric-space.

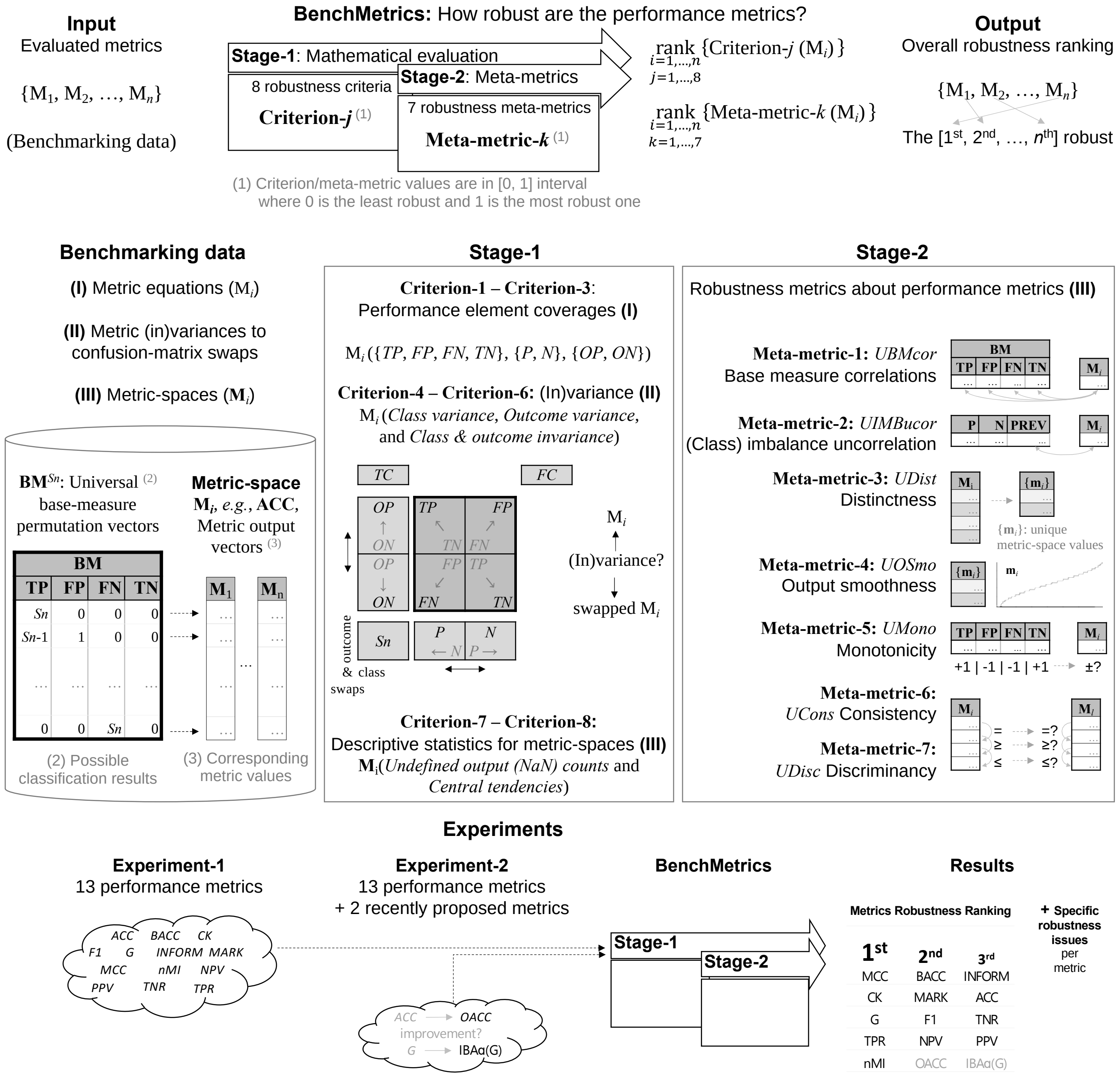

Input
Evaluated metrics
{M1, M2, …, Mn}
(Benchmarking data)
BenchMetrics: How robust are the performance metrics?
Stage-1: Mathematical evaluation
8 robustness criteria
Criterion-j (1)
Stage-2: Meta-metrics
7 robustness meta-metrics
Meta-metric-k (1)
rank {Criterion-j (Mi)} i=1,…,n j=1,…,8
rank {Meta-metric-k (Mi)} i=1,…,n k=1,…,7
Output
Overall robustness ranking
{M1, M2, …, Mn}
The [1st, 2nd, …, nth] robust
(1) Criterion/meta-metric values are in [0, 1] interval where 0 is the least robust and 1 is the most robust one
Benchmarking data
(I) Metric equations (Mi)
(II) Metric (in)variances to confusion-matrix swaps
(III) Metric-spaces (Mi)
BMSn: Universal (2) base-measure permutation vectors
Metric-space Mi, e.g., ACC, Metric output vectors (3)
BM
TP FP FN TN
Sn 0 0 0
Sn-1 1 0 0
0 0 Sn 0
M1 Mn
(2) Possible classification results
(3) Corresponding metric values
Stage-1
Criterion-1 – Criterion-3: Performance element coverages (I)
Mi ({TP, FP, FN, TN}, {P, N}, {OP, ON})
Criterion-4 – Criterion-6: (In)variance (II)
Mi (Class variance, Outcome variance, and Class & outcome invariance)
TC FC
OP ON TP FP TN FN
Sn P N
outcome & class swaps
Mi
(In)variance?
swapped Mi
Criterion-7 – Criterion-8: Descriptive statistics for metric-spaces (III)
Mi(Undefined output (NaN) counts and Central tendencies)
Stage-2
Robustness metrics about performance metrics (III)
Meta-metric-1: UBMcor Base measure correlations
Meta-metric-2: UIMBucor (Class) imbalance uncorrelation
P N PREV
Meta-metric-3: UDist Distinctness
{mi}: unique metric-space values
Meta-metric-4: UOSmo Output smoothness
Meta-metric-5: UMono Monotonicity
+1 | -1 | -1 | +1 ±?
Meta-metric-6: UCons Consistency
Meta-metric-7: UDisc Discriminancy
= =? ≥ ≥? ≤ ≤?
Experiments
Experiment-1
13 performance metrics
ACC BACC CK F1 G INFORM MARK MCC nMI NPV PPV TNR TPR
Experiment-2
13 performance metrics + 2 recently proposed metrics
ACC → OACC improvement? G → IBAα(G)
BenchMetrics
Stage-1
Stage-2
Results
Metrics Robustness Ranking
1st 2nd 3rd
MCC BACC INFORM
CK MARK ACC
G F1 TNR
TPR NPV PPV
nMI OACC IBAα(G)
+ Specific robustness issues per metric

Fig. 1 BenchMetrics: inputs, stages, outputs, evaluation criteria/meta-metrics for metrics and metric-spaces. The method was tested for the benchmarking data for 13 metrics (Experiment-1) and 15 metrics with two recently proposed metrics (Experiment-2). The evaluated metrics are ranked according to overall robustness values. The experiments also provide specific robustness issues indicated by low values in criteria or meta-metrics, all of which are described in this study.

3.1 Benchmarking data

This subsection introduces a new aspect of metrics named "metric-space" before describing the BenchMetrics in stages conducted on the metric-spaces.

Metric-space: metric's outcome distribution in pseudo-universal "base performance measure permutations"

All possible performance results (*i.e.* classifiers' outcomes) of a binary classification conducted on a dataset with a sample size (*Sn*) are the permutations of base measures (*TP*, *FP*, *FN*, *TN*), a total of which yields the same *Sn*. We define this vector with four elements in Definition 1 below and call it "base performance measure permutations" or shortly "base-measure permutations" ($\mathbf{BM}^{Sn}$) that provide a *pseudo-universal* space for analyzing metrics' outcomes.

A metric-space (**M**) defined in this study keeps the corresponding performances in terms of a specific metric (M) per each permutation. It allows us to analyze and compare how metrics summarize classification outcomes in the complete coverage of performance results. It is 'pseudo' because of the sample size (*Sn*) dependency. In this paper, we represent metric-spaces in bold (*e.g.*, **ACC** metric-space vector for ACC metric), single metric values in italic (*e.g.*, *ACC* = 0.9), and set or array of metric values in bold-italic (*e.g.*, ***BM*** = {*TP* = 7, *FP* = 1, *FN* = 0, *TN* = 2}). Definition 2 expresses a metric-space for a given *Sn*.

Definition 1 (Universal Base-Measure Permutations) A vector $\mathbf{BM}^{Sn}$ shows all possible base-measure permutations with repetition where each *i*th element of $\mathbf{BM}^{Sn}$ is $\mathbf{BM}_i^{Sn}: \boldsymbol{BM} \to \mathbb{Z}^{*4}$ and $\boldsymbol{BM} = \{TP, FP, FN, TN\}$ and $TP_i + FP_i + FN_i + TN_i = Sn$ and $0 \leq TP_i, FP_i, FN_i, TN_i \leq Sn$.

Definition 2 (Metric-Space) A metric-space **M** or $\mathbf{M}^{Sn}$ covers the outputs given by an M metric for all the elements of $\mathbf{BM}^{Sn}$ base-measure permutations for a sample size of *Sn*.

For example, there are 286 permutations of four base measures with repetition for ten samples, where the sum of the base measures is equal to ten, as shown in Fig. 4. An example permutation is ten true positives only (*TP* = 0, all others are zero). Another example might be seven true positives, one false positive, and two true negatives (*TP* = 7, *FP* = 1, *FN* = 0, *TN* = 2). Metrics summarize base measures (confusion matrix) into a single figure (*e.g.*, summarizing *TP* = 7, *FP* = 1, *FN* = 0, *TN* = 2 permutation as *ACC* = 0.9), and metric-spaces provide all possible performance metric values that are calculated for any metric covering each permutation (**F1**, **ACC**, and **MCC** examples are depicted in Fig. 4).

Metric-spaces and dataset size (Sn)

The size of base-measure permutations and corresponding metric-spaces increases exponentially[2] with *Sn*. For instance, the size is 2,667,126 permutations for a dataset with 250 samples. Metrics are the ratios (*e.g.*, *TC* / *Sn* where $0 \leq TC \leq Sn$) in a closed interval (either [0, 1] or [-1, 1]), and the sample size is reduced in the numerator/denominator of the metrics' equations. Hence the size reflects metric-space granularity that is the precision in transitions in different permutations. The metric-space's overall characteristics are the same for different *Sn* values (the density graphs shown in Fig. 3 are similar and descriptive statistics converges).

BenchMetrics analyzes the robustness of metrics over metric-spaces, not datasets (*e.g.*, ACC can be 1.0 for any dataset size). Hence, no classification cases including the extreme cases such as (the lowest performance case: *TP* = 0 and *TN* = 0 where *FC* = *FP* + *FN* > 0), (the extreme class imbalance case: *P* = 0 or *N* = 0), and (the extreme bias case: *OP* = 0 or *ON* = 0) are missing in the analysis. The mathematical characteristics (*e.g.*, metric-space distribution) and the relation between metric-spaces (e.g., **ACC** and **PREV** metric-spaces) are evaluated from different aspects. Note that we tested the related benchmarking criteria with different *Sn* values in our experiments. We observed that the results are the same (for two meta-metrics, namely *UBMcor* and *UIMBucor*) or converge as *Sn* increases. For the remaining five meta-metrics, we averaged meta-metric intermediate values calculated for metric-spaces generated for *Sn* = 25, 50, 75, 100, 125, 150, 175, 200, and 250 to eliminate the possible dataset-size effect. We limit the maximum sample size to 250 to keep the permutation granularity and calculation time in a reasonable precision and range. Calculation of the meta-metrics in metric-spaces up to a sample size of 250 (except for consistency and discriminancy meta-metrics) takes a maximum of one minute on an R version 3.5.2 (2018-12-20) platform on a Darwin 15.6.0 operating system with 2.3 GHz CPU and 16 GB RAM. The calculation of the complete set of proposed meta-metrics for a single metric, including consistency and discriminancy,

[2] Sample sizes (permutations/metric-space sizes): *Sn* = 25 (3,276); *Sn* = 50 (23,426); *Sn* = 75 (76,076); *Sn* = 100 (176,851); *Sn* = 125 (341,376); *Sn* = 150 (585,276); *Sn* = 175 (924,176); *Sn* = 200 (1,373,701); *Sn* = 250 (2,667,126)

takes 21 hours and 45 minutes. Note that detailed time test results and metric-space data for different sample sizes between 10 and 250 are provided in the online material.

3.2 Stage-1: Mathematical evaluation benchmarking

In this stage, we propose eleven criteria to evaluate different metrics from mathematical perspectives.

3.2.1 Criterion-1 – Criterion-3: Performance element coverages

By definition, a metric as a mathematical function M({*TP*, *FP*, *FN*, *TN*}, {*P*, *N*}, {*OP*, *ON*}) should not have a missing facade of fundamental performance element sets, namely four base measures, class measures, and outcome measures, respectively. Otherwise, they cannot be applied to completely summarize the confusion matrix and the number of classes and classification outputs. For example, as the name implies, True Positive Rate ($TPR = TP/P \ = TPR(\{TP\},\{P\},\{\emptyset\})$) reflects only the correct classification of positive-class performance, but it does not sense incorrect classification (*e.g.*, no *FP*) and negative-class performances (*e.g.*, no *N*). We provide the following three criteria that can help to distinguish the limitations of metrics by mathematical functional definition:

Criterion-1 (*Outcome/class coverage*): A metric function should not have outcome measures only (*i.e.* includes {*OP* or *ON*} without {*P* or *N*}) or class measures only (*i.e.* includes *P* or *N*} without *OP* or *ON*}).

Criterion-2 (*Class coverage*): A metric function should fully cover the class arguments ({*P or OP*} *with* {*N or ON*}) without excluding any class.

Criterion-3 (*Base-measure coverage*): A metric function should cover the base performance measures ({*TP*, *FP*, *FN*, *TN*}) without excluding any measure.

3.2.2 Criterion-4 – Criterion-6: Variance/invariance

Contrary to association measures, invariance (*i.e.* not differentiating the swaps among base measures) might not be a desirable characteristic of a robust performance metric. Any change making four base measures of the confusion matrix different should ideally be distinguished. Fig. 2 depicts the three types of swaps we used to assess a toy classification example's metrics' variance. A performance metric should be variant to class swap and variant to outcome swap because base measures become different, as given in Fig. 2 (b) and (c) with the original ones in Fig. 2 (a). Otherwise, the metric does not differentiate such classification results.

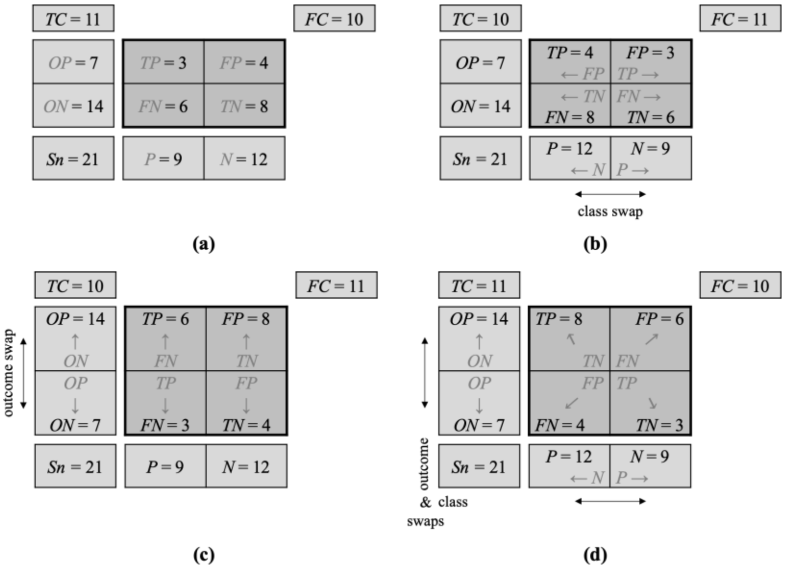


Fig. 2 Three types of swaps of **(a)** an original confusion matrix (base measures): **(b)** Criterion-4: Variant to class swap (horizontally: between *TP* and *FP* along with *FN* and *TN*), **(c)** Criterion-5: Variant to outcome swap (vertically: between *TP* and *FN* along with *FP* and *TN*), and **(d)** Criterion-6: Invariant to class-and-outcome swaps (diagonally: between *TP* and *TN* along with *FP* and *FN*)

To find the variance or invariance of a metric, the base measures in a metric's equation should be changed according to the type of swaps, as shown in Fig. 2 (b – d), and the original and swapped version equations are compared. For example, switching classes in $TPR = TP/P = TP/(TP + FN)$ makes the equation $FP/(FP + TN) = FP/N = FPR$, which is different from the original metric. Hence, TPR is a variant to class swap. Whereas, class-and-outcome exchanges in $MCC = (TP \cdot TN - FP \cdot FN)/\sqrt{P \cdot N \cdot OP \cdot ON}$ result in no variance $(TN \cdot TP - FN \cdot FP)/\sqrt{OP \cdot ON \cdot P \cdot N} = MCC$. Table 1 also shows the known metrics corresponding to each swap. We identified only two metrics that contradict these criteria: nMI and F1. F1 is not invariant to class and outcome swaps because it has no *TN* coverage as addressed in base measure coverage in Table 1. In the literature, Sokolova and Lapalme [13] suggest eight invariance properties, one of which (I_1 as stated) directly corresponds with our criterion, namely class-and-outcome swapping and examines six metrics (TPR, TNR, PPV, ACC, BACC, and F1). The remaining four properties indicate the variance by changing *TN*-only (I_2), *TP*-only (I_3), *FN*-only (I_4), and *FP*-only (I_5), which are quickly evaluated by our base-measure coverage criterion (Criterion-3). For example, F1 has "No *TN*" base measure coverage that corresponds to I_2 invariance. The remaining three properties scale all the base measures, class components separately (*P*: *TP* and *FN* – *N*: *TN* and *FP*), and outcome components separately (*OP*: *TP* and *FP* – *ON*: *TN* and *FN*), which are addressed by our Criterion-1, Criterion-2, and Criterion-3 simply.

3.2.3 Criterion-7 – Criterion-11: Descriptive statistics

The distribution and descriptive statistics such as range, mean, median, and standard deviation calculated for the metric-space of a metric give insights about the dispersions and transitions of metric outputs. Fig. 3 illustrates density graphs along with the range, mean, median, and mode statistics per metric. Each density graph shows the metric-space in terms of relative frequencies per equally spaced breaks in the metric's range. A fitted normal distribution curve over the mean is also attached where possible (**ACC**, **INFORM**, **BACC**, **CK**, and **MCC**). The most important findings shown in Fig. 3 are that the distributions are different, and not all the performance metrics show smooth and continuous transitions. The revealed difference could be another motivation to identify the most robust metric. The following criteria we defined are important for metric evaluation:

Criterion-7 (*Undefined (NaN) counts*): The number of undefined values (not-a-number, NaN) due to 0/0 divisions is listed in Table 1. The NaN count of **MCC** is the highest with proportional to *Sn*, whereas **ACC**, **F1**, and **CK** have zero, one, and two NaNs, respectively, regardless of *Sn*. Robust metrics are expected to calculate any base-measure permutations without any exception.

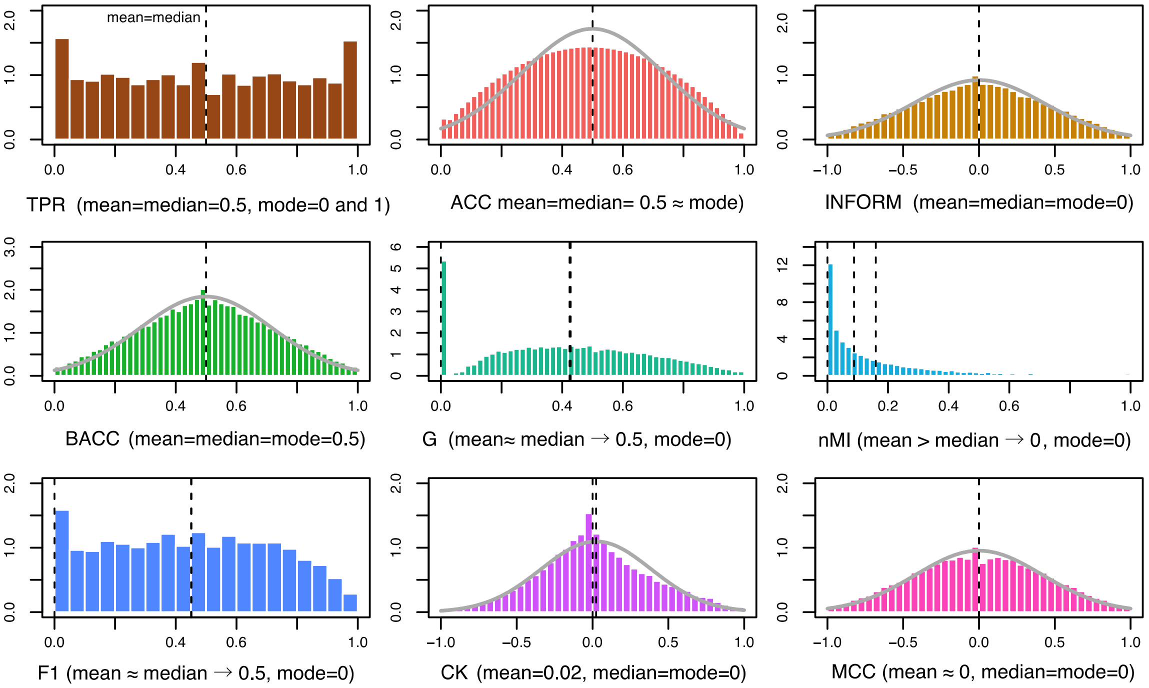


Fig. 3 Density graphs summarizing each of the nine metric-spaces (**TNR**, **PPV**, and **NPV** are the same as **TPR**; **MARK** is the same as **INFORM**) (the area under the distribution is one)

Table 1 Experiment-1: Stage-1 benchmarking results for 13 performance metrics according to 8 proposed criteria along with three informative criteria (Criterion-9 – Criterion-11)

Stage-1 Criteria	**CK**	**MCC**	**F1**	**INFORM**	**MARK**	**BACC**	**G**	**ACC**	**TPR**	**PPV**	**TNR**	**NPV**	**nMI**
1	Yes	Yes	Yes	**Class-only**	**Outcome-only**	**Class-only**	**Class-only**	**None**	**Class-only**	**Outcome-only**	**Class-only**	**Outcome-only**	Yes
2	Yes	Yes	Yes	Yes				**None**	***P*-only**		***N*-only**		Yes
3	Yes	*Yes*	**No *TN***	***TP, TN***				***TP, TN***	***TP*-only**		***TN*-only**		Yes
4	Yes	Yes	Yes	Yes				Yes (*MCR*)	Yes (*FPR*)	Yes (*FDR*)	Yes (*FNR*)	Yes (*FOR*)	***No (nMI)***
5	Yes	Yes	Yes	Yes				Yes (*MCR*)	Yes (*FNR*)	Yes (*FOR*)	Yes (*FPR*)	Yes (*FDR*)	***No (nMI)***
6	Yes	Yes	**No**	Yes				Yes	Yes				Yes
7	2	**4*Sn***	1	**2(*Sn*+1)**				0	***Sn*+1**				4
8	$\underline{\mathbf{\bar{M}} \neq \mathbf{\tilde{M}}} = \mathbf{Mo}$	$\mathbf{\bar{M}} \approx \mathbf{\tilde{M}} = \mathbf{Mo}$	$\mathbf{\bar{M}} \approx \mathbf{\tilde{M}} \neq \mathbf{Mo}$	$\mathbf{\bar{M}} = \mathbf{\tilde{M}} = \mathbf{Mo}$			$\mathbf{\bar{M}} \approx \mathbf{\tilde{M}} \neq \mathbf{Mo}$	$\mathbf{\bar{M}} = \mathbf{\tilde{M}} \approx \mathbf{Mo}$	$\mathbf{\bar{M}} = \mathbf{\tilde{M}} \neq \mathbf{Mo}$				$\underline{\mathbf{\bar{M}} \neq \mathbf{\tilde{M}}} \neq \mathbf{Mo}$
Stage-1 Rank	**1**		**3**	**4**				**8**	**9**				**13**
Other informative criteria (*i.e.* not used in the ranking of the metrics)													
9	**0.18**[(a)]	0.20[(a)]	0.22	0.21[(a)]	0.21[(a)]	0.2	0.23	0.26	0.29	0.29	0.29	0.29	**0.17**
10	Slightly positive/right skewed (0.16)	Symmetric (0)	Slightly positive/right skewed (0.05)	Symmetric (0)	Symmetric (0)	Symmetric (0)	Slightly positive/right skewed (0.18)	Symmetric (0)	Symmetric (0)	Symmetric (0)	Symmetric (0)	Symmetric (0)	**Highly positive/right skewed (1.69)**
11)	Platykurtic (-0.2)	Platykurtic (-0.6)	Platykurtic (-1.07)	Platykurtic (-0.6)	Platykurtic (-0.6)	Platykurtic (-0.6)	Platykurtic (-0.85)	Platykurtic (-0.86)	Platykurtic (-1.2)	Platykurtic (-1.2)	Platykurtic (-1.2)	Platykurtic (-1.2)	**Leptokurtic (2.75)**

Criteria **Criterion-1** Outcome/class (OP and ON *vs.* P and N) coverage; **Criterion-2** Class (P with N or OP with ON) coverage; **Criterion-3** Base-measure (TP, FP, FN, and TN) coverage;
Criterion-4 Variant to class swap[(b)]; **Criterion-5** Variant to outcome swap[(b)]; **Criterion-6** Invariant to class-and-outcome swaps;
Criterion-7 Undefined (NaN) count ; **Criterion-8** Central tendencies (mean-median difference)[(c)]
Informative Criteria
Criterion-9 Standard Deviation; **Criterion-10** Skewness; **Criterion-11** Kurtosis[(d)]

Notes: **(a)** When normalized into [0, 1].
(b) The corresponding metric name after swapping is displayed in braces.
(c) $\mathbf{\bar{M}}$: mean, $\mathbf{\tilde{M}}$: median, and **Mo**: mode of a metric-space
(d) Kurtosis types: Platykurtic: thin-tailed, Leptokurtic: fat-tailed, Mesokurtic (normal tail shape)

Criterion-8 (*Central tendencies*): The central tendency defined by mean, median, and mode should be examined for metric-spaces. Only **INFORM**, **MARK**, and **BACC** have precisely the same three central tendencies. However, a mean-median difference (*i.e.* arithmetic *vs.* positional average in sorted metric-space) was observed in **nMI** and **CK** (even though **CK** is symmetric). The difference could indicate an imbalance in mapping the uniform classification performance results (*i.e.* base-measure permutations) to the corresponding uniform output ranges of a metric-space.

Criterion-9 (*Standard deviation*): Informatively, the standard deviation of **nMI** and **CK** are the lowest, indicating low dispersion around their mean values, whereas others disperse over a higher range of values in metric-space as can be seen in Fig. 3.

Criterion-10 (*Skewness*) and **Criterion-11** (*Kurtosis*): The shape of distributions: Table 1 shows two measures to recognize the form of metric-space distribution and dispersion shown in the graphs in Fig. 3. Most metric-spaces are symmetric and platykurtic (thin-tailed) except **CK**, **F1**, **G**, and **nMI**. Note that **G** and **F1** metric-spaces exhibit unexpected distortions by yielding zero dominantly, which indicates the particular accumulation points in metric-space. Table 1 shows the results of the Stage-1 benchmarking along with the metrics' ranks. The underlined bold texts depict the deficiencies, and each criterion is taken as equally important. The last three criteria are informative and not included in benchmarking ranking.

3.3 Stage-2: Meta-metrics benchmarking

Stage-2 measures the robustness of performance metrics via our proposed meta-metrics (*i.e.* metrics about (performance) metrics). The meta-metrics in [0, 1] interval are calculated in metric-spaces (0 for the least and 1 for the most robust case). We obtained each meta-metric for the reviewed performance metrics such as Accuracy or MCC in different *Sn* sample sizes in our experiments. We observed that some meta-metric values are equal regardless of the sample size or converge consistently as *Sn* increases. For the latter case, we calculated the meta-metric values for several *Sn* values and obtained their averages as the final meta-metric value. Fig. 4 depicts six of the seven proposed meta-metrics calculated for some example metrics in a sample size of ten.

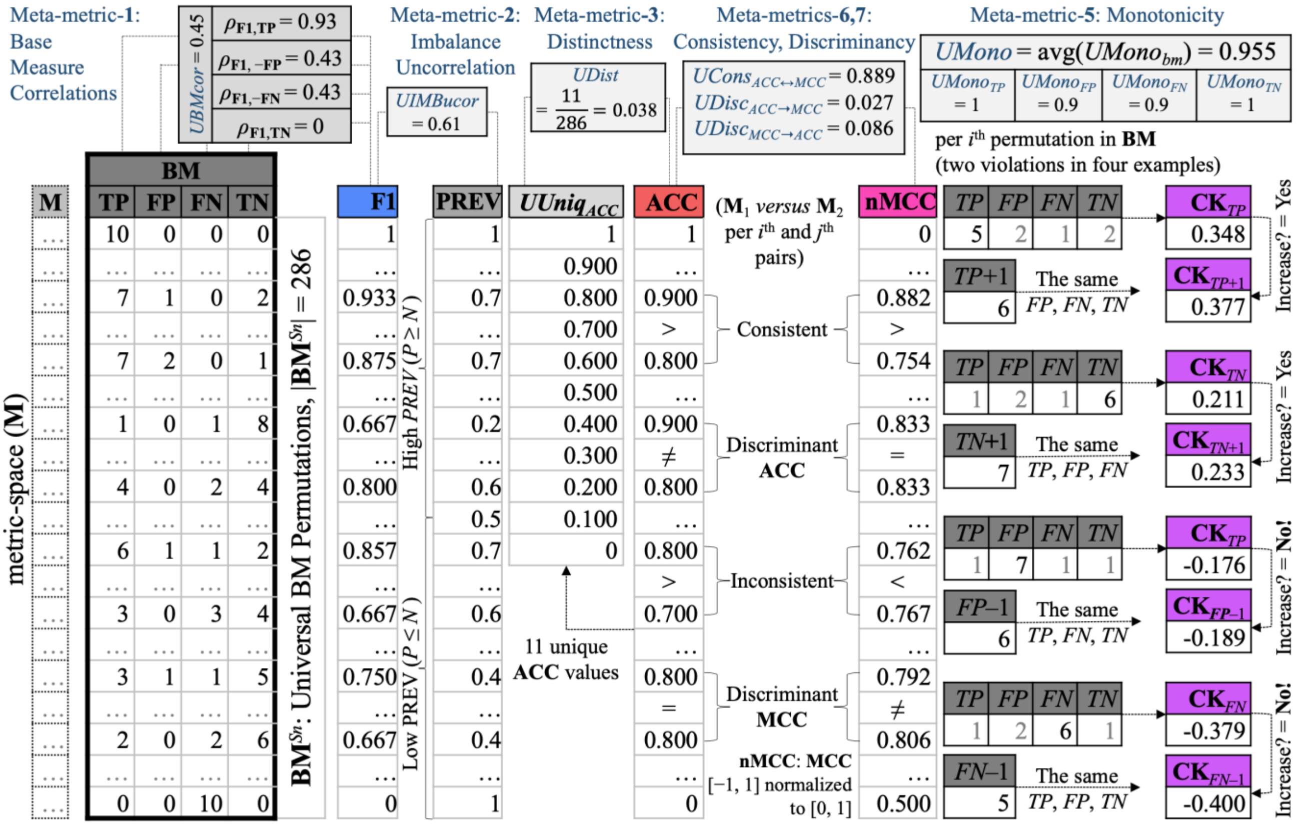


Fig. 4 Depiction of six of seven meta-metrics for 286 base-measure permutations (sample size 10): 1) *UBMcor* for F1 metric; 2) *UIMBucor* for F1; 3) *UDist* for ACC; 4) *UMono* for CK; and 5-6) *UCons* and *UDisc* for ACC versus nMCC (MCC normalized into [0, 1] interval) (refer to Section 3.3.4 and Fig. 3 for *UOsmo* meta-metric)

The following subsections describe and give formal definitions of each meta-metric along with Experiment-1 intermediate results.

3.3.1 Meta-metric-1: Base measure correlations (*UBMcor*)

The correlation between a metric-space and each base measure gives their degree of relationship. Robust metrics should be correlated with all base performance measures from an objective perspective unless otherwise required. The correlations with **FP** and **FN** must be negative for a performance metric (*i.e.* it should go higher as *FP* and *FN* go lower). Fig. 4 shows the **F1** metric-space with corresponding **BM** permutations as an example. The correlations with **TP**, **−FP**, **−FN**, and **TN**, along with the final *UBMcor* meta-metric value, are also displayed.

We used Spearman correlation (ρ, 'rho') in two meta-metrics, which is less sensitive to outliers than Pearson correlation that also assumes linearity among the metric-space and base measure spaces (or prevalence spaced for *UIMBucor* meta-metric described below). Spearman correlation significance level (α) is taken as 0.05. Table 2 lists Spearman's rho correlation values for all benchmarked metrics. Recall that underlined bold texts depict the deficiencies.

Table 2 Experiment-1: meta-metric *UBMcor* values [0, 1] and correlations with **TP**, **−FP**, **−FN**, **TN**

	ρ	ACC	MCC	INFORM	MARK	BACC	CK	G	F1	TPR	PPV	TNR	NPV	nMI
Correlations	**TP**	0.55*	0.55*	0.54*	0.54*	0.54*	0.53*	0.54*	0.93*	0.78*	0.78*	**0**	**0**	**-0.05***
	TN	0.55*	0.55*	0.54*	0.54*	0.54*	0.53*	0.54*	**0**	**0**	**0**	0.78*	0.78*	**-0.05***
	−FP	0.55*	0.55*	0.54*	0.54*	0.54*	0.55*	0.49*	0.43*	**0**	0.78*	0.78*	**0**	0.05*
	−FN	0.55*	0.55*	0.54*	0.54*	0.54*	0.55*	0.49*	0.43*	0.78*	**0**	**0**	0.78*	0.05*
	UBMcor	0.55	0.55	0.54	0.54	0.54	0.54	0.52	0.45	0.39	0.39	0.39	0.39	0.00

* correlation is significant at the 0.05 level

UBMcor meta-metric reveals that **F1** has zero correlation with **TN** values, whereas it is highly correlated with **TP** but lower correlated with false positives/negatives than true positives. **CK** has a lower correlation with true positives/negatives (*i.e.* more emphasis on performance errors than successes) compared to the others. ***G*** is class-balanced (*i.e.* correlations for **TP** *vs.* **TN** and **−FP** *vs.* **−FN** are the same), but it is lower correlated with negative false positives/negatives than true positives/negatives (0.49 < 0.54). **ACC**, **INFORM**, **MARK**, **BACC**, and **MCC** are ideally all balanced (*i.e.* absolute correlations for **TP** *vs.* **−FP** *vs.* **TN** *vs.* **−FN** are the same). **nMI** has the lowest correlations with base measures.

The meta-metric *UBMcor* is the arithmetic average of correlations of a metric-space (**M**) with each of the four base-measure-spaces (false positive/negatives are negated), as calculated in Eq. (1).

$$UBMcor = {}^{1}\!/_{4}\left(\rho_{\mathbf{M},\mathbf{TP}} + \rho_{\mathbf{M},-\mathbf{FP}} + \rho_{\mathbf{M},-\mathbf{FN}} + \rho_{\mathbf{M},\mathbf{TN}}\right) \tag{1}$$

3.3.2 Meta-metric-2: (Class) imbalance uncorrelation (*UIMBucor*)

Robust metrics should not be influenced by class imbalance (i.e., increasing/decreasing performance values by changing the class ratios). The correlation between corresponding sub-metric-spaces and two half-ranges of **PREV** ([0, 0.5] and [0.5, 1]) shows the degree of bias between classification performance and class imbalance. Fig. 4 above shows the **F1** metric-space and corresponding **PREV** values above and below balanced classes (*PREV* ≤ 0.5 and *PREV* ≥ 0.5). Meta-metric *UIMBucor* is calculated by for a metric-space **M**, as shown in Eq. (2), where $\mathbf{M}_a$ represents the sub-metric-space according to condition *a*.

$$UIMBucor = 1\text{-}\ {}^{\left(\left|\rho_{\mathbf{M}_{P\le N},\mathbf{PREV}_{P\le N}}\right| + \left|\rho_{\mathbf{M}_{P\ge N},\mathbf{PREV}_{P\ge N}}\right|\right)}\!\big/_{2} \tag{2}$$

Hence, for example, $\rho_{\mathbf{M}_{P\le N},\mathbf{PREV}_{P\le N}}$is the correlation between **M** and **PREV** sub-metric-spaces having corresponding pairs where $P \le N$ (*i.e.* *PREV* ≤ 0.5), vice versa. Note that absolute correlations are used because the direction of the correlation (*i.e.* whether correlated or inverse correlated) is not essential; hence the two correlations will not cancel out each other. As shown in Table 3, only **PPV**, **NPV**, **F1**, **nMI**, **CK**, and **G** are correlated with class imbalance regardless of the sample sizes.

Table 3 Experiment-1: meta-metric *UIMBucor* values [0, 1]

	TPR	TNR	ACC	INFORM	MARK	BACC	MCC	G	CK	nMI	F1	PPV	NPV
UIMBucor	1	1	1	1	1	1	1	**<u>0.97*</u>**	**<u>0.96*</u>**	**<u>0.91*</u>**	**<u>0.64*</u>**	**<u>0.55*</u>**	**<u>0.55*</u>**

* correlation is significant at the 0.05 level (the correlations of the metric-spaces (*e.g.*, **TPR**) with *UIMBucor* is zero and are not significant)

3.3.3 Meta-metric-3: Distinctness (*UDist*)

As each base-measure permutation differs from each other, a robust metric should differentiate these different cases in metric-space. Fig. 4 above depicts how *UDist* is calculated for ACC metric as an example. The number of unique values of the metric-space (*e.g.*, 11 unique values for **ACC**) is compared against the metric-space size (the number of unique values in **BM** permutations, *e.g.*, 286 for *Sn* = 10). The distinctness meta-metric defined formally below gives metrics' granularity in metric-space as listed in Table 4.

Definition 3 (Universal Distinctness) *UDist* measures the ratio of unique values in the metric-space of a metric *M* where **M**: $\mathbf{BM}^{Sn} \rightarrow \overline{\mathbb{R}}$ and $\mathbf{UUniq}$ is a finite set where **M**: $\mathbf{UUniq} \rightarrow \overline{\mathbb{R}}_{\geq 1}$ and $UDist = |\mathbf{UUniq}|/|\mathbf{BM}^{Sn}|$.

Table 4 Experiment-1: meta-metric *UDist* minimum, average, and maximum values [0, 1]

UDist	nMI	BACC	INFORM	MARK	MCC	CK	G	TPR	TNR	PPV	NPV	F1	ACC
Min	0.3	0.3	0.3	0.3	0.23	0.17	0.18	0.007	0.007	0.007	0.007	0.007	0.0001
Average	0.38	0.35	0.35	0.35	0.24	0.20	0.20	**<u>0.02</u>**	**<u>0.02</u>**	**<u>0.02</u>**	**<u>0.02</u>**	**<u>0.02</u>**	**<u>0.001</u>**
Max	0.4	0.4	0.4	0.4	0.24	0.24	0.20	0.06	0.06	0.06	0.06	0.06	0.008

Sample Size and Permutations: *Sn*=25 (3,276); *Sn*=50 (23,426); *Sn*=75 (76,076); *Sn*=100 (176,851); *Sn*=125 (341,376); *Sn*=150 (585,276); *Sn*=175 (924,176); *Sn*=200 (1,373,701); *Sn*=250 (2,667,126)

To the contrary of the first two meta-metrics, *UDist* values might differ per *Sn*. We calculated *UDist* values for nine different sample sizes (given in the footnotes of Table 4) and benchmarked the metrics according to their average values. While nMI has the most distinct metric-space, ACC has the least. Unexpectedly, F1 has the same level of distinctness as TPR, TNR, PPV, and NPV metrics.

3.3.4 Meta-metric-4: Output smoothness (*UOsmo*)

Output smoothness evaluates how a metric uniformly uses its output range. As each variation in corresponding base measures is a unit change, a metric-space should exhibit a smooth transition. Fig. 5 shows the transition of metric-spaces sorted in ascending order.

Unexpectedly, we discovered a repeating stepped transition in **ACC**. Moreover, as mentioned in the shape of distributions criteria in Stage-1, **G** and **F1** dominantly yield zero. Stepped changes indicate a robustness defect where a metric produces coarse resolution in steps or accumulates in some values. These behaviors degrade a metric's ability to differentiate different classification results (*e.g.*, two classifiers' performance are more likely to fall into the same amount than if a smoother metric is used). Eq. (3) is used to measure the smoothness without a need for visual inspection where $\mathbf{\Delta}^{Sn} = \{\delta_i \mid i = 2, \ldots, Sn\}$ and $\mathbf{\Delta}^{Sn}{}_{\text{absolute}} = \{|\delta_i| \mid i = 2, \ldots, Sn\}$ and $\delta_i = \mathbf{sM}_i - \mathbf{sM}_{i-1}$.

$$osmo^{Sn} = \text{std}(\mathbf{\Delta}^{Sn})/\text{avg}(\mathbf{\Delta}^{Sn}{}_{\text{absolute}}) \qquad (3)$$

sM denotes the sorted metric-space in increasing order, **sM**$_i$ denotes the i^{th} value of the sorted metric-space, and std and avg are the standard deviation and average (arithmetic mean) functions. The equation calculates the coefficient of variation for one lagged self-difference. The minimum the result, the maximum the smoothness is.

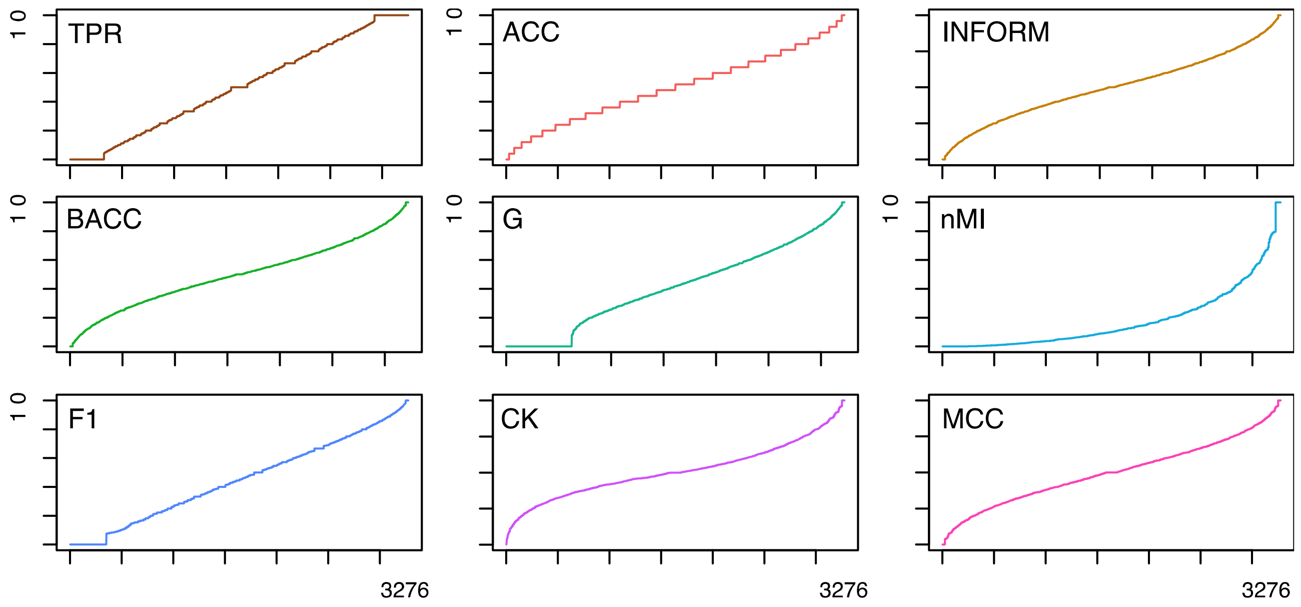


Fig. 5 Sorted metric-spaces' transitions. The transitions **MARK** and **INFORM** and **TNR**, **PPV**, **NPV**, and **TPR** are similar (*y*-axis shows the metric's outputs, and the *x*-axis shows the index of the elements in the metric-space, total 3,276 for *Sn* = 25)

We average the smoothness values calculated for the sample sizes between 25 and 250 (*see* Table 5's notes for the sample sizes) and Eq. (4) to get the *UOsmo* meta-metric for a metric M*k* by normalizing the smoothness values (*osmo*) among all the compared metric-spaces (METRICS, *e.g.*, benchmarked 13 metrics). Table 5 shows the smoothness and *UOsmo* meta-metric values for the compared metrics.

$$UOsmo_{\mathrm{M}k} = \frac{\operatorname{avg}\left(osmo_{\mathrm{M}k}^{Sn=25,\cdots,250}\right) - \min\limits_{\forall \mathrm{M}l}\left\{\operatorname{avg}\left(osmo_{\mathrm{M}l}^{Sn=25,\cdots,250}\right)\right\}}{\max\limits_{\forall \mathrm{M}l}\left\{\operatorname{avg}\left(osmo_{\mathrm{M}l}^{Sn=25,\cdots,250}\right)\right\} - \min\limits_{\forall \mathrm{M}l}\left\{\operatorname{avg}\left(osmo_{\mathrm{M}l}^{Sn=25,\cdots,250}\right)\right\}},\ \mathrm{M}k, \mathrm{M}l \in \mathrm{METRICS} \qquad (4)$$

Table 5 Experiment-1: meta-metric *UOsmo* values [0, 1] and the minimum, average, and maximum smoothness values per base measure

		INFORM	**MARK**	**BACC**	**CK**	**MCC**	**G**	**TPR**	**TNR**	**PPV**	**NPV**	**F1**	**nMI**	**ACC**
Min	Smoothness (*osmo*)[(a)]	2.07	2.07	2.07	2.92	3.02	3.79	3.39	3.39	3.39	3.39	4.02	6.94	5.25
Average		4.73	4.73	4.73	8.08	8.46	11.67	15.61	15.61	15.61	15.61	18.03	45.44	91.71
Max		9.79	9.79	9.79	16.74	18.94	27.07	41.73	41.73	41.73	41.73	47.70	135.93	409.47
	UOsmo	1	1	1	0.96	0.96	0.92	0.87	0.87	0.87	0.87	0.85	0.53	0

(a) Minimum, average, and maximum smoothness are calculated for Sn=25, 50, 75, 100, 125, 150, 175, 200, and 250

ACC and nMI have the least smooth metric-spaces, following Fig. 3, whereas CK and MCC have slightly unsmooth metric-spaces than INFORM, MARK, and BACC.

3.3.5 Meta-metric-5: Monotonicity (*UMono*)

A robust metric should be sensitive to small changes in classification performance. *UMono* meta-metric is calculated per four base measures by increasing *TP* and *TN* by one and decreasing *FP* and *FN* by one separately for each of **BM** permutations (hence raising classification performance in the smallest scale) and checking whether the new metric value does not decrease. Otherwise, it is considered an explicit violation in a metric-space. The formal definition is given in Definition 4. Our analysis reveals that the metrics do have 100% monotonicity except for **INFORM**, **MARK**, **BACC**, **nMI**, and **CK**, as listed in Table 6.

Table 6 Experiment-1: meta-metric *UMono* values [0, 1] per base measure (the metrics are sorted according to *UMono* values and the average of the four meta-metric sub-values: $UMono_{TP}$, $UMono_{TN}$, $UMono_{FP}$, $UMono_{FN}$)

UMono	TPR	TNR	PPV	NPV	ACC	G	F1	MCC	INFORM	MARK	BACC	CK	nMI
$UMono_{TP}$	1	1	1	1	1	1	1	1	**0.9990**	**0.9990**	**0.9990**	1	**0.5029**
$UMono_{TN}$	1	1	1	1	1	1	1	1	1	1	1	1	**0.5029**
$UMono_{FP}$	1	1	1	1	1	1	1	1	1	1	1	**0.9005**	**0.5032**
$UMono_{FN}$	1	1	1	1	1	1	1	1	1	1	1	**0.9005**	**0.5032**
UMono	1	1	1	1	1	1	1	1	0.9995	0.9995	0.9995	0.9502	0.5031

Definition 4 (Universal Monotonicity) $UMono_{bm}$ gives the ratio of cases where a metric-space **M** adjusts its performance value congruous with the unit changes (±1) by $bm \in \{TP,\ TN,\ FP,\ FN\}$ in metric-space. $\forall\ \mathbf{M}_i\colon \mathbf{BM}^{Sn} \to \overline{\mathbb{R}}$ and $\mathbf{M}_{i\pm}\colon \mathbf{BM}^{Sn\pm1} \to \overline{\mathbb{R}}$ where $i = 1, \ldots, Sn$:

$$\mathbf{M}_{i+}\colon \mathbf{BM}^{Sn+1} = \begin{cases}\{TP_i + 1,\ FP_i,\ \ FN_i,\ \ TN_i\}, & bm = TP\\ \{TP_i,\ FP_i,\ \ FN_i,\ \ TN_i + 1\}, & bm = TN\end{cases}$$
$$\mathbf{M}_{i-}\colon \mathbf{BM}^{Sn-1} = \begin{cases}\{TP_i,\ FP_i - 1,\ FN_i,\ \ TN_i\}, & bm = FP\\ \{TP_i,\ FP_i,\ \ FN_i - 1,\ \ TN_i\}, & bm = FN\end{cases}$$
$$\mathbf{Mono}_{bm} = \{(\mathbf{M}_i, \mathbf{M}_{i\pm})\colon \mathbf{M}_{i\pm} \geq \mathbf{M}_i\}$$
$$UMono_{bm} = |\mathbf{Mono}_{bm}|/|\mathbf{BM}^{Sn}|$$

CK −as parallel to *UBMcor* meta-metric shown in Table 2− has 90% monotonicity for *FP* and *FN* decrements (10% violations), and **BACC** has 99% monotonicity (1% violation) for *TP* and *TN* increments. For example, **CK** is −0.176 for *TP* = 1, *FP* = 7, *FN* = 1, *TN* = 1, as shown in Fig. 4 above. Decreasing *FP* only by one (*FP* = 7 – 1 = 6) should increase the performance, but **CK** yields −0.189 violating monotonicity (*i.e.* −0.189 < −0.176). Increasing *TP* only by one (*TP* = 1+1, *FP* = 7, *FN* = 1, *TN* = 1) yields −0.128 preserving monotonicity (−0.128 > −0.176). In the worst case, **nMI** monotonicity violations are almost exactly half-and-half.

3.3.6 Meta-metric-6 and 7: Inconsistency/consistency (*UICons*/*UCons*) and discriminancy (*UDisc*)

The meta-metrics formally defined below are proposed for comparing two metrics' robustness.

Definition 5 (Universal Consistency and Inconsistency) $UCons_{M_1 \leftrightarrow M_2}$ and $UICons_{M_1 \leftrightarrow M_2}$ give the agreement and disagreement in increments/decrements in metric-space of two metrics M_1 and M_2, respectively, where $\mathbf{M}_1, \mathbf{M}_2\colon \mathbf{BM}^{Sn} \to \overline{\mathbb{R}}$. $\forall\ \mathbf{M}_{1_i}, \mathbf{M}_{1_j}, \mathbf{M}_{2_i}, \mathbf{M}_{2_j}$ (different corresponding pairs of i^{th} and j^{th} values of $\mathbf{M}_1$ and $\mathbf{M}_2$) where *i*, *j* = 1, …, *Sn* and *i* ≠ *j*:

$$\mathbf{ICons}_{M_1 \leftrightarrow M_2} = \left\{\begin{matrix}(\mathbf{M}_{1_i}, \mathbf{M}_{1_j}), (\mathbf{M}_{2_i}, \mathbf{M}_{2_j})\colon \\ \Big(\big(\mathbf{M}_{1_i} > \mathbf{M}_{1_j}\big) \wedge \big(\mathbf{M}_{2_i} < \mathbf{M}_{2_j}\big)\Big) \vee \Big(\big(\mathbf{M}_{1_i} < \mathbf{M}_{1_j}\big) \wedge \big(\mathbf{M}_{2_i} > \mathbf{M}_{2_j}\big)\Big)\end{matrix}\right\}$$
$$UICons_{M_1 \leftrightarrow M_2} = \left|\mathbf{ICons}_{\mathrm{M}_1 \leftrightarrow \mathrm{M}_2}\right| / \binom{|\mathbf{BM}^{Sn}|}{2}$$
$$UCons_{M_1 \leftrightarrow M_2} = 1 - \mathbf{UICons}_{M_1 \leftrightarrow M_2}$$

Definition 6 (Universal Discriminancy) $UDisc_{M_1 \to M_2}$ gives the ratio of cases where the metric M_1 yields different values while the metric M_2 could not differentiate in metric-spaces where $\mathbf{M}_1, \mathbf{M}_2\colon \mathbf{BM}^{Sn} \to \overline{\mathbb{R}}$. $\forall\ \mathbf{M}_{1_i}, \mathbf{M}_{1_j}, \mathbf{M}_{2_i}, \mathbf{M}_{2_j}$ (different corresponding pairs of i^{th} and j^{th} values of $\mathbf{M}_1$ and $\mathbf{M}_2$) where *i*, *j* = 1, …, *Sn* and *i* ≠ *j*:

$$\mathbf{Disc}_{M_1 \to M_2} = \left\{\begin{matrix}\big(\mathbf{M}_{1_i}, \mathbf{M}_{1_j}\big), \big(\mathbf{M}_{2_i}, \mathbf{M}_{2_j}\big)\colon \\ (\mathbf{M}_{1_i} \neq \mathbf{M}_{1_j}) \ \wedge (\mathbf{M}_{2_i} = \mathbf{M}_{2_j})\end{matrix}\right\}$$
$$UDisc_{M_1 \to M_2} = \left|\mathbf{Disc}_{M_1 \to M_2}\right| / \binom{|\mathbf{BM}^{Sn}|}{2}$$

Fig. 4 above depicts the example cases on **ACC** and **nMCC's** real metric values (**MCC** normalized to [0, 1]) where Sn = 10. Among all possible i^{th} and j^{th} pairs, the first given example pairs are consistent because i^{th} values (*ACC* = 0.900 and *nMCC* = 0.882) are greater than j^{th} values (*ACC* = 0.800 and *nMCC* = 0.754) for both metrics. However, in the third example, the pairs are inconsistent because the i^{th} value is greater than the j^{th} value for *ACC* (0.800 > 0.700), but the i^{th} value is less than the j^{th} value for *nMCC* (0.762 < 0.767). For discriminancy, ACC is discriminant against *nMCC* in the second example, because *ACC* yields different values (0.900 ≠ 0.800) where *nMCC* yields the same value (0.833 = 0.833) for corresponding pairs. Likewise, *nMCC* is discriminant against *ACC* in the fourth example. Note that *UICons*/*UCons* and *UDisc* meta-metrics are based on Huang and Ling's two formal criteria for comparing two metrics [14]. The application of these criteria ("degree of consistency" and "degree of discriminancy") has become one of the most used comparative methods in the literature. Our improvement is transforming the degrees that are ranged differently per compared metrics into a fixed ratio in [0, 1] representing the cases concerning the universal **BM** permutations. Hence, our meta-metrics can be used for comparing more than two performance metrics, as shown in Tables 7 and 8.

Table 7 Experiment-1: *UCons* values per pairs of metrics and final *UCons* meta-metric values (the average of the meta-metric values per performance metric)

	MCC												
	0.96	**INFORM**											
	0.96	1.00	**BACC**										
	0.96	0.94	0.94	**CK**									
	0.96	0.91	0.91	0.94	**MARK**								
	0.90	0.91	0.91	0.89	0.89	**G**							
	0.88$^{(a)}$	0.88	0.88	0.87	0.88	0.86	**ACC**						
	0.79	0.79	0.79	0.78	0.79	0.81	0.83	**F1**					
	0.76	0.77	0.77	0.75	0.76	0.77	0.76	0.85	**TPR**				
	0.76	0.76	0.76	0.75	0.77	0.76	0.76	0.85	0.69	**PPV**			
	0.76	0.77	0.77	0.75	0.76	0.77	0.76	0.60	0.53	0.69	**TNR**		
	0.76	0.76	0.76	0.75	0.77	0.76	0.76	0.60	0.69	0.53	0.69	**NPV**	
	0.50	0.50	0.50	0.51	0.50	0.54	0.52	0.53	0.52	0.52	0.52	0.52	**nMI**
UCons:	**0.83**$^{(b)}$	0.83	0.83	0.82	0.82	0.81	**0.80**$^{(c)}$	0.75	0.72	0.72	0.70	0.70	0.51
Rank:	1	1	1	4	4	6	7	8	9	9	11	12	13

Examples: the cell marked with **(a)** (the consistency between **ACC** and **MCC**) is 88% ($UCons_{ACC\leftrightarrow MCC}$ = 0.88), *UCons* for **MCC** (the average meta-metric values for **MCC**) and **ACC** are the cell marked with **(b)** (0.83) and the cell marked with **(c)** (0.80), respectively.

Table 8 Experiment-1: *UDisc* values [0, 1] per ordered pairs of metrics. The metrics are sorted according to the average of the meta-metric values per metric. The arrows depict the direction of the discriminancy. Taking "↳ **F1** ↰" as an example, the first arrow "↳" shows $UDisc_{\mathrm{F1}\rightarrow\mathrm{ACC}}$, (*i.e.* **F1**'s discriminancy against **ACC**, as the first below and right metric) whereas the last arrow " ↰" shows $UDisc_{\mathrm{F1}\rightarrow\mathrm{MARK}}$, (*i.e.* **F1**'s discriminancy against **MARK**, as the first top and left metric).

	↳ **nMI**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.001	↳ **CK** ↰	0.000	0.001	0.001	0.001	0.002	0.001	0.002	0.002	0.002	0.002	0.001
	0.001	0.000	↳ **MCC** ↰	0.001	0.001	0.001	0.002	0.001	0.002	0.002	0.002	0.002	0.001
	0.001	0.001	0.001	↳ **BACC** ↰	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.001	0.001	0.001	0.000	↳ **INFORM** ↰	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.001	0.001	0.001	0.001	0.001	↳ **MARK** ↰	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.018	0.018	0.018	0.017	0.017	0.017	↳ **F1** ↰	0.014	0.018	0.018	0.007	0.007	**0.006**$^{(a)}$
	0.044	0.043	0.043	0.044	0.044	0.044	0.040	↳ **ACC** ↰	0.043	0.043	0.043	0.043	0.042
	0.029	0.029	0.029	0.029	0.029	0.027	0.028	0.028	↳ **TNR** ↰	0.019	0.029	0.019	0.019
	0.029	0.029	0.029	0.027	0.027	0.029	0.028	0.028	0.019	↳ **NPV** ↰	0.019	0.029	0.018
	0.029	0.029	0.029	0.029	0.029	0.027	0.019	0.028	0.029	0.019	↳ **TPR** ↰	0.019	0.019
	0.029	0.029	0.029	0.027	0.027	0.029	0.019	0.028	0.019	0.029	0.019	↳ **PPV** ↰	0.018
	0.039	0.038	0.038	0.038	0.038	0.034	**0.028**$^{(b)}$	0.037	0.029	0.028	0.029	0.028	**G** ↰
UDisc:	0.019	0.018	0.018	0.018	0.018	0.018	0.014	0.014	0.014	0.014	0.013	0.013	0.011
Rank:	1	2	2	2	2	2	7	7	7	7	11	11	13

The cell marked with **(a)** shows **G**'s discriminancy against **F1** ($UDisc_{\mathrm{G}\rightarrow\mathrm{F1}}$) is 0.6%.

The cell marked with **(b)** shows **F1**' discriminancy against **G** ($UDisc_{\mathrm{F1}\rightarrow\mathrm{G}}$) is 2.8%, shown in bold.

Table 7 shows the *UCons* values calculated for *Sn* = 25 per pair of the reviewed metric pairs and final *UCons* values that are the average of a metric with all the others (*e.g.*, $UCons_{\text{ACC}}$ = avg($UCons_{\text{ACC}\leftrightarrow\text{per other metrics}}$). **MCC**, **INFORM**, and **BACC** are the most consistent with the other metrics on average (83%), whereas **nMI** is the least consistent metric (51%). For individual pairs, **INFORM** and **BACC** are only 100% consistent (*i.e.* $UCons_{INFORM\leftrightarrow BACC}$ = 1.00). Table 8 shows the *UDisc* values per ordered pairs of metrics analyzed in 25 samples. **nMI**, the least consistent metric, is the most discriminant metric (about 1%). Interestingly, **MCC** is the most consistent and the third discriminant metric at the same time. The table also illustrates another important finding that all the metrics are highly discriminant (about 4%) with **ACC**. Note that *UDisc*, when indicated with two metrics, is a directional meta-metrics (*i.e.* the violation instances of $UDisc_{M_1 \rightarrow M_2}$ and $UDisc_{M_2 \rightarrow M_1}$ are different (though the values might be equal) as opposed to symmetric $UCons_{M_1 \leftrightarrow M_2}$ meta-metric.

Table 9 shows the overall results of the Stage-2 benchmarking along with the metrics' ranks. Stage-2 differentiates the positions of the benchmarked metrics, some of which are equal in the previous stages (*e.g.*, MCC and CK have the same ranks). According to overall meta-metrics benchmarking, MCC is ranked the first, whereas nMI and PPV are ranked the last.

Table 9 Experiment-1: Stage-2 benchmarking results for 13 performance metrics according to the seven proposed meta-metrics

Meta-Metrics / Metrics	MCC	BACC	INFORM	MARK	CK	ACC	TNR	TPR	G	F1	NPV	nMI	PPV
UBMcor	1	3	3	3	3	1	9	9	7	8	9	13	9
UIMBucor	1	1	1	1	9	1	1	1	8	11	12	10	12
UDist	5	2	3	3	6	13	8	8	7	8	8	1	8
UOSmo	4	1	1	1	4	13	7	7	6	11	7	12	7
UMono	1	9	9	9	12	1	1	1	1	1	1	13	1
UCons	1	1	1	4	4	7	11	9	6	8	12	13	9
UDisc	2	2	2	2	2	7	7	11	13	7	7	1	11
Overall Stage-2 Rank	1	2	3	4	5	6	7	8	9	10	11	12	12

4 Experiment-1 results and findings

Table 10 summarizes and combines Experiment-1 benchmark results from the two benchmark stages and gives a finalized ranking of the 13 performance metrics reviewed. The stages are weighted according to the complexity and coverage. We set the weights, as shown in Table 10, putting more weight in the meta-metrics stage.

The weights are heuristically determined with more value on Stage-2. Stage-1 summarizes each metric-space or characteristics of metric equations, whereas Stage-2 addresses specific critical robustness issues covering whole metric-spaces. Taking equal weights can lead to rank a metric at the top, even having small meta-metric values. Nevertheless, MCC and BACC are still the first and second robust metric if the weights are equal (the ranks are also the same for the remaining five metrics, namely MARK, TPR, NPV, PPV, and nMI). Taking equal weights diminishes the distinguishing rank between BACC – CK (2nd – 4th) and G – F1 (7th – 8th).

Table 10 The ranking of two benchmark stages and final benchmarking ranking results of Experiment-1

Benchmark Stages	Stage Elements	Weight	MCC	BACC	INFORM	CK	MARK	ACC	G	F1	TNR	TPR	NPV	PPV	nMI
Stage-1: Mathematical evaluation	9 of 11 criteria	1	1	4	4	1	4	8	4	3	9	9	9	9	13
Stage-2: Meta-metrics	Seven meta-metrics	2	1	2	3	5	4	6	9	10	7	8	11	12	12
	Final Benchmarking Ranking:		1	2	3	4	5	6	7	8	8	10	11	12	13

The followings are the main findings:

- MCC is the most robust performance metric.
- BACC is the second-best metric. CK could be interpreted as the third-best because BACC and INFORM are very similar to each other (*UCons* = 1, *UDisc* = 0).
- MCC is also better than CK in other aspects not included in benchmarking (*see* informative criteria in Table 1).
- Highly recommended or conventionally used metrics such as TPR, PPV, ACC, G, and F1 exhibit robustness issues and should be used cautiously if used alone in performance evaluation.
- The recommended nMI metric is also not proper to handle different cases encountered in a classification problem.

Some of the notable observations were obtained from the benchmarking:

In Stage-1:

i) Only **INFORM**, **MARK**, and **BACC** have the same mean, median, and mode values. **ACC** and **MCC** have very close central tendency measures (*see* Fig. 3).
ii) The metrics have a symmetric metric-space except for G, nMI, F1, and CK.
iii) **G** and **F1** metric-spaces exhibit an accumulation at zero.
iv) Only MCC, CK, F1, and nMI cover both outcome (*OP*, *ON*) and class measures (*P*, *N*).
v) TPR, PPV, TNR, and NPV are single-class-only metrics (*i.e. P*-only or *N*-only).
vi) All metrics are insensitive to one or more base measures except nMI, CK, and MCC.
vii) nMI and F1 exhibit some inconsistencies in swapping of base measures.
viii) nMI has a highly right-skewed metric-space.

In Stage-2:

i) **ACC**, **INFORM**, **MARK**, **BACC**, and **MCC** have a high correlation with individual base measures, whereas the others have either some imbalances or no associations.
ii) **nMI** does not exhibit any relationship with base measures.
iii) PPV, NPV, and F1 are moderately; G, CK, and nMI are slightly correlated with class imbalance (*i.e.* changing class ratios either increase or decrease the performance without changing the model).
iv) TPR, TNR, PPV, NPV, ACC, and F1 do not exhibit granular output coverage in metric-spaces (*see* Fig. 5).
v) nMI and ACC do not output smoothly in metric-spaces.
vi) All metrics are monotonic except INFORM, MARK, BACC, CK, and nMI. CK has a minor, and nMI has numerous monotonicity violations.
vii) BACC, INFORM, and MCC are the most consistent metrics among all the metrics.
viii) INFORM and BACC are the only metrics that are entirely consistent with each other.
ix) nMI is the least consistent and the most discriminating metric.
x) G is the least discriminant metric.

The robustness issues affect performance evaluation of multi-class classification. For example, macro-averaged F1 (shortly macro-F1) and micro-averaged F1 (micro-F1) are based on F1 with 8th robustness ranking. Macro-F1 is the arithmetic mean of F1s calculated for each class, whereas micro-F1 is the additive calculation of F1 by counting the total number of true class, false all-the-other class, and false class per each class, which are similar to binary *TP*, *FN*, and *FP*, respectively. These metrics commonly used in multi-class classification are apart from other metrics specific to multi-class, such as Hamming loss [29].

5 Evaluation of the proposed benchmarking method with the literature

We compared our benchmarking method with the other methods in the literature threefold. We first compared our methodology with the existing metrics evaluation methods. In the second step, we compared the evaluation strategies of the studies proposing new performance metrics. Finally, we re-tested BenchMetrics on recently proposed metrics and compared our results with their findings.

5.1 Comparison of BenchMetrics with the existing metrics evaluation methods

Table 11 gives details about the methods designed for metric comparisons in the literature, summarizes their limitations, and compares them with our benchmarking method. First of all, the compared studies examined only a few metrics. Second, while some focused on basic behaviors of performance metrics (*e.g.*, extreme cases such as comparing two classifiers' results with swapped confusion matrix), the other studies worked on experimental classifications corresponding to only a minimal part of metric-spaces and showed similarities from a pure perspective without using an explicit ranking.

Table 11 Comparison of our benchmarking method with the existing metrics evaluation methods

Compared Metrics, Year, Study,	Conclusion	Evaluation Method	Notes: Limitation of the Studies and the Comparison Results (**OCC** stands for "Our Corresponding Criteria")
ACC and AUCROC, 2005, [14]	AUCROC is recommended instead of ACC.	The simulated classifiers' performances applied on balanced and imbalanced synthetic datasets and three classifiers applied on 18 real-world datasets (with 61 <= *Sn* <= 8.124) are calculated in terms of ACC and AUCROC, and each paired metric value are compared for consistency and discriminancy.	**1)** Assessing the consistency and discriminancy among the compared metrics does not impose a superiority, especially in paired comparisons. For example, consistency between CK and ACC is meaningful only if both of the metrics are robust. Likewise, if either one or both metrics are not robust, then the discriminancies could not be interpreted. **2)** Our benchmark includes a large number of metrics; thus, the conclusions are more comprehensive. **3)** Our benchmark also indicates that CK is better than ACC, in line with this study. **OCC:** *UCons*/*UDisc* A similar method is used to measure the consistency and discriminancy among the metrics.
CK and ACC, 2008, [30]	CK is recommended instead of ACC.	The consistency and discriminancy are compared within "the desired region of operation" only (*i.e.* where *TPR* >= 0.5 and *FPR* <= 0.02). This is because the calculation of consistency and discriminancy degree has time and calculation costs, as defined in the above study.	
BACC, ACC, F1, TNR, TPR, and PPV, 2006, [4]	TNR and BACC are more appropriate metrics concerning the variance or invariance of changes in confusion matrix elements.	Checking whether the performance output depends on the following changes in confusion matrix: 1) exchange *TP* with *TN* and *FN* with *FP* 2) change only in *TN* 3) change only in *FP*, and 4) scale *TP* and *FP* along with *TN* and *FN*.	**4)** We reformulate those four changes to fit the classification performance evaluation context and make the assessments more comprehensible. **5)** Our benchmark shows that MCC and CK are the most robust metrics from the corresponding three criteria. But, TNR and BACC have the same inconsistencies with TPR, PPV, and ACC. **OCC:** 1) Invariant to class-and-outcome swaps (Criterion-6); 2 & 3) Base measure coverage (Criterion-3); 4) Outcome/class coverage (Criterion-1)
ACC, G, F1, FPR, FNR, NPV, PPV, AUCROC, and AUC-PR, 2009, [31]	Instead of comparing and rankings metrics, the study groups the compared metrics into two to four similar groups.	The performances of two decision tree classifiers applied on 35 real-world datasets with 200 <= *Sn* <= 20.000 and 65% < *PREV* < 99% based on different decision thresholds (0 < *t* < 1, default: 0.5) are calculated in terms of the compared nine metrics. The relations of the metric values are compared for 350 classifier-dataset runs in total: Comparison-1: Via correlations; Comparison-2: Via factor analysis (analyzing correlated metric values (observed variables) in terms of a small number of factors (unobserved variables).	**6)** Examining performance metrics based on a number of real-world datasets covers limited prevalence and metric-space cases. For example, in our benchmarking, there are 2,667,126 base-measure permutations for *Sn* = 250. Whereas, for example, the 350 cases correspond to only 0.013 % of all possible permutations. Thus, correlations and factors may not be representative. **7)** The comparisons simply show similar metrics that are redundant when they are used together. They do not sufficiently dictate a proper metric and do not reveal any robustness issues. For example, G and F1 are found similar in factor analysis, whereas in our benchmarking, G is slightly more robust in general than F1. **8)** The comparisons are limited as they are reliant on the performance of two decision tree classifiers. **OCC:** *UBMcor*, *UIMBucor*, *UMono*, *UCons*/*UDisc*
INFORM, ACC, G, and F1, 2002, [32]	F1 is the recommended metric.	They constructed performance trend graphics for different TPR, PPV, and PREV variations and observed whether the performances increase according to PREV.	**9)** Both techniques require visual inspection and manual interpretation and are not measurable as in our benchmark. **10)** For the former study, our benchmark shows that INFORM is better among the compared four metrics. **11)** For the latter study, MCC is more robust and in line with our benchmark, whereas F1 has robustness issues in our corresponding criteria. **OCC:** *UIMBucor*, *UCons* (with TPR and PPV), and *UCons* (with TPR and TNR)
MCC, BACC, ACC, F1, TNR, and PPV, 2018, [3]	MCC and F1 exhibit a more "realistic" estimation of classification performance.	They constructed performance trends graphics for different TPR and TNR variations and inverse cumulative distribution function plots per balanced and imbalanced datasets.	

Nevertheless, we addressed the proposed comparison techniques in BenchMetrics in a formal and easy to understand manner with measurable and comparable outputs. We improved the existing metric comparison approaches either by extending them or defining them in a classification performance context.

5.2 Comparison of BenchMetrics with the methods which were used to evaluate new metrics

Table 12 shows the recently proposed performance metrics and how they were compared with the other existing literature metrics.

Table 12 Comparison of our approach with the methods which were used to evaluate new metrics

The Study, Proposed New Metric, and its Description	Notes and Validation of the New Metric / Our Corresponding Criteria (**OCC**)
[33] SAR (an abbreviation of Squared error, Accuracy, and ROC area) $SAR = \frac{ACC + AUCROC + (1 - RMS)}{3}$ *SAR* combines Accuracy, Area Under ROC Curve, and Squared Error into one measure.	AUCROC and RMS (root mean square) are different from all the metrics summarizing base measures like ACC. RMS is for regression problems instead of classification. The proposed metric is validated via correlation analysis. **OCC:** *UCons*/*UDisc*$^{(a)}$
[5] Optimized Precision (OACC): $OACC = ACC - \frac{\lvert TPR - TNR \rvert}{TPR + TNR}$ OACC reduces the sub-optimal performance measurement of ACC due to the skewed datasets by adding a heuristic correcting factor that minimizes TPR and TNR difference while maximizing their totals.	The proposed metric is validated by comparing ACC and OACC outputs with class balanced and highly-imbalanced synthetic datasets (*SKEW*s are 1:1 and 1:9) along with a single real-world dataset (human DNA sequences). They inspected graphics showing the metrics' variance for theoretical *TPR* and *TNR* ranges using $ACC = TPR \cdot N + TNR \cdot P$ equations. *See* "note 9" in Table 11 for comparison. **OCC:** *UIMBucor*
[34] AUCROC:ACC $AUCROC{:}ACC = \begin{cases} AUCROC, & AUC\text{-}ROC \text{ pairs are different} \\ ACC, & \text{pairs are the same} \end{cases}$ AUCROC:ACC is a two-staged measure to enhance metric output differentiation.	The proposed metric is validated by examining the new metric's correlations with AUCROC and ACC separately, then comparing it with the best RMS values (AUCROC:ACC is positively correlated with RMS). *See* "note 1" in Table 11 for comparison. **OCC:** *UCons*/*UDisc*$^{(a)}$
[9] Standardized Relative Performance Metric (*SRPM*) $SRPM_i = \frac{RPM_i - \min(RPM_i)}{\max(RPM_i) - \min(RPM_i)}$ $RPM_i = \sum_{i=0}^{n} p_j^* f_{ij}$ where n: number of factors, f_{ij} is the factor score of the i^{th} instance on the j^{th} factor, and p^*_j is the normalized proportion of the variance to the j^{th} factor.	Performances are calculated in different metrics (ACC, G, F1, NPV, PPV, AUCROC, AUCPR: Area-Under-Precision-Recall-Curve) for 12 ML models on 35 real datasets. Factor analysis is applied to the metric values. A relative metric value calculated with factor scores and normalized proportions of the eigenvalues is standardized into [0, 1] range for the given number of factors. The choice of the number of factors requires expertise. As an empirical study, their work results are limited to the scope of the chosen classifiers and performance metrics. **OCC:** *UCons*/*UDisc*$^{(a)}$
[6] Index of Balanced Accuracy: $IBA_\alpha(G) = \big(1 + \alpha(TPR - TNR)\big)G$ IBA is a parametric metric like OACC that adjusts a known metric (here G), taking the difference between TPR and TNR.	1. The correlations of the new metric with TPR, TNR, ACC, and G are evaluated concerning class imbalance (ACC and G), and class focuses (TPR and TNR) 2. Invariance properties are used to evaluate the metric. *See* "notes 4 and 7" in Table 11 for comparison. **OCC:** *UCons*/*UDisc*$^{(a)}$, Criterion-6, Criterion-3, and Criterion-1

(a) The *UCons* and *UDisc* meta-metrics can be used to see the similarities and differences of the metrics comprising the composite new metrics. For example, *UCons* and *UDisc* meta-metric values for TPR, TNR, and G metrics can give the following insights in a composite new metric IBA$_a$(G): TPR – TNR relations: I) $UCons_{TPR \leftrightarrow TNR} = 0.53$, II) $UDisc_{TPR \to TNR} = UDisc_{TNR \to TPR} = 0.29$
G – TPR and TNR relations: III) $UCons_{G \leftrightarrow TPR} = UCons_{G \leftrightarrow TNR} = 0.77$, IV) $UDisc_{G \to TPR} = UDisc_{G \to TNR} = 0.19$, V) $UDisc_{TPR \to G} = UDisc_{TNR \to G} = 0.29$. Because TPR and TNR together are a factor of IBA$_a$(G) (*i.e. TPR – TNR*), we expect that the directional discriminancy meta-metrics between TPR and TNR (relation II) as well as between G and each of TPR/TNR (relations IV and V) the same. Otherwise, two contradicting terms are to be subtracting. Having consistency meta-metrics that are not close to 1 (relations I and III) can be interpreted as the composite metric differentiates from the dependent metrics (if the consistencies are close to 1, then there is no need to define another metric).

The first three of the proposed metrics are intended to minimize the class imbalance effect of ACC.

5.3 Experiment-2 results and findings

In Experiment-2, we tested BenchMetrics on 13 performance metrics with additional two recently proposed metrics namely Optimized Precision (OACC) and Index of Balanced Accuracy for G (IBA$_a$(G)). To avoid repeated information presented for Experiment-1, we specifically focused on:

- whether proposed metrics provide an improvement compared to their base metrics (ACC for OACC and G for IBA$_a$(G)) and
- comparison with MCC as the most robust metric in Experiment-1.

Table 13 lists the details of the Stage-1 benchmarking results of Experiment-2 (like Table 1 of Experiment-1 presented for the benchmarking of 13 performance metrics). The various positive or negative robustness issues (underlined bold texts depict negative ones) are revealed. Note that a coefficient in $IBA_{\alpha}(G) = \big(1 + \alpha(TPR - TNR)\big)$ is taken as 0.05 as suggested by [6].

Table 13 Experiment-2: Benchmarking Stage-1 results (*Sn* = 50) for the two new proposed metrics.

Stage-1 Criteria	ACC	OACC	G	IBA$_a$(G)	MCC
1 Outcome/class coverage	**None**	**Class-only**[1]	**Class-only**	**Class-only**[2]	Yes
2 Class coverage (*P* and *N*)	**None**	Yes[1]	Yes	Yes[2]	Yes
3 Base Measure Coverage	***TP, TN***	***TP, TN***	***TP, TN***	***TP, TN***	Yes
4 Variant to class swap	Yes	Yes	Yes	Yes	Yes
5 Variant to outcome swap	Yes	Yes	Yes	Yes	Yes
6 Invariant to class-and-outcome swaps	Yes	Yes	Yes	**No**	Yes
7 Undefined (NaN) count	0	**3*Sn*+1**	**2(*Sn*+1)**	**2(*Sn*+1)**	**4*Sn***
8 Central tendencies (mean-median difference)	$\bar{\mathbf{M}} = \tilde{\mathbf{M}} \approx \mathbf{Mo}$	$\underline{\bar{\mathbf{M}} \neq \tilde{\mathbf{M}}} \neq \mathbf{Mo}$	$\bar{\mathbf{M}} \approx \tilde{\mathbf{M}} \neq \mathbf{Mo}$	$\bar{\mathbf{M}} \approx \tilde{\mathbf{M}} \neq \mathbf{Mo}$	$\bar{\mathbf{M}} \approx \tilde{\mathbf{M}} = \mathbf{Mo}$
Other Informative Criteria					
9 Standard Deviation	0.23	0.23	0.26	0.26	0.21
10 Skewness	Symmetric	Slightly negative[3,4]	Slightly positive[5]	Slightly positive[5]	Symmetric
11 Kurtosis	Platykurtic[6]	Platykurtic[6]	Platykurtic[6]	Platykurtic[6]	Platykurtic[6]

(1) *OACC = f(TP, TN, P, N, TC, Sn)*, (2) *IBA$_a$(G) = f(TP, TN, P, N)*, (3) Left-skewed, (4) Distorting symmetry, (5) Right-skewed, (6) Thin-tailed

The followings are the summary of the findings according to focus expressed above:

i. OACC improved ACC on outcome/class and class coverages, but robustness issues appeared in undefined metric outputs and mean-median difference. It also distorts symmetry observed in ACC.
ii. IBA$_a$(G) has no improvement on G; it is not invariant in class-and-outcome swaps, which is only seen in F1 in the benchmarked metrics, as seen in Table 1.
iii. Evaluating the eight criteria in Stage-1, the robustness of OACC and IBA$_a$(G) is almost identical. Only Criterion-6 (invariant to class-and-outcome swaps) and Criterion-8 (central tendencies) are different mutually.
iv. MCC is more robust than the new metrics.

Table 14 shows the results of the Stage-2 benchmark of Experiment-2 according to the first five meta-metrics. Up arrows depict that a new metric improves the dependent metric (*i.e.* IBA$_a$(G) improves G or OACC improves ACC). Down arrows represent degradation. The following is a summary of the findings:

i. OACC improves ACC on distinctness and output smoothness but decreases the robustness for base measure correlations, imbalance uncorrelation, and monotonicity in a contradictory manner.
ii. IBA$_a$(G) improved G by increasing distinctness and output smoothness (no significant improvement for imbalance uncorrelation).

iii. IBA$_a$(G) is more robust than OACC, considering the base measure correlations, distinctness, and monotonicity.

iv. MCC is more robust than the new metrics as in Stage-1.

Table 15 lists the remaining meta-metrics in Stage-2, namely *UCons* and *UDisc*. We summarized them per each recently proposed metric instead of giving each pairwise meta-metric values among the metrics as in Table 7 and Table 8. Bold values depict higher meta-metric summary values. For example, the mean consistency of IBA$_a$(G) with the 13 benchmarked metrics (0.834) is higher than the mean consistency of ACC (0.773).

Table 14 Experiment-2: Benchmarking Stage-2 results (*Sn* = 50) for the two new proposed metrics in the literature (excluding the *UCons* and *UDisc* meta-metrics). Metrics are sorted in descending order per meta-metrics from the most robust one to the least. *Osmo* is the smoothness value.

UBMcor		*UIMBucor*		*UDist*		*UMono*		*Osmo*	
MCC	**0.78**	**MCC**	**1**	**IBA$_\alpha$(G) ▲**	**0.8**	**MCC**	**1**	**INFORM**	3.22
ACC	**0.78**	**INFORM**	1	**OACC ▲**	**0.412**	**ACC**	**1**	**MARK**	3.22
INFORM	0.77	**MARK**	1	**nMI**	0.382	**G**	**1**	**BACC**	3.22
MARK	0.77	**BACC**	1	**BACC**	0.333	**IBA$_\alpha$(G)**	1	**OACC ▲**	**4.91**
BACC	0.77	**ACC**	**1**	**INFORM**	0.332	**F1**	1	**MCC**	**5.26**
CK	0.77	**TPR**	1	**MARK**	0.332	**TPR**	1	**CK**	5.28
G	**0.75**	**TNR**	1	**MCC**	**0.232**	**TNR**	1	**IBA$_\alpha$(G) ▲**	**6.44**
IBA$_\alpha$(G)	**0.75**	**IBA$_\alpha$(G) ≈**	**0.98**	**CK**	0.202	**PPV**	1	**G**	**6.98**
OACC ▼	**0.73**	**G**	**0.97**	**G**	**0.196**	**NPV**	1	**TPR**	7.82
F1	0.72	**OACC ▼**	**0.97**	**TPR**	0.033	**INFORM**	0.998	**TNR**	7.82
TPR	0.69	**CK**	0.96	**TNR**	0.033	**MARK**	0.998	**PPV**	7.82
PPV	0.69	**nMI**	0.91	**PPV**	0.033	**BACC**	0.998	**NPV**	7.82
TNR	0.69	**F1**	0.64	**NPV**	0.033	**CK**	0.948	**F1**	9.15
NPV	0.69	**PPV**	0.55	**F1**	0.033	**OACC ▼**	0.76	**nMI**	19.7
nMI	0.5	**NPV**	0.55	**ACC**	**0.002**	**nMI**	0.517	**ACC**	**21.62**

Table 15 Summary of the pairwise *UCons* and *UDisc* meta-metrics in Experiment-2 per OACC and IBA$_a$(G) with the 13 benchmarked metrics (minimum, mean, standard deviation (SD), and maximum values) for Stage-2 with *Sn* = 20

New Metric(s) *vs.* Reviewed Metrics	Consistencies / Discriminancies with each reviewed metric	Min	Mean ± SD	Max
OACC	$\{UCons_{\mathrm{OACC}\leftrightarrow\mathrm{M1,\,\ldots,\,M13}}\}$	0.511	0.773 ± 0.090	0.899
	$\{UDisc_{\mathrm{OACC}\leftrightarrow\mathrm{M1,\,\ldots,\,M13}}\}$	0.002	**0.022 ± 0.020**	**0.052**
	$\{UDisc_{\mathrm{M1,\,\ldots,\,M13}\leftrightarrow\mathrm{OACC}}\}$	0	**0.003 ± 0.001**	**0.004**
IBA$_a$(G)	$\{UCons_{\mathrm{IBA\alpha G}\rightarrow\mathrm{M1,\,\ldots,\,M13}}\}$	**0.551**	**0.834 ± 0.110**	**0.992**
	$\{UDisc_{\mathrm{IBA\alpha G}\rightarrow\mathrm{M1,\,\ldots,\,M13}}\}$	0.002	0.014 ± 0.020	0.051
	$\{UDisc_{\mathrm{M1,\,\ldots,\,M13}\rightarrow\mathrm{IBA\alpha G}}\}$	0	0.042 ± 0.014	0.053

OACC vs. IBA$_a$(G)	Consistency / Discrimininancy between two new metrics	Meta-metric value
	$UCons_{\mathrm{IBA\alpha G}\leftrightarrow\mathrm{OACC}}$	0.898
	$UDisc_{\mathrm{IBA\alpha G}\rightarrow\mathrm{OACC}}$	0.001
	$UDisc_{\mathrm{OACC}\rightarrow\mathrm{IBA\alpha G}}$	**0.046**

Among the paired metric values in metric-space, OACC and IBA$_a$(G) are 89.8% consistent. However, IBA$_a$(G) is more consistent with the 13 benchmarked metrics on average, whereas OACC is more discriminant than both the benchmarked metrics (2.2%) and IBA$_a$(G) (4.6%). Briefly, IBA$_a$(G) is more consistent, and OACC is more discriminant.

Combining Stage-1 and Stage-2, IBA$_a$(G) is more robust than OACC. However, neither of them is as robust as MCC. Both Experiment-1 and Experiment-2 show that BenchMetrics can be used to analyze and compare the robustness of any proposed performance metrics derived from a confusion matrix for single-threshold ML models or crisp binary-classifiers. The findings of BenchMetrics for the metrics comprising graphical performance metrics are directly valid, namely *TPR*s *vs. TNR*s for *AUCROC* and *TPR*s *vs. PPV*s for *AUCPR*, respectively. Section 7 discusses the robustness of the instruments based on the probabilistic interpretation of classification error or loss.

6 Evaluation of meta-metrics

This section demonstrates the usability of meta-metrics in assessing the robustness of metrics via an approach different from those evaluated in Table 11 and Table 12, such as comparing metric values of some ML classifiers tested on real-world datasets and manual analysis of metric graphics. This approach is to analyze and compare metric values calculated for the controlled synthetic classifiers. Evaluating metrics via ML classifiers such as Support Vector Machines or Decision Trees is not straightforward or conclusive. They bring extra factors to consider, such as the randomness of models, dataset dependency, different tuning of parameters, etc. Synthetic classifiers are consistent and controlled. Therefore, they successfully reveal the pure behavior of the metrics.

To the best of our knowledge, only Boughorbel et al. [7] used synthetic classifiers to compare ACC, F1, MCC, and AUCROC. They examined the classifiers defined for $\boldsymbol{PREV} < 0.5$ values only and interpreted the relationship between metrics and prevalence via the graphics. In this evaluation, we tested BenchMetrics to evaluate the metrics on those synthetic classifiers to see whether the meta-metrics give insights about the metrics' robustness without manual interpretation.

Extending Boughorbel's classifiers, we covered the full ***PREV*** range and clearly defined the metric-class-imbalance relation via the proposed *UIMBucor* meta-metric. The specifications of the synthetic classifiers are as follows:

- *SC-1 (Stratified Random):* Stratified random classifier makes a random prediction (*i.e.* independent from the input instance) by taking into consideration the test dataset's class distribution (*i.e.* the probability of predicting as positive is *PREV*).
- *SC-2 (Random):* Simple random classifier makes a random prediction independent from the test dataset's class distribution (*i.e.* the probability of predicting as positive is 0.5 that is independent of *PREV*).

Fig. 6 shows each synthetic classifier's performances in terms of thirteen metrics calculated for 41 synthetic datasets with a sample size of 10,000. The datasets' class ratios are generated per *PREV* level from 0 to 1 in 0.025 increments, making various binary examples with known labels. For example, the first dataset ($\boldsymbol{PREV}_1 = 0$) consists of negative examples only (10,000 negatives); the second one ($\boldsymbol{PREV}_2 = 0.025$) has 250 positives and 9,750 negatives, and so on.

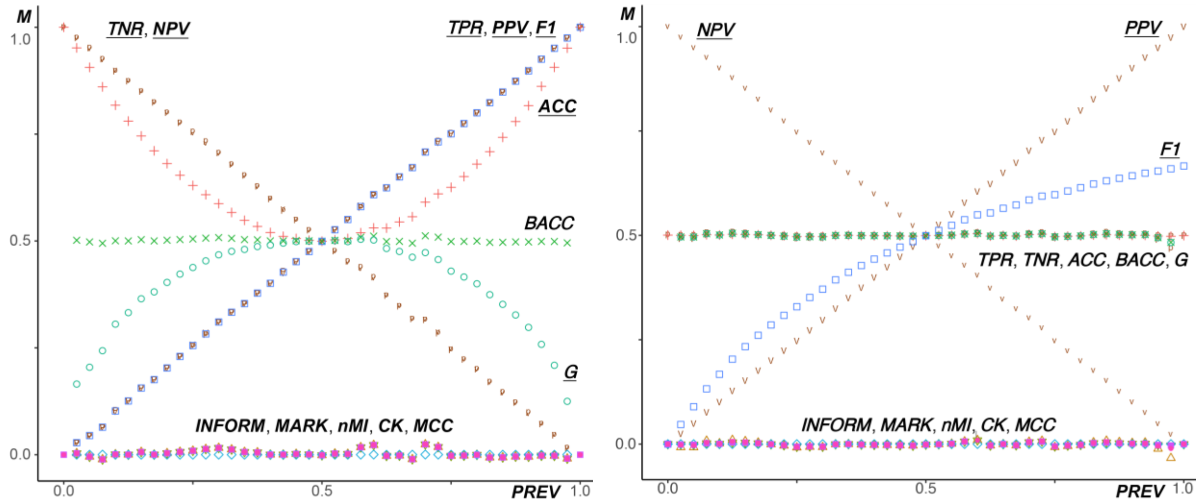
Each synthetic classifier predicts each example in the datasets according to the given specification above. The sum of the results of those predictions yields the classifiers' performance per dataset in terms of *TP*, *FP*, *FN*, and *TN* that makes an array of 41 base measures (***BM***, not a base-measure permutation: $\mathbf{BM}^{Sn}$). The performances are calculated per synthetic classifier for 41 datasets in terms of thirteen metrics (yielding the corresponding array of metric values, *e.g.*, ***ACC***, not a metric-space).

As shown in *UIMBucor* values (the bold-underlined ones: F1, NPV, PPV, TNR, and TPR in SC-1 and F1, PPV, and NPV in SC-2) in Fig. 6 (c) and (d), *UIMBucor* meta-metric successfully identifies the metrics directly affected by class imbalance that is also observed in the graphs. Note that the other six meta-metrics are not included in this evaluation because they show the entire metric-space aggregated statistics.

The overall results of Stage-2 BenchMetrics (including all seven meta-metrics) tested on SC-1 and SC-2 also support MCC's robustness. This evaluation shows that meta-metrics does not require manual analysis and give measurable insights even within a small section of permutations in overall metric-spaces.

7 Robustness of graphical and probabilistic classification error/loss instruments

As described above, BenchMetrics is proposed to assess the metrics derived from a confusion matrix for single-threshold ML models or crisp binary-classifiers. Although graphical performance metrics are not based on a single instance of a confusion matrix, they are calculated by varying a decision threshold (*i.e.* full operating range of a classifier) for different metric pairs in a specific binary-classification application [35]. Therefore, BenchMetrics is also valuable for graphical performance metrics by assessing robustness through the dependent metrics (TPRs *vs*. TNRs for AUCROC and TPRs *vs*. PPVs for AUCPR).



(a) SC-1: Stratified random performance metric values

(b) SC-2: Random performance metric values

Metrics	*UIMBucor*
BACC	0.89
INFORM	0.89
MARK	0.89
nMI	0.64
CK	0.80
MCC	0.80
G	**0**
TNR	**0**
TPR	**0**
ACC	**0**
F1	**0**
NPV	**0**
PPV	**0**

(c) *UIMBucor* for SC-1

Metrics	*UIMBucor*
nMI	0.84
TPR	0.89
BACC	0.85
G	0.85
INFORM	0.85
MCC	0.64
CK	0.48
MARK	0.48
TNR	0.43
ACC	0.38
F1	**0**
NPV	**0**
PPV	**0**

(d) *UIMBucor* for SC-2

Fig. 6 Classification performance metrics' trends for the synthetic classifiers on datasets with varying *PREV* and corresponding *UIMBucor* meta-metric value distributions.

8 Probabilistic error/loss instruments and not-applicability of BenchMetrics

Like graphical-based instruments, probabilistic error/loss instruments, such as MSE or LogLoss, do not depend on a confusion matrix. They summarize the deviation from the true probability or prediction uncertainty. Although they are preferred in regression problems rather than classification problems or chosen for multi-class classification rather than binary, they can be reported in binary classification (*e.g.*, neural network classification models), usually with one or more confusion matrix based metrics, as a "reliability metric" [8, 35] instead of a "performance metric".

Contrary to zero-one loss metrics (*e.g.*, *MCR*, *FPR*, *FNR*, *FDR*, and *FOR*), probabilistic error/loss instruments evaluate the performance error of scoring or non-crisp classifiers that label instances with a reported or attached belief value (score, probability or likelihood) according to a decision boundary. For example, instead of labeling an instance as positive (one) or negative (zero) absolutely (also known as "hard label"), a classifier model with a $\Theta = 0.50$ internal decision-boundary value (the right side is for positive labels, the left side is for negative ones) in [0, 1] interval can label a positive instance ($c_i = 1$) as positive correctly with a $p_i = 0.85$ score (also known as "soft label"). It labels another instance ($c_j = 0$) as negative correctly with a $p_j = 0.40$ score. Hence, we can interpret the probabilistic classification error as a distance function for those instances such that the former labeling is more probable than the latter ($|0.85 - 0.50| = 0.35 > 0.10 = |0.40 - 0.50|$ where $\Theta = 0.5$).

BenchMetrics could not be applied to assess probabilistic performance instruments' robustness because it mainly analyzes the performance metric-spaces representing all the possible base-measure permutations. One permutation would be $TP = 10$, $FP = 0$, $FN = 0$, $TN = 10$ for $Sn = 20$. In this best-case scenario, a probabilistic error/loss metric, such as *MAE*, could be any value in [0, 0.5)[3]. In other permutations, the possible range is in [0, 1]. Because there is no relation between classification results in terms of base measures with any error/loss values, BenchMetrics' methods on probabilistic instruments adding extra dimension to the metric-space are not applicable.

9 Discussion and implications

The base-measure permutations provide a pseudo-universal metric-space for analyzing and comparing metrics' outcomes without missing any case contrary to the studies in the literature, investigating limited cases. The permutation vector and metric-space size increase exponentially with the classification dataset size (286 permutations for $Sn = 10$ and 2,667,126 for $Sn = 250$). Increasing *Sn* only increases the granularity of the transitions among permutations (the minimum change of base measures is always ±1). The distribution and descriptive statistics remain almost the same, as shown in metric-space density graphs in Fig. 3. *UBMcor* and *UIMBucor* benchmarking meta-metrics are not affected by *Sn*. For the remaining five benchmarking meta-metrics, yielding different but consistent values for other *Sn*, we averaged the values for nine *Sn* values between 25 and 250, which corresponds to 3,276 and 2,667,126 permutations, to calculate the final meta-metric value. Although the granularities exhibited by those permutations are sufficient to conduct the proposed analysis, one practical barrier was the long calculation time of *UCons* and *UDisc* meta-metrics (about 22 hours to calculate a metric with twelve other metrics). These meta-metrics compare every pair of the reviewed metrics. The time constraint is also valid for creating the base-measure permutations. However, we optimized the basic algorithm to generate the base-measures according to our Definition 1. Hence, the improvement of generating and calculating base-measure permutations and metric-spaces and calculating meta-metrics for larger sizes or conducting the benchmark for larger sizes, possibly in a high computation power platform is a future research.

Contradictory, it could be argued that the benchmarking highlights subtle issues in some metrics (*e.g.*, monotonicity violations in CK) that cannot be seen in practice or a well-prepared classification study. In our opinion, the issues re-summarized in Section 4 cannot be ignored as they may arise in several areas such as online machine learning classifications, decision-making applications, including "what if" scenarios, and artificial general intelligence in the future where the classification performance possibilities are diverse.

The theoretical implications of this study are (1) to define objective and measurable criteria for metrics, (2) to establish a well-founded testing methodology for their comparison, and the practical implications are (1) to increase awareness regarding the problematic metrics and their specific weaknesses, and (2) to support researchers in terms of selecting a robust metrics for their classification application domain in addition to the conventional ones used in the field when necessary. A further implication of this transition into robust metrics is the possibility that different ML approaches in the literature may come to prominence in specific domains requiring using those metrics as a standard practice.

BenchMetrics is designed to produce scaled scores for ranking and comparison and implemented as an API (Application Programming Interface) to prepare the related data, calculate the scores, and provide the test results. Hence, other researchers can extend it by including new criteria. Additional statistical analysis, for example, could contribute to revealing informative characteristics of the distributions (besides "skewness" and "kurtosis" in

[3] For ten negative samples (*e.g.*, $i = 1, \ldots, 10$): $c_i = 0$ and example $p_i = 0.49$ then $| c_i - p_i | = 0.49$. For remaining ten positive samples (*e.g.*, $i = 11, \ldots, 20$) : $c_i = 1$ and example $p_i = 0.51$ then $| c_i - p_i | = 0.49$. Hence, $MAE = 0.49$.

Section 3.2.3) such as modality of the probability distributions (*i.e.* multimodal distributions having two or more modes or peaks) [36].

As an up-to-date improvement in scientific studies, initiatives such as OpenAIRE by the European Union and Zenodo by CERN aim to encourage common, responsible, and reproducible open research approaches where research data become available to all researchers. Inline with their objectives, this study aims to develop a common standard for evaluating performance since scientific progress cannot be achieved unless objective comparison methodologies are determined clearly and followed by all researchers.

We expect that the metrics' rankings and their robustness issues revealed will guide researchers to evaluate classification performances straightforwardly by choosing the right metrics and further contributing toward responsible and reproducible open research by establishing the common best practices in performance evaluation and reporting. As several ML studies are conducted, and the industry starts to develop and use the classifiers practically in many areas, using precise and concise instruments would become a precondition for evaluating performances and claiming an improvement. Finally, we plan to generalize the proposed BenchMetrics method into multi-class performance evaluation metrics as a future study.

10 Conclusions

This study examined binary-classification performance metrics' behaviors from a broader perspective to reveal the problematic issues and showed the most robust metric. A new comprehensive method called BenchMetrics is proposed to evaluate and compare performance metrics. Contrary to the existing approaches, our method comprising two stages presents new concepts with formal definitions such as metric-space, meta-metrics (metrics about performance metrics), base-measure permutations, and variance/invariance in base-measure swapping to analyze the metrics. BenchMetrics spots the weak and robust issues of individual metrics, metric pairs, and a group of metrics in an objective and comparable manner. BenchMetrics Stage-1 comprises eight criteria to evaluate the metrics from a mathematical perspective, and Stage-2 presents seven novel meta-metrics to analyze a metric-space based on different robustness requirements. All criteria and meta-metrics yield a robustness value normalized into [0, 1] interval to enable comparison of metrics per criteria or meta-metric as well as provide a ranking for the metrics per stage or an overall final benchmarking.

BenchMetrics was tested on thirteen performance metrics that are commonly used or recommended in the literature (Experiment-1). To the best of our knowledge, this is the first time that such a wide range of metrics have been reviewed in this scope, and one metric has been suggested with substantial justification. BenchMetrics spotted specific cases where a metric behaves unexpectedly (*e.g.*, yielding high-performance values in a higher number of false classifications or metric-space distribution anomalies clearly demonstrated in Fig. 5). For instance, TPR, ACC, nMI, F1, and CK exhibit significant robustness issues, which need to be taken into consideration if they needed to be used in an experiment.

Lack of a systematic benchmarking methodology specifically brought a limitation for the researchers who propose new performance metrics. The previous techniques reveal only limited aspects of performance metrics, which have been already addressed with BenchMetrics. In this regard, we re-tested two recently proposed performance metrics, namely Optimized Precision (OACC) and Index of Balanced Accuracy (IBAα(G)), along with the previously benchmarked thirteen metrics (Experiment-2). This experiment showed limited improvements in these new metrics and conversely, introduced robustness issues unaddressed in the literature. Experiment-2 benchmarking showed that MCC is still the most robust metric, including the recent ones. Future new metric propositions should exhibit more robustness from MCC when tested by BenchMetrics. MCC has been previously shown superior to numerous metrics in the literature [37]. Reporting MCC with conventional metrics (ACC, TPR, PPV, and F1) in recent research in some fields such as biology and bioinformatics (*e.g.*, diagnose cancer diseases via ML-based classification) [38, 39] can be an indication of the trust to MCC as a robust metric. BenchMetrics offers a systematic and well-founded approach for metric choice in specific domains instead of heuristic reasoning.

Finally, we demonstrated the effectiveness of meta-metrics via synthetic classifiers. Our method produces similar results mathematically to those synthetic classifiers which require visual interpretation on plots where marginal differences cannot be observed properly. Note that the developed online benchmarking open-source software library and an interactive platform to run BenchMetrics with example data/graphs are provided for the researchers who wish to see the details or conduct their benchmarks.

In conclusion, this study proposes a new comprehensive benchmarking method to analyze the robustness of performance metrics and ranks 15 performance metrics in the literature. Researchers can use MCC as the most robust metric for general objective purposes to be on the safe side. Otherwise, they can select a metric among others required or enforced by their domain of interest, considering the ranks and specific robustness issues revealed by the benchmark.

Appendix A Developed online research tools and data

- An online interactive BenchMetrics experimentation platform

  Platform: Code Ocean, **Address:** https://doi.org/10.24433/CO.1564477.v3
- BenchMetrics open-source performance metrics benchmarking software library (API)

  Repository: GitHub, **Address:** https://github.com/gurol/benchmetrics
- Binary-classification performance metric-spaces data

  Repository: Mendeley Data, **Address:** http://dx.doi.org/10.17632/64r4jr8c88.2
- Binary-classification performance-metrics benchmarking data

  Repository: Mendeley Data, **Address:** http://dx.doi.org/10.17632/2g36672s5f.4

Appendix B Binary-classification performance instrument list

Table 16 Performance measures and metrics (names, alternative names, abbreviations, and equations)

Measure — Abbreviation		Equation
True Positive	*TP*	
False Positive	*FP*	
False Negative	*FN*	
True Negative	*TN*	
Positive	*P*	$TP + FN$
Negative	*N*	$TN + FP$
Outcome Positive	*OP*	$TP + FP$
Outcome Negative	*ON*	$TN + FN$
True Classification	*TC*	$TP + TN$
False Classification	*FC*	$FP + FN$
Sample Size	*Sn*	$P + N = OP + ON$ $= TC + FC$ $= TP + FP + FN + TN$
Prevalence	*PREV*	P/Sn
Bias	*BIAS*	OP/Sn
Cohen’s Kappa Chance	*CKc*	$\frac{P \cdot OP + N \cdot ON}{Sn^2} =$ $PREV \cdot BIAS +$ $(1 - PREV) \cdot (1 - BIAS)$
Determinant	*DET*	$TP \cdot FN - FP \cdot FN$
Class Entropy	*HC*	$-\sum_{c=P,N} \frac{c}{Sn} log_2 \frac{c}{Sn}$ $= -\sum_{m=PREV,1-PREV} m\, log_2\, m$
Outcome Entropy	*HO*	$-\sum_{o=OP,ON} \frac{o}{Sn} log_2 \frac{o}{Sn}$ $-\sum_{m=BIAS,1-BIAS} m\, log_2\, m$

Metric — Abbreviation		*Equation*
True Positive Rate (recall, sensitivity, hit rate, recognition rate)	*TPR*	TP/P
True Negative Rate (inverse recall, specificity)	*TNR*	TN/N
Positive Predictive Value (precision, confidence)	*PPV*	TP/OP
Negative Predictive Value (inverse precision)	*NPV*	TN/ON

Metric — Abbreviation		Equation
Joint Entropy	*HOC*	$-\sum_{oc=TP,FP,FN,TN} \frac{oc}{Sn} log_2 \frac{oc}{Sn}$
Mutual Information	*MI*	$\frac{TP}{Sn} log_2 \frac{TP/Sn}{PREV \cdot BIAS} +$ $\frac{FP}{Sn} log_2 \frac{FP/Sn}{(1 - PREV) \cdot BIAS} +$ $\frac{FN}{Sn} log_2 \frac{FN/Sn}{PREV \cdot (1 - BIAS)} +$ $\frac{TN}{Sn} log_2 \frac{TN/Sn}{(1 - PREV) \cdot (1 - BIAS)}$
Accuracy (efficiency, rand index, fraction correct)	*ACC*	TC/Sn
Misclassification Rate	*MCR*	$1 - ACC$
Informedness (±) (Youden’s index, delta P’, Peirce skill score)	*INFORM*	$TPR + TNR - 1$
Markedness (±) (delta P, Clayton skill score, predictive summary index)	*MARK*	$PPV + NPV - 1$
Balanced Accuracy (strength)	*BACC*	$\frac{TPR + TNR}{2} = \frac{TP \cdot N + TN \cdot P}{2P \cdot N}$
G-metric (G-mean, Fowlkes-Mallows index)	*G*	$\sqrt[2]{TPR \cdot TNR} = \sqrt[2]{{TP \cdot TN}/{P \cdot N}}$
Normalized Mutual Info	*nMI*	${MI}/{(HC + HO)/2}$
F-metric (F-score, F-measure, positive specific agreement)	*F1*	$\frac{2TP}{2TP + FC} = \frac{2PPV \cdot \mathrm{TPR}}{PPV + TPR}$
Cohen’s Kappa (±) (Heidke skill score, quality index)	*CK*	$\frac{ACC - CKc}{1 - CKc} = \frac{DET}{(P \cdot N \cdot OP \cdot ON)/2}$
Matthews Correlation Coefficient (±) (Phi correlation coefficient, Cohen’s index, Yule phi)	*MCC*	$\sqrt{INFORM \cdot MARK}$ $= \frac{DET}{\sqrt{P \cdot N \cdot OP \cdot ON}}$

Note that more information can be found in [40]. See Table 12 for the recently proposed metrics’ equations.

References

1. Luque A, Carrasco A, Martín A, de las Heras A (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 91:216–231. https://doi.org/10.1016/j.patcog.2019.02.023

2. Staartjes VE, Schröder ML (2018) Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? J Neurosurg Spine 29:611–612. https://doi.org/10.3171/2018.5.SPINE18543

3. Brown JB (2018) Classifiers and their metrics quantified. Mol Inform 37:1–11.

https://doi.org/10.1002/minf.201700127

4. Sokolova M (2006) Assessing invariance properties of evaluation measures. Proc Work Test Deployable Learn Decis Syst 19th Neural Inf Process Syst Conf (NIPS 2006) 1–6

5. Ranawana R, Palade V (2006) Optimized precision - a new measure for classifier performance evaluation. In: 2006 IEEE International Conference on Evolutionary Computation. IEEE, Vancouver, BC, Canada, pp 2254–2261

6. Garcia V, Mollineda R a., Sanchez JS (2010) Theoretical analysis of a performance measure for imbalanced data. 2006 IEEE Int Conf Pattern Recognit 617–620. https://doi.org/10.1109/ICPR.2010.156

7. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One 12:1–17. https://doi.org/10.1371/journal.pone.0177678

8. Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recognit Lett 30:27–38. https://doi.org/10.1016/j.patrec.2008.08.010

9. Seliya N, Khoshgoftaar TM, Van Hulse J (2009) Aggregating performance metrics for classifier evaluation. In: IEEE International Conference on Information Reuse and Integration, IRI. pp 35–40

10. Liu Y, Zhou Y, Wen S, Tang C (2016) A strategy on selecting performance metrics for classifier evaluation. Int J Mob Comput Multimed Commun 6:20–35. https://doi.org/10.4018/ijmcmc.2014100102

11. Brzezinski D, Stefanowski J, Susmaga R, Szczęch I (2018) Visual-based analysis of classification measures and their properties for class imbalanced problems. Inf Sci (Ny) 462:242–261. https://doi.org/10.1016/j.ins.2018.06.020

12. Hu B-G, Dong W-M (2014) A study on cost behaviors of binary classification measures in class-imbalanced problems. Comput Res Repos abs/1403.7:

13. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002

14. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17:299–310. https://doi.org/10.1109/TKDE.2005.50

15. Forbes A (1995) Classification-algorithm evaluation: five performance measures based on confusion matrices. J Clin Monit Comput 11:189–206. https://doi.org/10.1007/BF01617722

16. Pereira RB, Plastino A, Zadrozny B, Merschmann LHC (2018) Correlation analysis of performance measures for multi-label classification. Inf Process Manag 54:359–369. https://doi.org/10.1016/j.ipm.2018.01.002

17. Straube S, Krell MM (2014) How to evaluate an agent's behavior to infrequent events? Reliable performance estimation insensitive to class distribution. Front Comput Neurosci 8:1–6. https://doi.org/10.3389/fncom.2014.00043

18. Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5:1–11. https://doi.org/10.5121/ijdkp.2015.5201

19. Tharwat A (2020) Classification assessment methods. Appl Comput Informatics ahead-of-p:1–13. https://doi.org/10.1016/j.aci.2018.08.003

20. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:. https://doi.org/10.1186/s12864-019-6413-7

21. Brzezinski D, Stefanowski J, Susmaga R, Szczech I (2020) On the dynamics of classification measures for imbalanced and streaming data. IEEE Trans Neural Networks Learn Syst 31:1–11. https://doi.org/10.1109/TNNLS.2019.2899061

22. Baldi P, Brunak S, Chauvin Y, et al (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424. https://doi.org/10.1093/bioinformatics/16.5.412

23. Hu B-G, He R, Yuan X-T (2012) Information-theoretic measures for objective evaluation of classifications. Acta Autom Sin 38:1169–1182. https://doi.org/10.1016/S1874-1029(11)60289-9

24. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010

25. Valverde-Albacete FJ, Peláez-Moreno C (2014) 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. PLoS One 9:1–10. https://doi.org/10.1371/journal.pone.0084217

26. Shepperd M (2013) Assessing the predictive performance of machine learners in software defect prediction function. In: The 24th CREST Open Workshop (COW), on Machine Learning and Search Based Software Engineering (ML&SBSE). Centre for Research on Evolution, Search and Testing (CREST), London, pp 1–16

27. Schröder G, Thiele M, Lehner W (2011) Setting goals and choosing metrics for recommender system evaluations. In: UCERSTI 2 Workshop at the 5th ACM Conference on Recommender Systems. Chicago, Illinois, pp 1–8

28. Delgado R, Tibau XA (2019) Why Cohen's kappa should be avoided as performance measure in classification. PLoS One 14:1–26. https://doi.org/10.1371/journal.pone.0222916

29. Ma J, Zhou S (2020) Metric learning-guided k nearest neighbor multilabel classifier. Neural Comput Appl 1–15. https://doi.org/10.1007/s00521-020-05134-9

30. Fatourechi M, Ward RK, Mason SG, et al (2008) Comparison of evaluation metrics in classification applications with imbalanced datasets. In: 7th International Conference on Machine Learning and Applications (ICMLA). pp 777–782

31. Seliya N, Khoshgoftaar TM, Van Hulse J (2009) A study on the relationships of classifier performance metrics. In: 21st IEEE International Conference on Tools with Artificial Intelligence, ICTAI. pp 59–66

32. Joshi MV (2002) On evaluating performance of classifiers for rare classes. In: Proceedings IEEE International Conference on Data Mining. IEEE, pp 641–644

33. Caruana R, Niculescu-Mizil A (2004) Data mining in metric space: an empirical analysis of supervised learning performance criteria. Proc 10th ACM SIGKDD Int Conf Knowl Discov Data Min 69–78. https://doi.org/1-58113-888-1/04/0008

34. Huang J, Ling CX (2007) Constructing new and better evaluation measures for machine learning. IJCAI Int Jt Conf Artif Intell 859–864

35. Japkowicz N, Shah M (2011) Evaluating learning algorithms: A classification perspective. Cambridge University Press, Cambridge

36. Contreras-Reyes JE (2020) An asymptotic test for bimodality using the Kullback-Leibler divergence. Symmetry (Basel) 12:1–13. https://doi.org/10.3390/SYM12061013

37. Shi L, Campbell G, Jones WD, et al (2010) The Microarray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827–838. https://doi.org/10.1038/nbt.1665

38. Rohani A, Mamarabadi M (2019) Free alignment classification of dikarya fungi using some machine learning methods. Neural Comput Appl 31:6995–7016. https://doi.org/10.1007/s00521-018-3539-5

39. Azar AT, El-Said SA (2014) Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl 24:1163–1177. https://doi.org/10.1007/s00521-012-1324-4

40. Canbek G, Sagiroglu S, Taskaya Temizel T, Baykal N (2017) Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In: 2017 International Conference on Computer Science and Engineering (UBMK). IEEE, Antalya, Turkey, pp 821–826