IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 April 2023, accepted 15 May 2023, date of publication 22 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278996

TOPICAL REVIEW

Common Problems With the Usage of F-Measure and Accuracy Metrics in Medical Research

LUIGI LAVAZZA, (Senior Member, IEEE), AND SANDRO MORASCA, (Member, IEEE)
Dipartimento di Scienze Teoriche e Applicate, Università degli Studi dell'Insubria, 21100 Varese, Italy

Corresponding author: Luigi Lavazza (luigi.lavazza@uninsubria.it)

This work was supported in part by the "Fondo di Ricerca d'Ateneo" funded by the Università degli Studi dell'Insubria.

ABSTRACT *Problem* Binary classifiers are widely used in medical research, especially for diagnoses. They are usually evaluated via performance metrics computed based on confusion matrices. Accuracy and F-measure are among the most frequently used performance metrics, but they make implicit assumptions and do not take into account important characteristics of classifiers. As a consequence, evaluations based on Accuracy or F-measure may turn out to be incorrect, unreliable, and inadequate for the specific application context. The usage of Accuracy and F-measure is particularly critical in the medical domain, where selecting a sub-optimal classifier may lead to incorrect diagnoses, with potentially serious or even fatal consequences. *Aim* We investigated whether the improper or naive usage of Accuracy and F-measure can lead to partial or incorrect evaluations. If this is the case, we need a procedure to reinterpret the conclusions reported in research articles, whenever possible. *Method* After discussing a few important properties of Accuracy and F-measure, we examine a set of representative research articles, to assess their conclusions, and illustrate a procedure to reinterpret those conclusions. *Results* It appears that the examined research articles yield conclusions that are largely affected by the used performance metrics, which in some cases lead to very misleading conclusions. The application of the proposed procedure allows the retrieval of confusion matrices and the derivation of reliable indications of classifiers' performances. *Conclusion* F-measure and Accuracy should be used with care, being aware of their characteristics and limits. We recommend that future evaluations of binary classifiers be provided with the complete confusion matrices, so that users can formulate evaluations based on specific contexts and priorities.

INDEX TERMS Accuracy, binary classifiers, F-measure, F-score, performance metrics.

I. INTRODUCTION

Binary classification is increasingly used in the biomedical field to create new diagnostic models. Given the importance of obtaining correct diagnostic indications, evaluating the performance of binary classifiers is of paramount importance.

The performance of a binary classifier is fully represented via a so-called confusion matrix (as discussed in the next section), which is used to derive several "performance metrics." Each of them is defined to provide a simple and direct evaluation of performance, from different points of view and for different purposes. Once a performance metric is chosen, a binary classifier is considered "good enough" if its value for that performance metric is "high enough" according to some predefined minimum level of performance above which a classifier is considered satisfactory. Similarly, a binary classifier is considered better than another if it has a higher value for that performance metric.

Among the proposed performance metrics, Accuracy and the F-measure are widely used in medical research [1], [2], [3]. However, as we show in the next section, both have limitations and drawbacks, so that evaluations based on them may be incorrect, partial, or valid only in a very specific context.

To assess to what extent the naive usage of performance metrics can be misleading, we analyze a set of research articles dealing with diagnostic tests. We also propose a procedure to reinterpret the results reported via Accuracy,

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.

TABLE 1. A confusion matrix.

	Act. Negative	Act. Positive	
Est. Neg.	TN	FN	$EN = TN + FN$
Est. Pos.	FP	TP	$EP = FP + TP$
	$AN = TN + FP$	$AP = FN + TP$	$n = AN + AP$ $= EN + EP$

F-measure, and other performance metrics. The procedure involves retrieving the original confusion matrix from the published values of the performance metrics. The knowledge of the confusion matrix makes it possible to verify the consistency of the published results and compute additional metrics, which can shed new light on the published results. We applied the proposed procedure to the aforementioned set of scientific articles, to demonstrate its steps and the results it can provide.

Our article provides three main contributions.

- It shows the risks of improper usage of performance metrics, thus increasing the awareness of the problem and connected risks.
- It illustrates a method to evaluate the real performance of published classifiers, so that many already published results can be correctly reinterpreted.
- It provides suggestions concerning the proper evaluation of binary classifiers.

The importance of this study also derives from the nature of the biomedical field: misinterpreting classification results can lead to preferring a diagnostic model that is actually worse than another model that achieves better measures, with possibly quite serious consequences for patients' health.

The rest of the paper is organized as follows. Section II provides some basic information concerning performance metrics. Section III illustrates the proposed procedure to derive confusion matrices from published performance metrics, to compute additional performance metrics as well as retrieve characteristics of the used test sets. In Section IV a set of six representative research papers is analyzed: our analysis shows that inaccurate use of performance metrics can lead to wrong conclusions. The application of the proposed technique makes it possible to reinterpret the conclusions provided by published articles. Section V accounts for related work and positions our study in the context of ongoing research. Section VI discusses how the illustrated work can be applied to situations not addressed in our study. Finally, Section VII draws some conclusions and outlines future work.

In this article, we address specifically the medical domain; nonetheless, the proposed considerations and practices can be applied to binary classifiers in any domain. In fact, there are multiple domains where the consequences of bad classifications can have quite serious effects.

II. BACKGROUND

The performance of a binary classifier can be assessed based on a confusion matrix, whose schema is shown in Table 1.

TABLE 2. Performance Metrics and Prevalence.

Formula	Terms
$PPV = \frac{TP}{EP}$	Positive Predictive Value, Precision
$TPR = \frac{TP}{AP}$	True Positive Rate, Recall, Sensitivity
$FPR = \frac{FP}{AN}$	False Positive Rate, fall-out, false alarm ratio
$TNR = \frac{TN}{AN}$	True Negative Rate, Specificity
$FM = \frac{2\ PPV\ TPR}{PPV + TPR}$	F-score, F-measure, F-1 score
$\phi = \frac{TP\ TN - FP\ FN}{\sqrt{AN\ AP\ EN\ EP}}$	Phi, Matthews Correlation Coefficient
$ACC = \frac{TP+TN}{n}$	Accuracy
$\rho = \frac{AP}{n}$	Prevalence

In Table 1, AP and AN are the numbers of actual positives and actual negatives, respectively. EP and EN are the numbers of estimated positives and estimated negatives, respectively. It is $AP + AN = EP + EN = n$, where n is the total number of subjects in the considered dataset.

TN is the number of true negatives, i.e., the number of subjects that are correctly estimated negative; TP is the number of true positives, i.e., the number of subjects that are correctly estimated positive; FN is the number of false negatives, i.e., the number of subjects that are positive but are incorrectly estimated negative; FP is the number of false positives, i.e., the number of subjects that are negative but are incorrectly estimated positive.

We denote by $\rho = \frac{AP}{n}$ the *prevalence* of positive subjects in a dataset. Note that prevalence is tightly linked to the concept of data imbalance: $\rho = 0.5$ indicates a perfectly balanced dataset, while ρ close to zero or one indicates a large imbalance, in favor of negatives, or, respectively, positive subjects.

Sometimes, the confusion matrix is populated with relative values, i.e., $\frac{TN}{n}$, $\frac{FP}{n}$, $\frac{FN}{n}$, $\frac{TP}{n}$. We denote such values as tn, fp, fn and tp, respectively. Similarly, $ep = \frac{EP}{n}$, $en = \frac{EN}{n}$, $ap = \frac{AP}{n}$, $an = \frac{AN}{n}$. Note that $ap = \rho$ and $tn + fp + fn + tp = 1$.

A. PERFORMANCE METRICS

The performance of a given binary classifier applied to a given dataset is always completely represented by the confusion matrix. However, researchers quite often prefer to summarize performance via a single number rather than report the values of the four cells of the confusion matrix. So, several performance metrics have been defined as functions of the cells of the confusion matrix. Table 2 shows some of the most widely used performance metrics.

Several performance metrics provide a focused but partial view of performance. For instance, in Table 2, PPV takes into account data from a single row of a confusion matrix, while TPR is computed using the data from a single column. FM depends on all cells except TN. Other performance metrics, like ϕ, take into account all four cells of a confusion matrix.

B. PROBLEMS WITH THE F-MEASURE

The F-measure was originally defined to evaluate the performance of information retrieval techniques [4]. The information retrieval domain is characterized by a huge (and often unknown) rate of actual negatives. For instance, a Web search based on a typical query returns hundreds or thousands of relevant pages, out of more than 10^{13} total pages. It is clearly irrelevant whether the correctly ignored pages are 10^{13} or 10^{14}, or even more. So, the F-measure does not take into account the number TN (or rate *tn*) of true negatives.

Being the geometric mean of *Precision* and *Recall*, which are also quite popular performance metrics, the F-measure is often perceived as a convenient means for obtaining an overall evaluation of binary classifier performance. Its value can also be computed according to (1), which clearly shows that the F-measure does not depend on TN.

$$FM = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}} = \frac{2\,TP}{AP + EP} = \frac{2TP}{2TP + FN + FP} \quad (1)$$

The F-measure as an overall performance metric has drawn a lot of criticisms, because of its numerous drawbacks [5], [6], [7].

The first problem directly descends from the fact that FM does not take into account TN. Let us consider confusion matrices CM_a and CM_b shown below, obtained for two different datasets. CM_a and CM_b only differ in the number of true negatives

CM_a	
TN=30	FN=10
FP=40	TP=50

CM_b	
TN=100	FN=10
FP=40	TP=50

Both have the same value $FM_a = FM_b = \frac{2 \cdot 50}{2 \cdot 50 + 10 + 40} \simeq 0.67$. So, the fact that 70 more true negatives are correctly classified in CM_b than in CM_a is ignored.

Take now a third confusion matrix CM_c, concerning a third dataset.

CM_c	
TN=5	FN=10
FP=39	TP=51

It is $FM_c \simeq 0.68$; thus, according to FM, one should conclude that the performance represented by CM_c is slightly better than those represented in CM_a and CM_b. However, though one more actual positive is classified correctly in CM_c, when it comes to classifying actual negatives CM_c performs quite poorly.

Based on these examples and (1), it appears that although ignoring true negatives is acceptable and even useful in domains like information retrieval, FM is not an adequate metric for quantifying the overall performance of a binary classifier, since it does not use all available information about the classification results. This is one of the main criticisms made to FM by previous studies [5], [6], [7].

Another important limitation of FM is that knowing the value of FM is not sufficient to evaluate how good the performance of the given diagnostic test is. A binary classifier is typically assessed by comparing its performance against the performance of a baseline model. The whole point of learning a binary classifier is to take advantage as much as possible of the information about the characteristics of each individual subject with the goal of estimating whether he or she is actually positive or negative. An alternative is to blindly carry out the estimations by random classification, which can be taken as the simplest baseline model. Clearly, it makes little sense in general to adopt a diagnostic test that performs not better than a random classifier, where each actually positive subject has a probability ρ of being estimated positive. A binary classifier that instead uses information about the features of subjects (e.g., the results of a diagnostic test) should estimate actual positives with a higher probability than ρ and, conversely, actual negatives with a lower probability than ρ. The expected (i.e., mean) values of TPR and PPV for a random classification are both equal to ρ [8]: by using these values in (1), we obtain FM=ρ as well, so, when evaluating a classifier, we should compare its FM against ρ. Thus, the knowledge of the value of FM by itself is not sufficient to tell whether a classifier performs better than even random estimation [7]. For instance, suppose that FM = 0.85 for a binary classifier. Even though specific guidelines for the interpretation of the values of FM do not exist, FM=0.85 can be considered a high value for FM. However, in a dataset in which ρ=0.9, FM=0.85 would denote a worse-than-random (hence unacceptable) performance. Even with ρ=0.8, FM=0.85 should be considered as a mediocre performance.

Comparing the performance of a diagnostic test with a baseline is a commendable practice, regardless of the specific performance metric used. However, with some metrics the comparison is immediate: for instance, random classifiers have $\phi = 0$ on average, for any value of ρ.

Despite the problems described above, the F-measure is widely used in the field of medical diagnostics. This is clearly dangerous, since there is the risk that unwarranted conclusions are drawn from otherwise correctly conducted studies.

C. PROBLEMS WITH ACCURACY

The main problem with Accuracy is that it penalizes false positives and false negatives in the same way. As an example, let us consider the following confusion matrices, obtained for a dataset with AP=AN=100:

CM_d	
TN=80	FN=20
FP=20	TP=80

CM_e	
TN=70	FN=10
FP=30	TP=90

$ACC = \frac{160}{200} = 0.8$ in both cases: the fact that CM_e involves fewer false negatives and more false positives than CM_d is not considered. However, when the consequences of incorrect diagnoses are taken into consideration, we may find that the relative importance of false negatives is much greater than the importance of false positives. In fact, the cost associated with false positives is usually due to some additional diagnostic tests, which in general reveal that the subjects are actually

negative, so that there are no further costs. Instead, the cost of false negatives may be huge: when fatal diseases are concerned, incorrectly diagnosing a subject as negative may delay additional checks and treatments, possibly resulting in longer and more expensive therapies and even causing serious harm to the subjects.

In conclusion, the binary classifier whose results are in CM_e is in most cases preferable to a binary classifier whose results are in CM_d. Note that costs vary depending on the specific situations at hand: in some cases, it is also possible that false positives cost more than false negatives. Therefore, the only way to let people accurately compute misdiagnoses costs is by supplying the confusion matrix, rather than an overall performance metric.

Random classifiers that estimate a given subject positive with probability ρ have $ACC = \rho^2 + (1-\rho)^2$, on average. When $\rho = 0.1$, $\rho^2+(1-\rho)^2$ is 0.82; hence, even a high value of ACC like 0.85 would actually indicate a performance that is barely better than random classification's.

III. A PROCEDURE TO RE-INTERPRET PUBLISHED PERFORMANCE METRICS

The proposed method is based on the derivation of confusion matrices from published performance metrics.

As illustrated above, performance metrics are obtained by combining elements of the confusion matrix. Therefore, if we have enough performance metrics obtained from the same confusion matrix, it is possible to write a system of equations whose solution yields the elements of the confusion matrix.

We illustrate the procedure by means of an example. Suppose that a paper reports that a given diagnostic test achieved ACC=0.706, TPR=0.430, FPR=0.031, and PPV=0.930. Using the definition of these metrics, we obtain the following system of equations where tp, fp, tn, and fn are the unknowns

$$\begin{cases} tp + tn = ACC = 0.706 \\ \dfrac{tp}{tp+fn} = TPR = 0.430 \\ \dfrac{fp}{tn+fp} = FPR = 0.031 \\ \dfrac{tp}{tp+fp} = PPV = 0.930 \end{cases}$$

By solving the system of equations, we obtain the following rates confusion matrix

tn=0.495	fn=0.279
fp=0.016	tp=0.211

Note that the knowledge of even three performance metrics instead of four would have allowed us to derive the rates confusion matrix, because $tn + fp + fn + tp = 1$, so one cell of the rates confusion matrix can always be computed as a function of the other three (e.g., $tp = 1 - tn - fp - fn$).

The knowledge of the rates confusion matrix makes it possible to carry out a few analyses, in addition to the ones from the paper that reported only the values of ACC, TPR, FPR, and PPV, as we now discuss.

A. COMPUTING ADDITIONAL PERFORMANCE METRICS

If a paper bases its conclusions on performance metrics like Accuracy and F-measure, the risks described in the previous sections apply. In such cases, it is a good idea to compute additional performance metrics, selected among the most reliable ones, and check whether the original conclusions are also supported by these additional performance metrics. To this end, several authors recommend using ϕ [7], [9]. In fact, ϕ (see the definition in Table 2) is an effect size measure, which quantifies how far the classification performed by a binary classifier is from random classification, in which each subject has the same probability of being estimated positive. The value of ϕ obtained with random classification is 0. ϕ is also related to the χ^2 statistic, since $|\phi| = \sqrt{\frac{\chi^2}{n}}$.

A recent study [7] showed that, in the software engineering domain, around 22% of the results published in 38 considered studies would be reversed if ϕ is selected as a performance metric instead of the F-measure. This kind of risk is not limited to a specific domain, since it stems from the very definitions of the involved performance metrics.

In general, given two binary classifiers C_a and C_b, it is possible that C_a performs better than C_b according to the F-measure or Accuracy, while it performs worse than C_b according to ϕ. Similarly, when considering a single binary classifier C_a, it is possible that C_a achieves a quite high value of F-measure or Accuracy, while it gets a rather low ϕ. In these cases, we should trust ϕ or, even better, look for further evidence about the real performance of the considered binary classifier(s). To this end, it can be quite helpful to look at the values of all the four cells of the confusion matrices, instead of a single summarizing performance metric.

B. CONSIDERING COST

As already mentioned, the costs of false positives and false negatives are different, in general. Nonetheless, Accuracy assigns equal importance to false positives and false negatives. In fact, Accuracy can be defined as

$$ACC = \frac{TP+TN}{n} = \frac{AP - FN + AN - FP}{n} = \frac{n - FN - FP}{n}$$

or, equivalently, as $ACC = 1 - fn - fp$.

Instead, the F-measure does not account equally for false negatives and false positives. In fact, the definition of F-measure in (1) can be rewritten as follows

$$FM = \frac{2(AP-FN)}{2(AP-FN)+FN+FP} = \frac{1}{1+\frac{FN+FP}{2(AP-FN)}}.$$

Increasing or decreasing FN by some amount has more impact on FM than does increasing or decreasing FP by the same amount [10].

C. INTERNAL CONSISTENCY CHECKS

Having derived the confusion matrices concerning a diagnostic test, a few checks can be carried out, especially concerning the characteristics of the test set.

AP and AN can be respectively computed as $AP = TP + FN$ and $AN = TN + FP$, and prevalence ρ (the fraction of positive subjects) as $\rho = \frac{AP}{AP+AN}$. To be clear, ρ is here the prevalence of the test set, since it has been derived from the performance metrics obtained by applying a given diagnostic test to the test set.

When analyzing the comparison of multiple binary classifiers, we can check if all of them were assessed with test sets with the same prevalence. Ideally, multiple binary classifiers should be assessed using the same test set, to have fully meaningful comparisons. Finding that all test sets used for the comparison of multiple binary classifiers in a scientific study have the same prevalence does not guarantee that the same test set was used. However, finding different prevalence values is sufficient to conclude that different test sets were used, which may impair the validity and usefulness of the comparison.

IV. ANALYSIS OF SELECTED PAPERS

In this section, we use a few recent scientific papers from well-known journals and conferences to illustrate the possible problems deriving from using Accuracy and the F-measure, and, when possible, show how to provide sounder, more complete interpretations of the published results.

The papers we selected as examples are representative of the usage of F-measure or Accuracy (possibly in combination with other metrics) and report enough data to make the re-interpretation of results possible.

Note that we use the published data, which are typically rounded to the third decimal digit. This implies that our results are generally accurate up to the second decimal digit. For instance, in the first row of Table 4, it is $tn+fp+fn+tp = 1.001$, instead of 1. This level of imprecision does not affect our considerations.

In discussing each case, we use the terminology used in the considered paper. So, for instance, we may write F-score instead of F-measure, after the original paper.

A. CASE 1

We here consider a study that proposes *"using a novel combination of short-term and long-term features from different timescales to develop an automatic newborn cry diagnostic system to differentiate the cry audio signals (CASs) of healthy infants from those with respiratory distress syndrome (RDS)"* [11].

The study reports the performance metrics of classifiers, obtained using Support Vector Machines, using inspiration episodes as well as classifiers using expiration episodes. We consider only the classifiers using individual and combined feature sets for the expiration dataset, whose performance data reported in the original paper are in Table 3.

The conclusions of the study state *"The combination of long-term (melody[1] and rhythm) and short-term (MFCCs) features was found to provide a better classification performance for differentiating the CAS of healthy infants from infants with RDS in comparison to using short-term features alone, particularly for the expiration episodes. The best improvements of the results (F-score) that we achieved were 10.3% in the expiration episode."*

[1]Tilt features were used in the study to parameterize the melody features in the CAS.

TABLE 3. Performance of classifiers (Table 4 of [11]).

Expiration	ACC	TPR	FPR	PPV	FM
MFCC	0.706	0.430	0.031	0.930	0.588
Tilt	0.555	0.634	0.520	0.536	0.581
Rhythm	0.445	0.355	0.469	0.417	0.384
MFCC&Tilt	0.733	0.602	0.143	0.800	0.687
MFCC&Rhythm	0.722	0.462	0.031	0.934	0.618
MFCC&Tilt&Rhythm	0.738	0.602	0.132	0.811	0.691

The values of the performance metrics reported allowed us to compute the cells of the rates confusion matrix, based on which we computed ϕ for every classifier. Table 4 shows both the rates confusion matrices and the values of ϕ.

TABLE 4. Confusion matrices and ϕ derived from Table 4 of [11].

Expiration	tn	fp	fn	tp	ϕ
MFCC	0.495	0.016	0.279	0.211	0.48
Tilt	0.247	0.267	0.178	0.308	0.12
Rhythm	0.273	0.241	0.313	0.172	-0.12
MFCC&Tilt	0.440	0.073	0.194	0.293	0.48
MFCC&Rhythm	0.497	0.016	0.262	0.225	0.50
MFCC&Tilt&Rhythm	0.447	0.068	0.193	0.291	0.49

The data in Table 4 cast a new light on the results of the classifiers.

If we base the evaluation on ϕ, as many authors suggest [7], [9], [10], we can observe that the improvements obtained using Tilt and Rhythm together with MFCC are marginal (MFCC alone achieves $\phi = 0.48$, together with Tilt and/or Rhythm the classification reaches $\phi = 0.5$). We also gather evidence that Tilt or Rhythm alone provides performances that are hardly better than random estimation (ϕ being close to zero). The increase of F-score obtained using Tilt and Rhythm does not correspond to an equally large improvement of ϕ. This is possible because prevalence ρ in the test set is close to 0.5 [10]. In addition, according to ϕ, the best result is achieved by MFCC&Rhythm, not by MFCC&Tilt&Rhythm, as stated in the conclusions of the paper.

At any rate, we can base the evaluation of the results directly on the confusion matrices, rather than on a performance metric like F-score or ϕ. Let us consider the performance of the classifier that uses MFCC&Tilt&Rhythm against the performance of the classifier that uses only MFCC. The former classifier increases tp and decreases fn, at the expense of an increase of fp and a decrease of tn. It achieves a high F-score because the decrease of tn is not considered in the computation of the F-score. However, ϕ does consider all of the confusion matrix elements: in this case, the decrease of fn occurs with an increase of fp, and the

increase of *tp* with a decrease of *tn*, so that ϕ does not change much.

However, one could observe that increasing *tp* and decreasing *fn* is what really counts for most diagnostic tests. Accordingly, the classifier based on Tilt alone, which maximizes *tp* and minimizes *fn*, could be considered a very good classifier, even though it gives low values of both F-score and ϕ, because of the relatively high number of false positives.

So, how good is the Tilt-based classifier in practice? The answer to this question depends on the cost associated with false positives: if such costs are low, classifications based on Tilt could turn out to be preferable to all other classifications. In general, as costs are quite context-sensitive, it is advisable that studies like the one considered publish confusion matrices (like in Table 4) so that anybody can perform contextualized evaluations, by applying the most likely costs in the considered situation. As for costs, we need to highlight that *global* costs should be evaluated, taking into account not only misclassifications, but also the cost of true positives (i.e., the cost of the needed treatment) as well as a possibly limited budget.

In conclusion, the usage of the F-score in the considered study resulted in hiding a number of interesting considerations, and in leading to overoptimistic conclusions.

B. CASE 2

In this section, we consider a paper that *"proposes a new framework to automatically identify or confirm COVID-19 in cough audio signals based on six machine learning algorithms"* [12]. The paper also proposes and evaluates the usage of genetic algorithm (GA) in combination with the ML techniques.

The paper reports different types of performance metrics, including F-score, Precision, Recall, and Accuracy. The confusion matrices obtained by applying different machine learning techniques are also given: this made the computation of other performance metrics, like ϕ, straightforward.

Table 5 summarizes the confusion matrices that describe the performances of classifiers obtained using using ML techniques. It is a transcription of Fig. 7 of the original paper [12], in which we added the computation of ϕ.

TABLE 5. Confusion matrices and ϕ for ML techniques [12].

	TN	FP	FN	TP	ϕ
LR	4	12	1	124	0.42
LDA	8	8	2	123	0.60
KNN	12	4	2	123	0.78
CART	5	11	6	119	0.31
NB	9	7	19	106	0.33
SVM	5	11	0	125	0.54

Table 6 summarizes the confusion matrices that describe the performances of classifiers obtained using GA-ML techniques. It is a transcription of Fig. 8 of the original paper [12], in which we added the computation of ϕ.

TABLE 6. Confusion matrices and ϕ for GA-ML techniques [12].

	TN	FP	FN	TP	ϕ
LR	5	11	0	125	0.54
LDA	9	7	1	124	0.69
KNN	13	3	0	125	0.89
CART	8	8	3	122	0.56
NB	8	8	4	121	0.53
SVM	7	9	0	125	0.64

In addition, the original paper reported that in combination with GA, all the ML techniques achieve accuracy over 90%.

The authors conclude that *"The results showed that the KNN algorithm provides the best results based on different evaluation metrics compared with the other algorithms in the detection and diagnosis process."* The statement appears correct, in that KNN achieves the highest values of ϕ with both ML ($\phi = 0.78$) and GA-ML ($\phi = 0.89$) techniques.

However, the authors also report results from other similar papers, for the sake of comparison. Among others, the paper reports the results, expressed via Accuracy, achieved in a study by Pahar et al. [13], who noticeably used the same dataset, i.e., the Coswara dataset. Based on the reported results, the authors state *"It is clear that the proposed framework has more than an advantage over all of the conventional methods. Therefore, our proposed framework is presented for cough detection using a hybrid method of genetic and ML models to improve the performance of existing machine learning models in COVID19 detection."* Unfortunately, this statement appears overoptimistic when results are evaluated and compared via sounder performance metrics. As we did in Case 1, we reconstructed the confusion matrices of the results by Pahar et al. [13] and then computed ϕ, as shown in Table 7.

The comparison of Table 7 with Tables 5 and 6 shows that, according to ϕ, Resnet50 appears to perform better (although marginally) than any method proposed in the considered study. It is also apparent that CNN's performance is as good as the best model proposed in the considered study.

Reconstructing the original classifications also allowed us to carry out an additional check. Even though the two papers claim to use the same dataset as the test set, ρ is 0.887 in the paper by Hemdan et al. [12], while ρ is close to 0.5 in the paper by Pahar et al. [13]. This casts serious doubts on the soundness of the comparison, which seems to involve performance metrics computed on different datasets.

C. CASE 3

In this section, we deal with a study that aims "*to predict Chronic Kidney Disease (CKD)*" [14]. The study investigated different techniques to build classifiers, namely artificial neural network (ANN), C5.0, logistic regression (LR), Chi-square automatic interaction detection (CHAID), linear support vector machine (LSVM), K-nearest neighbors (KNN) and random tree (RT). All the mentioned techniques were applied with three types of feature selections: correlation-based feature selection (CFS), Wrapper method, and LASSO.

TABLE 7. Confusion matrices and ϕ for the results of [13].

Classifier	Best Feature Hyperparameters	tn	fp	fn	tp	ϕ
LR	M=13, F=1024, S=120	0.28	0.03	0.21	0.48	0.55
LR	M=26, F=1024, S=70	0.30	0.13	0.21	0.36	0.33
KNN	M=26, F=2048, S=100	0.30	0.09	0.16	0.45	0.49
KNN	M=26, F=1024, S=70	0.27	0.11	0.15	0.47	0.46
SVM	M=39, F=2048, S=100	0.32	0.17	0.11	0.41	0.45
SVM	M=26, F=1024, S=50	NA	NA	NA	NA	NA
MLP	M=26, F=2048, S=100	0.43	0.06	0.06	0.45	0.75
MLP	M=13, F=1024, S=100	0.42	0.16	0.08	0.34	0.53
CNN	M=26, F=1024, S=70	0.50	0.05	0.01	0.44	0.89
CNN	M=39, F=1024, S=50	0.53	0.05	0.01	0.41	0.89
LSTM	M=13, F=2048, S=70	0.49	0.04	0.02	0.45	0.88
LSTM	M=26, F=2048, S=100	0.51	0.05	0.02	0.43	0.87
Resnet50	M=39, F=1024, S=50	0.46	0.04	0.01	0.50	0.91
Resnet50	M=26, F=1024, S=70	0.39	0.04	0.01	0.56	0.90

In addition, classifiers were built with and without using SMOTE to balance the dataset.

In the paper, a number of performance metrics were used, namely Precision, Recall, Accuracy, F-measure, AUC (Area Under the ROC Curve) and the Gini coefficient. Nonetheless, the authors use only Accuracy both to compare the classifiers obtained in their study and make comparisons with the results reported in the literature. Also the conclusions are based on Accuracy: "*It was observed that LSVM achieved the highest accuracy of 98.86% in SMOTE with full features*" and "*LSVM achieved the highest accuracy in all experiments as compared to other classifiers algorithms*."

This study drew our attention because it reports several performance metrics, but it bases its evaluations and conclusions uniquely on Accuracy. This is particularly interesting, because basing the evaluation on F-measure leads to different conclusions: KNN with CFS achieves a value of the F-measure (0.985), which is slightly greater than the F-measure (0.983) of the best-performing model according to the authors (LSVM with SMOTE and full features).

Noticeably, in the conclusions, the authors also state that "*Logistic and KNN did not give suitable results*," which does not appear true, as far as KNN is concerned, when the F-measure is taken into account.

Thus, we proceeded to derive the confusion matrices from the published data, along the same lines as in the cases described above. Based on the confusion matrices, we computed ϕ. Considering ϕ, we can confirm the conclusions of the study, since LSVM with SMOTE and full features achieved the highest ϕ, namely 0.967. However, we can also add a few considerations, based on ϕ:

- CHAID with SMOTE and feature selection achieves $\phi = 0.958$, i.e., a performance level that is only marginally smaller than LSVM's.
- The paper reports that "LSVM with penalty L2 gave a better result in all techniques." This statement is not true when the comparison is based on ϕ: CHAID gets higher ϕ than LSVM in a few cases. So, even though it is true that LSM always achieves good performance, it is not always the *best* performance.

At any rate, we must not necessarily regard ϕ as the ultimate performance metric. In fact, looking into the confusion matrices themselves supports a definitely more accurate and complete evaluation. Let us consider the confusion matrices of the considered methods when SMOTE and selected features are used. The rates confusion matrices are given in Table 8 (note that the study did not use LR and KNN in this case).

TABLE 8. Confusion matrices and ϕ for results obtained with SMOTE and selected features [14].

	tn	fp	fn	tp	ϕ	ACC	FM
ANN	0.566	0.066	0.015	0.354	0.836	0.919	0.897
C5.0	0.523	0.108	0.005	0.364	0.788	0.887	0.866
CHAID	0.588	0.017	0.004	0.391	0.958	0.980	0.974
LSVM(L1,05)	0.626	0.005	0.010	0.359	0.967	0.985	0.979
LSVM(L2,05)	0.626	0.005	0.010	0.359	0.967	0.985	0.979
Random Tree	0.523	0.104	0.004	0.369	0.798	0.892	0.873

Table 8 shows that CHAID achieves slightly smaller values of F-score, Accuracy, and ϕ, with respect to LSVM. Nonetheless, it achieves more true positives and fewer false negatives. The performance metrics for CHAID are penalized by a relatively high fraction of false positives. To draw a reliable conclusion about which method is best, one should consider the cost of false positives: if the cost of false positives is small enough, CHAID is preferable to LSVM.

At any rate, while performing the computations to derive confusion matrices from the data retrieved from the original paper, we found a few problems (as described in Appendix A) that cast some doubts on the reliability of the data, although they do not affect the validity of the observations above.

D. CASE 4

We now analyze the results of a study aimed at "*demonstrating that it is possible to reliably identify cases of chronic spinal cord injury or disease (SCI/D) in a primary care electronic medical records database using a detailed case definition, a comprehensive keyword search strategy, and a rigorous manual chart review process*" [15].

The original paper reports the confusion matrices for all of the proposed algorithms, and uses the F-score (along with its components Recall and Precision) to identify the best results. Using the confusion matrices, we computed ϕ, as shown in Table 9.

TABLE 9. Confusion matrices and ϕ for the best results of [15].

TN	FP	FN	TP	AP	AN	FM	ϕ
667	10	37	89	126	677	0.791	0.765
670	7	41	85	126	677	0.780	0.759
667	10	39	87	126	677	0.780	0.754
666	11	40	86	126	677	0.771	0.744
670	7	43	83	126	677	0.769	0.748

Table 9 shows that

- High values of the F-score correspond to high values of ϕ. This is consistent with the fact that ρ is rather small ($\rho = \frac{AP}{n} = \frac{126}{126+677} \simeq 0.157$) [10].
- The algorithm featuring the highest F-score also has the highest ϕ.
- The algorithm that has the highest F-score and ϕ is also the one that maximizes TP and minimizes FN.
- In all cases, $\rho \simeq 0.157$. This indicates that the various algorithms were tested with the same dataset (or datasets having the same prevalence).

In conclusion, in the considered case, looking at the confusion matrices and ϕ confirms the considerations based on the F-score, but with greater confidence on the correctness of the evaluations.

However, by considering the confusion matrix, it is possible to note two interesting characteristics of the proposed diagnostic test. First, the test tends to estimate negative more subjects than the actual negatives, i.e., for all rows of Table 9 it is $TN + FN > AN$. Second, more than 30% of the positive subjects are incorrectly estimated negative: this may have important adverse consequences.

E. CASE 5

In this section, we consider a paper that proposed and evaluated the usage of fecal immunochemical tests (FIT), which is commonly used for screening, for detection of *H. pylori*, the main risk factor for gastric cancer [16].

The authors compared the results obtained via 1) ELISA stool antigen test in standard feces tube (SAT), 2) ELISA stool antigen test in FIT tube (Hp-FIT), and 3) blood sampling (Serological), by using performance metrics PPV, NPV, Sensitivity (alias TPR), Specificity (alias TNR) and Accuracy. Based on the collected results, the authors observe that[2] "*SAT and Hp-FIT showed comparable overall accuracy 71.1% vs. 77.6%, respectively; sensitivity of SAT was 91.8% versus 94.2%. Serology scored low with an overall accuracy of 49.7%.*" Based on these observations, the authors conclude that FIT can be used with high accuracy and sensitivity for diagnosing *H. pylori* and is rated as the most convenient test.

This case is a bit different from the previous ones, in that those concerned the usage of binary classifiers (mostly derived via machine learning techniques), while the present case deals with the direct evaluation of the performance of different diagnostic tests. At any rate, performance metrics are used in this case for the same purpose and in the same way as in previous cases.

From the published performance metrics, the confusion matrix could be easily derived. It is shown in Table 10, along with ϕ.

TABLE 10. Confusion matrix and ϕ for the results of [16].

	fp	*fn*	*tp*	*tn*	ϕ
Hp-FIT	0.20	0.02	0.30	0.48	0.60
SAT	0.26	0.03	0.30	0.42	0.50
Serological	0.47	0.03	0.26	0.23	0.23

Based on Table 10, we can add some arguments that support the conclusions of the authors: first, Hp-FIT achieves the highest value of ϕ; second, Hp-FIT minimizes *fn* and maximizes *tp*; third, although the performances of Hp-FIT and SAT are very close as far as *fn* and *tp* are concerned, Hp-FIT achieve better *fp* and *tn* than SAT.

In this case, our technique increases the confidence that the conclusions reported in the original paper are well supported by the collected data.

F. CASE 6

Gupta et al. [17] developed a Neural Architecture Search (NAS) method to find the best convolutional architecture capable of detecting pneumonia from chest X-rays. They proposed a Learning by Teaching framework inspired by the teaching-driven learning methodology from humans, and conducted experiments on a pneumonia chest X-ray dataset with over 5000 images. The proposed method achieved AUC=97.6%; Gupta et al. state that this "improves upon previous NAS methods by 5.1% (absolute)."

In Table 1 of their paper, Gupta et al. provide several performance metrics, including TPR, TNR, FM and ACC. These metrics were obtained from "fivefold cross validation"; the reported metrics are "the mean and standard deviation of the five test performance numbers." Gupta et al. compared the results of their methods (LBT-DARTS and LBT-PC-DARTS) with the results of other 18 methods. The standard deviation was greater than 0.01 only in 6 cases out of 80, and also in those cases it was just slightly greater than 0.01; hence, in what follows, we consider only the mean values, which appear to be reliable representations of the obtained performances.

As in previous cases, we computed *tp*, *fn*, *tn* and *fp* for each method, based on the provided performance metrics. The results are in Table 11.

The observation of Table 11 reveals that *ap* ranges from 0.435 to 0.729. This implies that the considered methods were tested with datasets having different prevalence: as a consequence, the reliability of results is dubious.

It can be noticed that the data that we computed include some error, due to the relatively low precision of the original data: for instance, it can be noticed that *ap*+*an* is not always 1, being slightly greater than 1 in all the rows of Table 11.

[2]The authors also report confidence intervals, which we do not use here, although the same kind of reasoning we propose can be applied to confidence intervals as well.

TABLE 11. Table from [17] with derived *tp*, *fn*, *tn*, *fp*.

Method	TPR	TNR	FM	ACC	fp	fn	tn	tp	ap	an
VGG19	0.927	0.924	0.93	0.927	0.037	0.040	0.447	0.512	0.553	0.484
InceptionV3	0.918	0.922	0.914	0.926	0.044	0.040	0.518	0.443	0.482	0.561
DenseNet121	0.938	0.917	0.924	0.931	0.046	0.031	0.505	0.464	0.495	0.551
AlexNet	0.925	0.927	0.921	0.927	0.041	0.036	0.515	0.448	0.485	0.556
VGG16	0.909	0.941	0.918	0.925	0.033	0.043	0.532	0.425	0.468	0.566
Xception	0.907	0.923	0.936	0.921	0.023	0.068	0.271	0.661	0.729	0.294
GoogLeNet	0.907	0.925	0.918	0.934	0.037	0.050	0.460	0.490	0.540	0.497
LeNet5	0.846	0.859	0.854	0.891	0.074	0.084	0.452	0.464	0.548	0.526
Kermany et al.	0.928	0.922	0.925	0.93	0.041	0.037	0.481	0.481	0.519	0.522
Stephen et al.	0.924	0.927	0.924	0.937	0.039	0.039	0.491	0.470	0.509	0.530
Siddiqi	0.947	0.931	0.927	0.935	0.042	0.023	0.565	0.412	0.435	0.607
Liang et al.	0.895	0.917	0.899	0.923	0.047	0.051	0.515	0.434	0.485	0.562
Meta Pseudo Label	0.906	0.923	0.917	0.918	0.038	0.051	0.456	0.493	0.544	0.494
Liu et al.	0.92	0.927	0.924	0.924	0.037	0.042	0.475	0.483	0.525	0.513
Kundu et al.	0.924	0.916	0.918	0.919	0.045	0.039	0.493	0.469	0.507	0.538
Cha et al.	0.921	0.913	0.914	0.92	0.047	0.040	0.497	0.463	0.503	0.545
DARTS	0.889	0.892	0.901	0.898	0.050	0.065	0.411	0.524	0.589	0.460
LBT-DARTS	0.93	0.932	0.928	0.933	0.037	0.035	0.505	0.461	0.495	0.541
PC-DARTS	0.932	0.909	0.918	0.914	0.050	0.034	0.496	0.470	0.504	0.546
LBT-PC-DARTS	0.959	0.967	0.971	0.97	0.011	0.028	0.323	0.649	0.677	0.334

Nonetheless, the differences among *ap* values are too large to be due only to the precision of our computations.

The observation of Table 11 reveals also that the method by Siddiqi appears to feature better *fn* than LBT-PC-DARTS, hence it could be preferable, if the cost of false negatives is prominent. Unfortunately, in this case the precision of the computed *fn* does not allow us to be sure that Siddiqi's method is preferable. As already noted, the authors themselves should provide the confusion matrices, so that the readers can draw their conclusions about which method is preferable.

G. FINAL REMARKS ON THE ANALYSIS OF PAPERS

We observed that several papers (including many not discussed here) use F-score and Accuracy, because these metrics had been used in similar previous papers. The motivation is that authors want to compare the results yielded by the new techniques and models they propose with the results that had been achieved previously. This mechanism seems to perpetuate the usage of performance metrics that, as shown above, have serious drawbacks.

In addition, several papers propose evaluations based on datasets having different prevalence. This phenomenon was observed both within the same paper and when comparing results from different papers.

The application of the proposed procedure to the six cases discussed above shows that it was possible to uncover problems with the evaluation of binary classifiers. In addition, we could also compute reliable indicators of performance. Specifically, in some cases, we were able to perform sound comparisons of results published in different papers.

V. RELATED WORK

The relevance of Accuracy and the F-measure in medical informatics has been documented in a systematic literature review by Hasan and Yao [18]. They found that "accuracy and specificity are among the most popular performance metrics used" and "as of 2018, the F-measure began to attract the researchers' attention and then used dramatically for performance measurement."

The characteristics of performance metrics have been—and still are—studied by several researchers, both empirically and theoretically. The pros and cons of performance metrics are also the subject of study by researchers.

Chicco and Jurman described the advantages of the Matthews correlation coefficient (i.e., ϕ) over FM and ACC [9]. They observed that ACC and FM can dangerously show overoptimistic inflated results, especially on imbalanced datasets. Instead, ϕ is considered a more reliable performance metric: they show the benefits of ϕ by explaining some mathematical properties, and then the qualities of ϕ in six synthetic use cases and in a real genomic scenario.

In a more recent paper, Lavazza and Morasca described analytically the relationship that links FM and ϕ (and ρ and EP) [10], thus providing the theoretical foundations that support the observations by Chicco and Yourman [9].

Chicco et al. also discussed why ϕ is more reliable than other performance metrics for the evaluation of binary classifiers. Specifically, they addressed balanced accuracy, bookmaker informedness, and markedness [19], Cohen's Kappa and Brier score [20] and the diagnostic odds ratio [21].

Some preliminary study concerning the relationship between the ROC AUC and ϕ was studied by Lavazza et al. [34]. Specifically, they studied the relationship when ϕ is constant for all the threshold values corresponding to the ROC curve points.

VI. ADDITIONAL CONSIDERATIONS

In this section, we deal with some additional aspects of classifiers and their evaluation. Though not completely related to the previous sections, they show how our work could be expanded and be the object of further work.

A. DEALING WITH COSTS

Usually, the binary classifiers that we build are not perfect, i.e., they involve a number of incorrect classifications (false positives and false negatives). Since incorrect classifications are associated to some cost, a reasonable way for evaluating binary classifiers consists in computing the cost of incorrect classifications.

Misclassification cost (MC) was used to evaluate the practical usefulness of binary classifiers, in different fields [22], [23], [24], [25]. It is defined as follows:

$$MC = C_{FN}FN + C_{FP}FP$$

where C_{FN} and C_{FP} are the costs per false negative and false positive, respectively.

MC shows how to take into account the cost of false positives and false negatives explicitly. However, other cost models could be adopted, depending on the specific context.

In a way, cost is the ultimate performance metric: if the values of C_{FN} and C_{FP} are known, then there is no need to use other performance metrics.

It can be argued that the exact knowledge of C_{FN} and C_{FP} is hardly ever possible. While that is true, we observe that such exact knowledge is in general not really necessary. For instance, when comparing the performances of two binary classifiers, what one needs to know is whether the ratio between C_{FN} and C_{FP} is above or below a specified value. It is much more likely that this information is available.

B. VARIABLE THRESHOLD CLASSIFICATION

The considerations reported in the above sections concern proper binary classifiers, which, given a subject, always classify it as either positive or negative

There are several situations when a binary classifier is obtained based on other model or knowledge. For instance, a function $f(X)$ that estimates the probability p that a subject X is positive (hence, the probability 1-p that X is negative) can be used to define a binary classifier, e.g., by considering X positive if and only if $f(X) > t$, where t is a threshold that can be determined in different ways. In cases like this, one is often interested in evaluating the classifications that can be obtained via different values of t. This is especially the case then the "best" value of t is not known, while it is possible to devise a reasonable range for t.

In these cases, it is common practice to represent the performance of the family of classifiers obtained by varying t via a ROC curve [26]. ROC curves represent a different concept with respect to the performance metrics discussed in the previous sections, since a ROC curve represents different performances, depending on the threshold chosen. In fact, every point of a ROC curve corresponds to a specific confusion matrix.

The Area Under the Curve (AUC) is often used to summarize the performance of the family of classifiers into a single number. Although it is widely used, the AUC has a few drawbacks. Among these is the fact that AUC is computed for the entire ROC curve, i.e., for all the possible values of the threshold t, even though some values of t are not suitable for being used in practice [8], [27], [28], [29]. To overcome this problem, it has been proposed to use partial AUC, which considers only a section of the curve and a portion of the ROC space [8], [30], [31], [32].

It is known that AUC is the mean TPR of the classifiers obtained form all the possible values of threshold t [33]. Some research concerning the relationship between AUC and ϕ is ongoing [34].

C. MULTICLASS CLASSIFICATION

Many diagnostic activities involve multiclass classification, i.e., subjects are classified into three or more classes. For instance, several tests for dementia detection consider three classes: mild cognitive impairment (MCI), dementia, and normal. In these cases, when evaluating a classifier, one should consider that classes are not equally distant from each other. For instance, given a subject that has dementia, an incorrect classification of the subject as normal is a bigger error than classifying the subject as having MCI.

Therefore, most performance metrics, namely those involving false positives and false negatives, cannot be applied as-is to multiclass classification. Instead, performance metrics should involve some sort of weighting, to account for the distance among classes.

The considerations reported in the previous sections are conceptually applicable to multiclass classification as well. Specifically, reporting the confusion matrix is even more important than for binary classification, since each element of the matrix can be assigned a difference importance (or cost), which implicitly determines the distance among classes. For instance, using the example mentioned above, in the confusion matrix you have (among others) two cells, representing the subjects having dementia and classified, respectively, as normal or having MCI. By assigning a weight or a cost to these cells, you are also defining the distance between MCI and dementia and normal and dementia.

VII. CONCLUSION

Several performance metrics have been proposed to evaluate binary classifiers and diagnostic tests. These metrics have the merit of summarizing performance (which is fully represented by a confusion matrix) into a single number, which appears easier to understand and makes comparisons straightforward.

However, most performance metrics make implicit assumptions and hide important characteristics of classifiers, also in relation to how they are evaluated. For instance, F-measure ignores the true negatives, while Accuracy and ϕ assume that false positives and false negatives have the same cost. Similarly, the relationships that link performance metrics to each other or to the prevalence of the dataset are largely ignored.

In this paper, we have discussed a few properties of Accuracy and F-measure, and we have shown how the improper or naive usage of these performance metrics can

lead to partial or incorrect evaluations. We have also proposed a technique to reinterpret published evaluations based on F-measure, Accuracy, and other metrics (including Precision, Recall, etc.). The proposed technique is based on reconstructing the confusion matrix from the available performance metrics. In fact, the confusion matrix supports the evaluation of performance from different points of view (e.g., the cost of incorrect diagnoses) and according to a specific context.

When applied to a few research papers, the proposed re-interpretation technique showed that some papers present overoptimistic interpretations of the obtained results, while others draw erroneous or incomplete conclusions from the obtained results, especially when comparing the latter with previously published results.

Finally, we recommend that new evaluations be provided with the complete confusion matrices, so that readers can formulate personal evaluations, based on specific contexts and priorities.

APPENDIX A
PROBLEMS WITH CASE 3

The original paper by Chittora et al. [14] provides Accuracy (ACC), Precision (PPV) and Recall (TPR) metrics. We derived the confusion matrices by solving the following set of equations, three of which are definitions, and the fourth is a property of confusion matrices:

$$\begin{cases} tp + tn = ACC \\ \dfrac{tp}{tp+fn} = TPR \\ \dfrac{tp}{tp+fp} = PPV \\ tp + tn + fp + fn = 1 \end{cases}$$

The system of equations can be solved as follows:

$$\begin{cases} tn = ACC - tp \\ fn = tp\,\dfrac{1-TPR}{TPR} \\ fp = tp\,\dfrac{1-PPV}{PPV} \\ tp + ACC - tp + tp\,\dfrac{1-PPV}{PPV} + tp\,\dfrac{1-TPR}{TPR} = 1 \end{cases}$$

Last equation yields

$$tp = \frac{1-ACC}{\frac{1-PPV}{PPV} + \frac{1-TPR}{TPR}}$$

so, tp can be computed, because ACC, PPV, and TPR are known. The value of tp can then be used to compute tn, fn and fp.

Based on the obtained results, we spotted two problems, described below.

A. VARIABLE PREVALENCE OF TEST SETS

We computed the prevalence of the test sets as $\rho = ap = tp + fn$ (being tp and fn relative values, ap is the rate of actual positives in the test population).

When considering the confusion matrix derived from Table 4 of the original paper (i.e., the table that reports the results obtained with no feature selection and without applying SMOTE), we have that

- Methods ANN, C5.0, LR, CHAID, LSVM and RT were tested with test sets having ρ in the [0.362, 0.375] range.
- Method KNN was tested with a test set having $\rho = 0.504$.

This observation triggers an obvious question: why was method KNN tested using an almost perfectly balanced test set, while the other methods were tested using definitely more imbalanced data? Moreover, why not use the same test set for all methods, given that adopting a unique test dataset would make the comparison of results from different methods more reliable?

B. INCONSISTENCY

Table 5 of the original paper (reporting results achieved with CFS and without SMOTE) reports the following performance metrics for KNN: PPV (Precision)=0.9705, TPR (Recall)=1, ACC (Accuracy)=0.5317.

$TPR = \frac{tp}{tp+fn} = 1$ implies that $fn = 0$. $ACC = tp + tn = (ap - fn) + (an - fp) = 1 - fn - fp$. Being $fn = 0$, we have $ACC = 1 - fp = 0.5317$, hence $fp = 1 - 0.5317 = 0.4683$. Finally, $PPV = \frac{tp}{tp+fp} = 0.9705$, hence $tp = \frac{fp\ PPV}{1-PPV} = \frac{0.4683 \cdot 0.9705}{1-0.9705} \approx 15.4$, which is not possible: being a relative value, tp cannot be greater than one.

We found this type of inconsistency for the performance metrics concerning LR and KNN in Tables 5, 6, and 7 of the original paper (LR and KNN do not appear in the following Tables).

As already noted, in the conclusions of the study ("*Logistic and KNN did not give suitable results*") might have been affected by the inconsistencies described above.

REFERENCES

[1] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artif. Intell. Med.*, vol. 128, Jun. 2022, Art. no. 102289, doi: 10.1016/j.artmed.2022.102289.

[2] L. Devnath, P. Summons, S. Luo, D. Wang, K. Shaukat, I. A. Hameed, and H. Aljuaid, "Computer-aided diagnosis of coal workers' pneumoconiosis in chest X-ray radiographs using machine learning: A systematic literature review," *Int. J. Environ. Res. Public Health*, vol. 19, no. 11, p. 6439, May 2022, doi: 10.3390/ijerph19116439.

[3] R. Kaur, J. Anupama Ginige, and O. Obst, "A systematic literature review of automated ICD coding and classification systems using discharge summaries," 2021, *arXiv:2107.10652*.

[4] C. J. van Rijsbergen, *Information Retrieval*. Malaysia: Butterworth, 1979.

[5] J. Hernandez-Orallo, P. A. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *J. Mach. Learn. Res.*, vol. 13, pp. 2813–2869, Oct. 2012.

[6] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.

[7] J. Yao and M. Shepperd, "The impact of using biased performance metrics on software defect prediction research," *Inf. Softw. Technol.*, vol. 139, Nov. 2021, Art. no. 106664, doi: 10.1016/j.infsof.2021.106664.

[8] S. Morasca and L. Lavazza, "On the assessment of software defect prediction models via ROC curves," *Empirical Softw. Eng.*, vol. 25, no. 5, pp. 3977–4019, Sep. 2020, doi: 10.1007/s10664-020-09861-4.

[9] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[10] L. Lavazza and S. Morasca, "Comparing ϕ and the F-measure as performance metrics for software-related classifications," *Empirical Softw. Eng.*, vol. 27, no. 7, pp. 1–37, Dec. 2022, doi: 10.1007/s10664-022-10199-2.
[11] F. Salehian Matikolaie and C. Tadj, "On the use of long-term features in a newborn cry diagnostic system," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101889, doi: 10.1016/j.bspc.2020.101889.
[12] E. E.-D. Hemdan, W. El-Shafai, and A. Sayed, "CR19: A framework for preliminary detection of COVID-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications," *J. Ambient Intell. Humanized Comput.*, Feb. 2022, doi: 10.1007/s12652-022-03732-0.
[13] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104572, doi: 10.1016/j.compbiomed.2021.104572.
[14] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasinski, L. Jasinski, R. Gono, E. Jasinska, and V. Bolshev, "Prediction of chronic kidney disease—A machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021, doi: 10.1109/ACCESS.2021.3053763.
[15] J. Shepherd, K. Tu, J. Young, J. Chishtie, B. C. Craven, R. Moineddin, and S. Jaglal, "Identifying cases of spinal cord injury or disease in a primary care electronic medical record database," *J. Spinal Cord Med.*, vol. 44, no. 1, pp. S28–S39, Sep. 2021, doi: 10.1080/10790268.2021.1971357.
[16] S. A. V. Nieuwenburg, M. C. Mommersteeg, L. M. M. Wolters, A. J. van Vuuren, N. Erler, M. P. Peppelenbosch, G. M. Fuhler, M. J. Bruno, E. J. Kuipers, and M. C. W. Spaander, "Accuracy of H. Pylori fecal antigen test using fecal immunochemical test (FIT)," *Gastric Cancer*, vol. 25, no. 2, pp. 375–381, Mar. 2022, doi: 10.1007/s10120-021-01264-8.
[17] A. Gupta, P. Sheth, and P. Xie, "Neural architecture search for pneumonia diagnosis from chest X-rays," *Sci. Rep.*, vol. 12, no. 1, Jul. 2022, doi: 10.1038/s41598-022-15341-0.
[18] N. Hasan and Y. Bao, "Understanding current states of machine learning approaches in medical informatics: A systematic literature review," *Health Technol.*, vol. 11, no. 3, pp. 471–482, May 2021, doi: 10.1007/s12553-021-00538-6.
[19] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, no. 1, pp. 1–22, Feb. 2021, doi: 10.1186/s13040-021-00244-z.
[20] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.
[21] D. Chicco, V. Starovoitov, and G. Jurman, "The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment," *IEEE Access*, vol. 9, pp. 47112–47124, 2021, doi: 10.1109/ACCESS.2021.3068614.
[22] T. M. Khoshgoftaar and E. B. Allen, "Classification of fault-prone software modules: Prior probabilities, costs, and model evaluation," *Empirical Softw. Eng.*, vol. 3, no. 3, pp. 275–298, 1998, doi: 10.1023/A:1009736205722.
[23] S. Nanda and P. Pendharkar, "Linear models for minimizing misclassification costs in bankruptcy prediction," *Int. J. Intell. Syst. Accounting, Finance Manage.*, vol. 10, no. 3, pp. 155–168, 2001, doi: 10.1002/isaf.203.
[24] S. Lombardi, M. Gorgoglione, and U. Panniello, "The effect of context on misclassification costs in e-commerce applications," *Exp. Syst. Appl.*, vol. 40, no. 13, pp. 5219–5227, Oct. 2013, doi: 10.1016/j.eswa.2013.03.009.
[25] Y. Xiong and R. Zuo, "Effects of misclassification costs on mapping mineral prospectivity," *Ore Geol. Rev.*, vol. 82, pp. 1–9, Apr. 2017, doi: 10.1016/j.oregeorev.2016.11.014.
[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[27] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecology Biogeography*, vol. 17, no. 2, pp. 145–151, Mar. 2008, doi: 10.1111/j.1466-8238.2007.00358.x.
[28] S. Mallett, S. Halligan, M. Thompson, G. S. Collins, and D. G. Altman, "Interpreting diagnostic accuracy studies for patient care," *BMJ*, vol. 345, no. jul02 1, pp. e3999–e3999, Jul. 2012, doi: 10.1136/bmj.e3999.
[29] A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, and A. Holzinger, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 329–341, Jan. 2023, doi: 10.1109/TPAMI.2022.3145392.
[30] D. K. McClish, "Analyzing a portion of the ROC curve," *Med. Decis. Making*, vol. 9, no. 3, pp. 190–195, Aug. 1989, doi: 10.1177/0272989X8900900307.
[31] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, no. 3, pp. 745–750, Dec. 1996, doi: 10.1148/radiology.201.3.8939225.
[32] H. Yang, K. Lu, X. Lyu, and F. Hu, "Two-way partial AUC and its properties," *Stat. Methods Med. Res.*, vol. 28, no. 1, pp. 184–195, Jan. 2019, doi: 10.1177/0962280217718866.
[33] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, Oct. 2009, doi: 10.1007/s10994-009-5119-5.
[34] L. Lavazza, S. Morasca, and G. Rotoloni, "On the reliability of the area under the ROC curve in empirical software engineering," in *Proc. 27th Int. Conf. Eval. Assessment Softw. Eng.*, Oulu, Finland, 2023.

LUIGI LAVAZZA (Senior Member, IEEE) received the Laurea degree in electronic engineering from Politecnico di Milano, in 1984.

From 1985 to 1990, he worked in industry. In 1990, he joined Cefriel (https://www.cefriel.com/?lang=en). From 1996 to 2005, he was a Research Assistant with Politecnico di Milano. Since 2005, he has been an Associate Professor with the University of Insubria, Varese, Italy. His research interests include empirical software engineering, software metrics and software quality evaluation, software project management and effort estimation, software process modeling, measurement and improvement, and open source software. He was involved in several international research projects. He served as a reviewer for EU projects. He has also served on the PC member for several of international software engineering conferences and in the editorial board for international journals. He is the coauthor of over 170 scientific articles, published in international journals, and in the proceedings of international conferences or in books.

Prof. Lavazza is a member of the ACM and a fellow of the IARIA.

SANDRO MORASCA (Member, IEEE) was an Associate Professor and an Assistant Professor with Politecnico di Milano, Milan and Como, Italy, and a Faculty Research Assistant and later a Visiting Scientist with the Department of Computer Science, University of Maryland, College Park. He is currently a Professor in computer science with Dipartimento di Scienze Teoriche e Applicate, Universitá degli Studi dell'Insubria, Como and Varese, Italy. He has been actively carrying out research in empirical software engineering, software quality, machine learning, software verification, open source software, web services, and specification of concurrent and real-time software systems. He has published over 40 journal articles and over 100 conference papers. He has been involved in several national and international projects and has served on the program committees and editorial boards of international software engineering conferences and journals.

Open Access funding provided by 'Università degli Studi dell'Insubria' within the CRUI CARE Agreement