IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 4, 2021, accepted March 20, 2021, date of publication March 24, 2021, date of current version March 31, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068614

The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment

DAVIDE CHICCO[1], VALERY STAROVOITOV[2], AND GIUSEPPE JURMAN[3]
[1]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON M5T 3M7, Canada
[2]National Academy of Sciences of Belarus, 220012 Minsk, Belarus
[3]Fondazione Bruno Kessler, 38123 Trento, Italy

Corresponding author: Davide Chicco (davidechicco@davidechicco.it)

The work of Valery Starovoitov was supported in part by the Belarusian Republican Foundation for Fundamental Research under Grant F20RA-014.

ABSTRACT To assess the quality of a binary classification, researchers often take advantage of a four-entry contingency table called *confusion matrix*, containing true positives, true negatives, false positives, and false negatives. To recap the four values of a confusion matrix in a unique score, researchers and statisticians have developed several rates and metrics. In the past, several scientific studies already showed why the Matthews correlation coefficient (MCC) is more informative and trustworthy than confusion-entropy error, accuracy, F_1 score, bookmaker informedness, markedness, and balanced accuracy. In this study, we compare the MCC with the diagnostic odds ratio (DOR), a statistical rate employed sometimes in biomedical sciences. After examining the properties of the MCC and of the DOR, we describe the relationships between them, by also taking advantage of an innovative geometrical plot called *confusion tetrahedron*, presented here for the first time. We then report some use cases where the MCC and the DOR produce discordant outcomes, and explain why the Matthews correlation coefficient is more informative and reliable between the two. Our results can have a strong impact in computer science and statistics, because they clearly explain why the trustworthiness of the information provided by the Matthews correlation coefficient is higher than the one generated by the diagnostic odds ratio.

INDEX TERMS Matthews correlation coefficient, diagnostic odds ratio, binary classification, confusion matrix, supervised machine learning, confusion tetrahedron.

I. INTRODUCTION

In scientific research, the goal of many studies is often to correctly predict elements that can have two conditions, like for example individuals that can be categorized as sick or healthy, or as likely to survive or at risk of death. In computational statistics and machine learning, these problems are commonly called *binary classifications*, and the two possible conditions are usually coded as 1 and 0, or true and false, or positive and negative.

Since both the elements of the ground truth dataset and the predicted elements can be belong to each of the two classes, the binary classification can generate four distinct categories, usually recapped in to a table called two-class *confusion matrix*. A positive element correctly predicted as positive is called true positive (TP), and a negative element correctly labeled negative is called true negative (TP). Since the classifier could have made some mistakes in the classification, other two categories are possible: a negative element wrongly predicted positive is called false positive (FP), while a positive element mistakenly classified as negative is called false negative (FN).

These four categories are the elements of a traditional 2×2 confusion matrix, a statistical table that is extremely common in studies involving applied machine learning and statistics. Several rates that summarize the four categories of the confusion matrix exist nowadays; none of them, however, has reached consensus in the computer science.

We propose the Matthews correlation coefficient (MCC) [1] as potential candidate for this role. In the past, we showed

The associate editor coordinating the review of this manuscript and approving it for publication was Victor S. Sheng.

why the MCC is more informative and truthful than confusion entropy (CEN) error [2], [3], than F_1 score [4] and accuracy [5], [6], and then bookmaker informedness [7], markedness [7], and balanced accuracy [7], [8]. In the present study, we compare this rate with the diagnostic odds ratio (DOR), another metric which is sometimes employed in the biomedical sciences [9].

In particular, a recent study by Rácz *et al.* [10] described a multi-level comparison of a number of performance metrics for binary classification. Surprisingly to us, this article affirmed that DOR resulted being the top performing metric, together with markedenss (MK), in the analyses they described. Starting from this remarkable result, we decided to study the diagnostic odds ratio and its similarities, differences, and relationships with the Matthews correlation coefficient. We already compared MCC and MK in another study [8].

We organize the rest of the article as follows. After this Introduction and its literature review, we report the details and the information about the two analyzed rates and describe the relationships between them (section II). We then describe some use cases where the MCC and the DOR produce discordant outcomes (section III), and finally outline some discussion and conclusions (section IV).

A. LITERATURE REVIEW

The Matthews correlation coefficient was originally introduced by Brian W. Matthews in a biochemistry article in the 1970s [1]. In early 2000s, the MCC became popular in the computer science community thanks to a highly-impactful review published by Pierre Baldi and colleagues [11].

As the MCC acquired popularity, studies comparing this metric with other rates started to appear in the scientific literature. Jurman and colleagues [3], in fact, published a study comparing this coefficient with cross-entropy (CEN) error.

Since then, an increasing number researchers has employed this rate to assess binary classifications, even if its usage has often been less frequent than the usage of other rates, such as accuracy and F_1 score.

In 2017, the MCC obtained new notoriety: Bourghorbel *et al.* [12] published a study based on the benefits of assessing the performances of several machine learning methods on imbalanced datasets with MCC.

Few years later, two of us authors continued the series of the comparisons between the MCC and other rates, by explaining the advantages of the Matthews correlation coefficient first over accuracy and F_1 score [5], and then over balanced accuracy, bookmaker informedness, and markedness [8].

In the recent years, scientists employed the MCC for several applications in biomedical sciences. Abhishek and Hamarneh [13], for example, used it for the segmentation skin lesion in pathological images, while Saqlain *et al.* [14] took advantage of it for heart failure diagnosis. Additionally, the MCC has been the key-metric of multiple studies related to the prediction of patient prognosis and diagnosis from electronic health records [15], [16]. In software engineering, instead, Yao and Martin [17] published an article explaining the benefits of using the Matthews correlation coefficient in software defection prediction.

Even if it is unclear when the diagnostic odds ratio (DOR) was first introduced in the scientific literature, it is known that its popularity raised since the publication of a study by Glas and colleagues [9], where the authors highlight some potential assets of this rate, from their point of view. After that, multiple researchers employed the DOR to assess binary predictions in their studies, mainly in the biomedical sciences.

Multiple researchers took advantage of the diagnostic odds ratio in meta-analyses and systematic reviews. Doust and colleagues [18], for example, employed the DOR for a systematic review of articles predicting heart failures assessed with natriuretic peptides. The DOR was employed also in a systematic review and a meta-analysis study to predict sepsis diagnosis carried out by Tang *et al.* [19]. In a hepatological meta-analysis, Tsochatzis *et al.* [20] took advantange of the DOR to detect the fibrosis severity in chronic liver diseases. In particular, they used the DOR to evaluate elastography performance in each study and disease stage. In another hepatology meta-analysis, Yoon and colleagues [21] used the DOR to assess the capability of IgG4 immunohistochemistry to detect autoimmune pancreatitis. Picot *et al.* [22] performed a meta-analysis and a systematic review on the usage of loop-mediated isothermal amplification (LAMP) to diagnose malaria.

The diagnostic odds ratio has not only been employed in meta-analyses and a systematic reviews in the past, but also on studies presenting single cohort and single-experiment biomedical results. Tedeschini and coauthor [23], for example, published a DOR-based study on mental health.

In a multi-disciplinary study crossing mental health, signal processing, and robotics, Kokot *et al.* [24] utilized computational intelligence methods to diagnose autism in children, and they measured their results through the diagnostic odds ratio. Children were also the patients involved in a biomedical engineering study of Aungaroon and colleagues [25], who performed SISCOM tests to detect the epileptogenic zone in patients diagnosed with epilepsy, and measured their results by the DOR.

In a dermatology article, Kamyab-Hesari *et al.* [26] reported the DOR-measured results of the detection of alopecia from scalp biopsies made by dermapathologists. Nwoye and coauthors [27] utilized machine learning modesl to diagnose schizophrenia from electronic health records, and measured their performance through the diagnostic odds ratio.

Moving from biomedicine to environmental sciences, we report the article of Rahmati and coauthors [28] which employs several statistical rates, including the DOR, to computationally evaluate geo-environmental models.

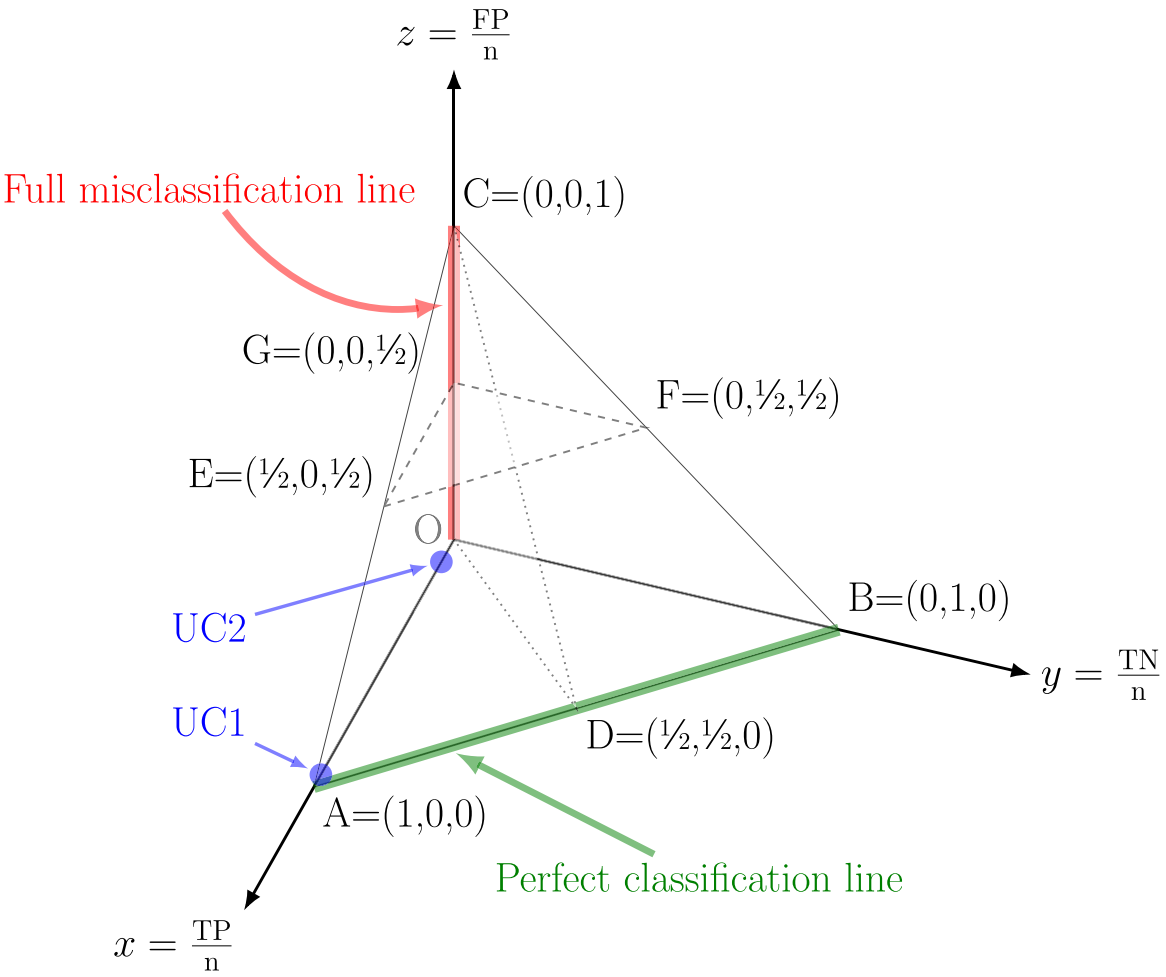


FIGURE 1. The confusion tetrahedron. The full misclassification line includes all the (classes of) confusion matrices $\begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}$, while the perfect classification line corresponds to matrices $\begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$, for $0 \leq \alpha, \beta \leq 1$. Blue dots mark the position of the use cases UC1 and UC2.

The scientific literature also includes some articles describing the properties and the limitations of the DOR. In a review on statistical rates, Šimundić [29] included the diagnostic odds ratio, and briefly explained its properties. Böhning and colleagues [30] explained why the DOR should not be used to find the best cut-off threshold for a diagnostic test. To the best of our knowledge, no study comparing the diagnostic odds ratio and the Matthews correlation coefficient exists in literature; we fill this gap with the present article.

II. METHODS

In this section, we first introduce the Matthews correlation coefficient (subsection II-A), then we introduce the diagnostic odds ratio (subsection II-B), and finally we report and discuss some statistical correlations between these two rates (subsection II-C). We indicate FP for false positives, FN for false negatives, TP for true positives, and TN for false negatives.

A. MATTHEWS CORRELATION COEFFICIENT (MCC)

Consider a generic confusion matrix $M = \begin{pmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{pmatrix}$ and let $n = \text{TP} + \text{TN} + \text{FP} + \text{FN}$ be the total number of samples.

1) DEFINITION

The Matthews correlation coefficient (MCC) [1], [3], [5] is the 2×2 case of ϕ coefficient [31], and is defined as:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{1}$$

$$(\text{worst value} = -1; \text{best value} = +1)$$

2) RANGE

MCC ranges between -1 and $+1$, with -1 for perfect misclassification ($\text{TP} = \text{TN} = 0$) and 1 for perfect classification ($\text{FP} = \text{FN} = 0$); $\text{MCC} = 0$ indicates random classification ($\text{TP} \cdot \text{TN} = \text{FP} \cdot \text{FN}$).

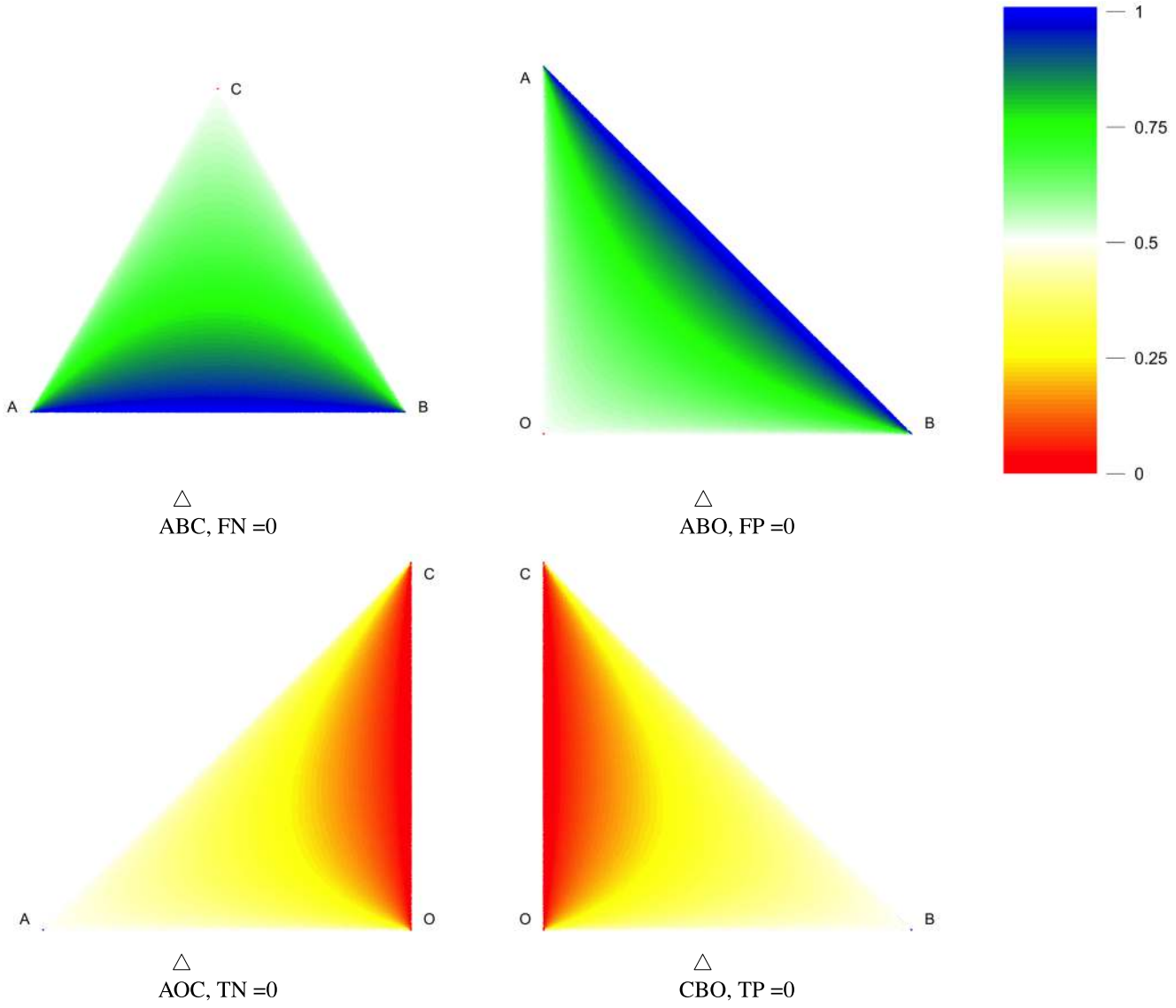


FIGURE 2. nMCC on the faces of the confusion tetrahedron. On the same triangles, nDOR is uniformly maximally blue (nDOR = 1) on the whole of $\overset{\triangle}{ABC}$ and $\overset{\triangle}{ABO}$, and uniformly maximally red (nDOR = 0) on the whole of $\overset{\triangle}{AOC}$ and $\overset{\triangle}{CBO}$. All graphs include 1M points. To avoid distortion in the planar projection of $\overset{\triangle}{ABC}$ we used the isometric mapping induced by barycentric coordinates.

3) UNDEFINED CASES AND SOLUTIONS

MCC is undefined when a whole row or column of M is zero. However, MCC can be naturally extended to cover all these cases:

- MCC $= +1$ for $M = \begin{pmatrix} \text{TP} & 0 \\ 0 & 0 \end{pmatrix}$ with TP > 0 or $M = \begin{pmatrix} 0 & 0 \\ 0 & \text{TN} \end{pmatrix}$ with TN > 0;
- MCC $= 0$ for $M \in \left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix}, \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ b & a \end{pmatrix}, \begin{pmatrix} 0 & b \\ 0 & a \end{pmatrix} \right\}$ for $a, b > 0$;
- MCC $= -1$ for $M = \begin{pmatrix} 0 & \text{FN} \\ 0 & 0 \end{pmatrix}$ with FN > 0 or $M = \begin{pmatrix} 0 & 0 \\ \text{FP} & 0 \end{pmatrix}$ with FP > 0.

4) NORMALIZATION

For comparison's purposes, the normalized version, ranging in [0, 1] can be used:

$$\text{nMCC} = \frac{\text{MCC} + 1}{2} \tag{2}$$

(worst value $= 0$; best value $= 1$)

where the values MCC $= \{-1, 0, +1\}$ are mapped to nMCC $= \{0, 0.5, 1\}$, respectively.

5) NOTE

MCC has its roots in the ϕ coefficient, introduced by Pearson, Boas and Yule independently in the early 90's as a measure of association [31]–[34] to express the linear correlation of an underlying bivariate discrete distribution between

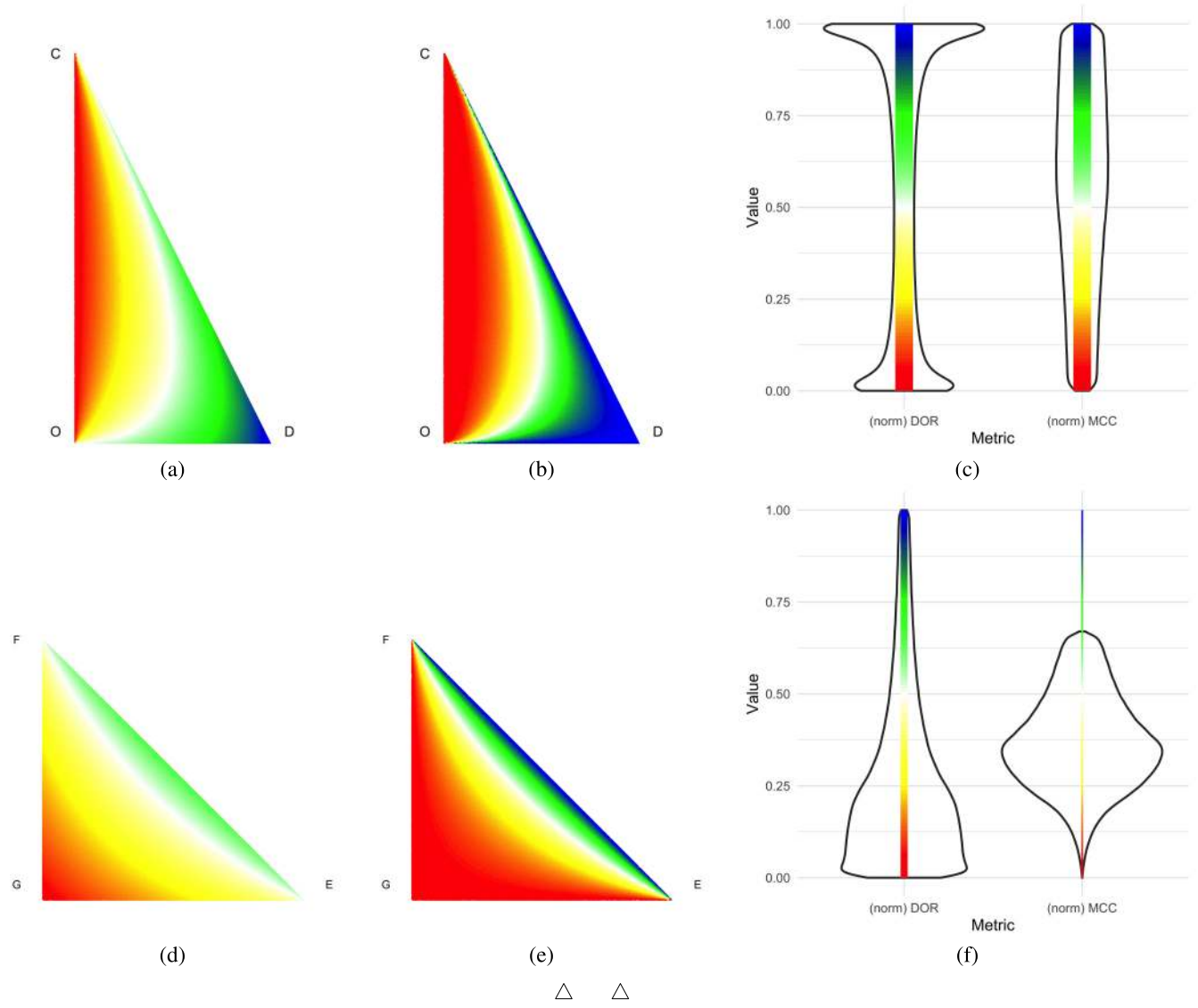


FIGURE 3. nMCC (a), (d) and nDOR (b), (e) on the triangles $\overset{\triangle}{\text{DOC}}$, $\overset{\triangle}{\text{GEF}}$ and (c), (f) violin plots of the corresponding values of nMCC and nDOR.

n	PCC
5	0.7495
10	0.8860
15	0.9157
20	0.9287
25	0.9359
50	0.9494
75	0.9535
100	0.9550

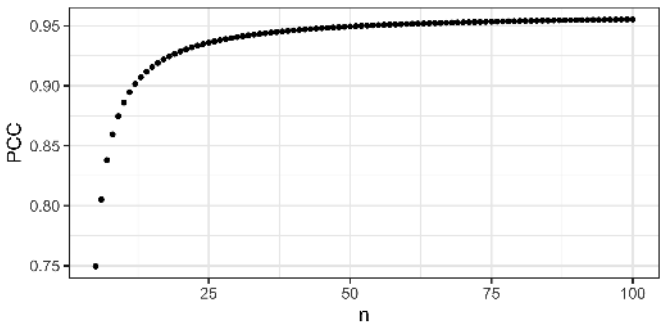


FIGURE 4. Pearson correlation between nMCC and nDOR for the complete subset of confusion matrices with FP, FN > 0 for datasets with n samples, $5 \leq n \leq 100$.

two variables. In details, both ϕ and MCC have a simple relation with the chi-square statistic for a contingency table, namely $\phi = |\text{MCC}| = \sqrt{\frac{\chi^2}{n}}$, where n is the total number of observations. Several other statistical metrics stem from variants of the ϕ coefficient: among them, the Cramér's V (or Cramér's χ_C) emerges as the most relevant. Introduced by Cramér in [35], V is also defined in terms of the chi-square

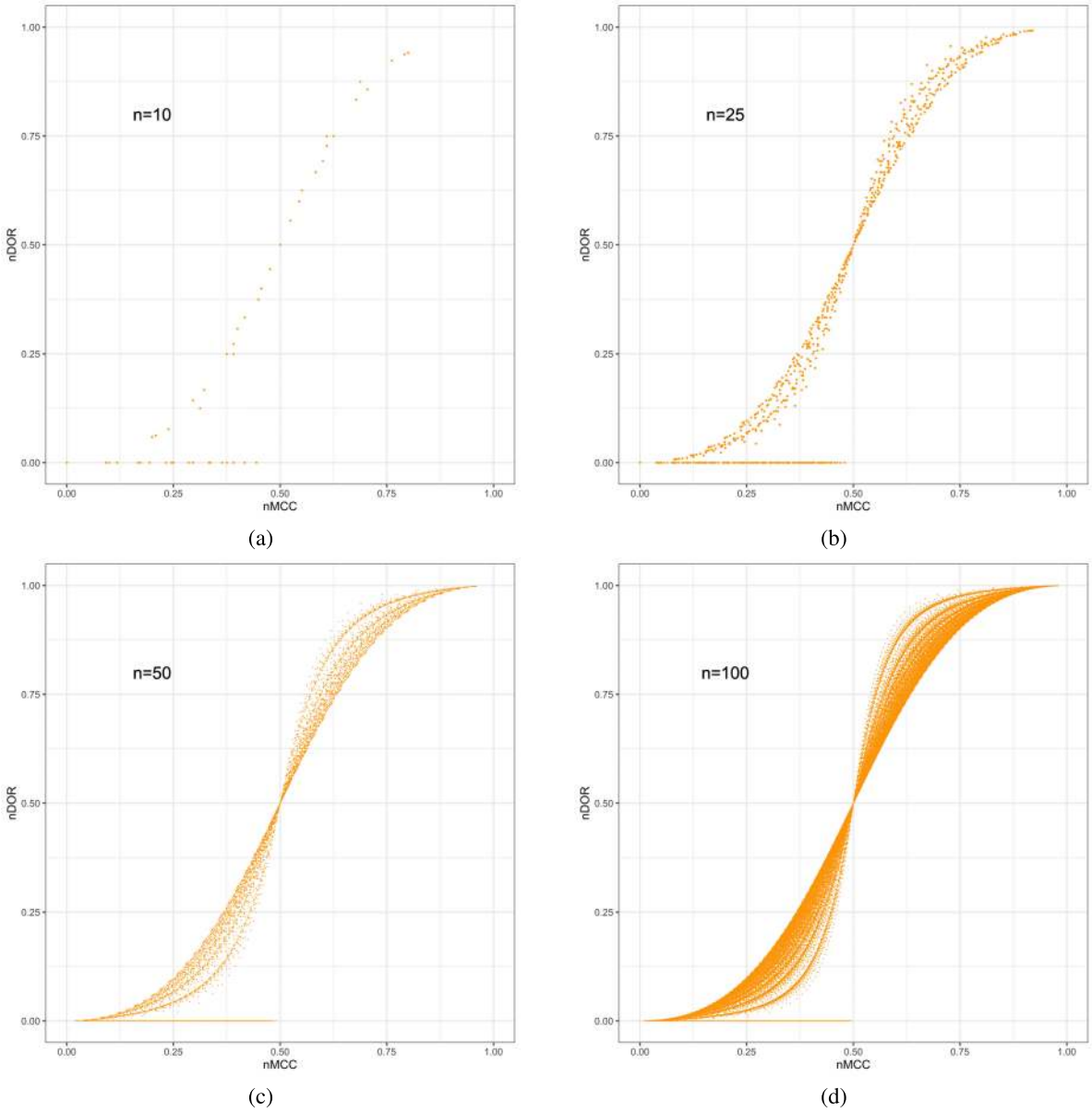


FIGURE 5. Scatterplot of (nMCC, nDOR) points corresponding to all confusion matrices with a) $n = 10$, b) $n = 25$, c) $n = 50$ and d) $n = 100$ samples with FP, FN > 0. Point size in panels a) and b) are larger than point size in panels c) and d).

statistics as $V = \sqrt{\dfrac{\chi^2/n}{\min(k-1, r-1)}}$, for k, r the number of data columns and data rows, respectively. By definition, on the 2×2 case a contingency table coincides with a confusion matrix, and thus the equality $\mathrm{MCC}^2 = \phi^2 = V^2$ holds [36], the three measures only numerically differing by the association sign, but having different statistical interpretations.

B. DIAGNOSTIC ODDS RATIO (DOR)

1) DEFINITION

The diagnostic odds ratio (DOR) [9], [37] is defined as:

$$\mathrm{DOR} = \frac{\mathrm{TP} \cdot \mathrm{TN}}{\mathrm{FP} \cdot \mathrm{FN}} = \frac{\mathrm{TPR} \cdot \mathrm{TNR}}{(1 - \mathrm{TPR}) \cdot (1 - \mathrm{TNR})} \quad (3)$$

(worst value $= 0$; best value $= +\infty$)

2) RANGE

DOR ranges in $[0, +\infty)$, being lower bounded by zero in the cases where there are no correctly classified samples in one of the two classes (TP or TN or both are zero), while it is not upper limited; the higher the value, the better the classifier's performance. DOR $= 1$ indicates random classification.

3) UNDEFINED CASES AND SOLUTIONS

DOR is undefined when all samples in one (or both) class are correctly classified (FP or FN or both are zero). However, this behaviour is meaningful only in the case of perfect classification (FP $=$ FN $= 0$), while if only one of the two entries FP, FN vanishes, the DOR metric being undefined does not provide an interpretable indication about the classification task. A possible solution is adding $\frac{1}{2}$ to all entries of the

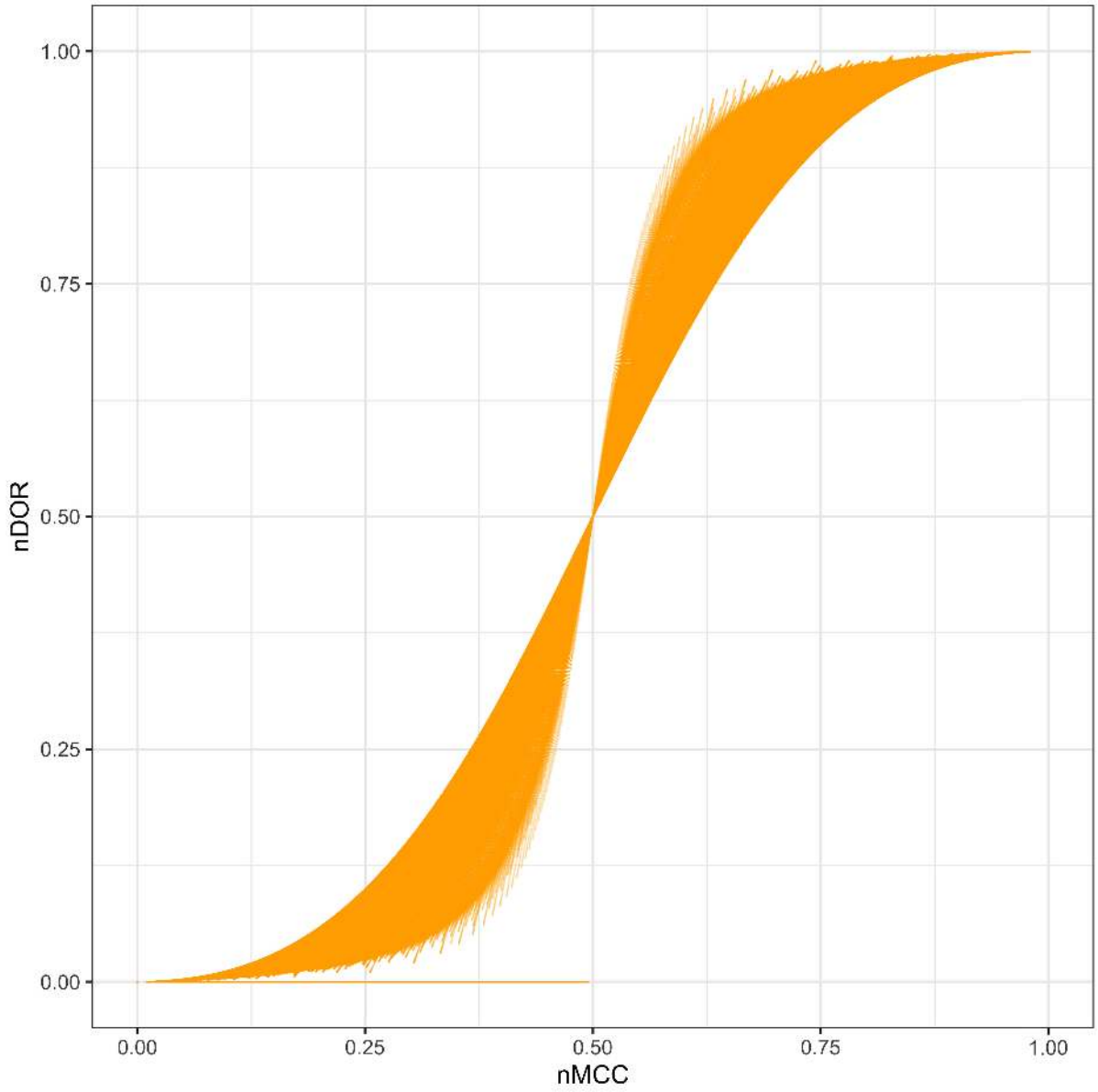


FIGURE 6. Scatterplot of 4,249,560 (nMCC, nDOR) points corresponding to all confusion matrices with $5 \leq n \leq 100$ samples with FP, FN > 0.

confusion matrix, either in the undefined cases [38], [39] (but this introduces a bias) or in all cases [40]: however, both solutions provide only an approximation of the true DOR measure. Alternatively, functions of the original DOR might be used, as in [37], where the authors considered the expression $\mathrm{DOR}^{\star} = \dfrac{\log_{10}(\log_{10}(\mathrm{DOR}))}{1.4}$.

4) NORMALIZATION

Again, for fair comparison's purposes, a normalized version ranging in [0, 1] can be used:

$$\mathrm{nDOR} = \frac{\mathrm{DOR}}{\mathrm{DOR} + 1} \tag{4}$$

where the values $\mathrm{DOR} = \{0, 1, +\infty\}$ are mapped to $\mathrm{nDOR} = \{0, 0.5, 1\}$, respectively. Furthermore, since

$$\mathrm{nDOR} = \frac{\mathrm{TP} \cdot \mathrm{TN}}{\mathrm{TP} \cdot \mathrm{TN} + \mathrm{FP} \cdot \mathrm{FN}} \tag{5}$$

nDOR is also defined when FP or FN (or both) vanishes, provided that $\mathrm{TP} \cdot \mathrm{TN} > 0$.

C. RELATIONS BETWEEN MCC AND DOR

1) INVARIANCE

- Both MCC and DOR (and thus also nMCC and nDOR) are invariant for class swapping P $\leftrightarrow$ N, meaning that exchanging the positives with the negatives would not change the value of the score.

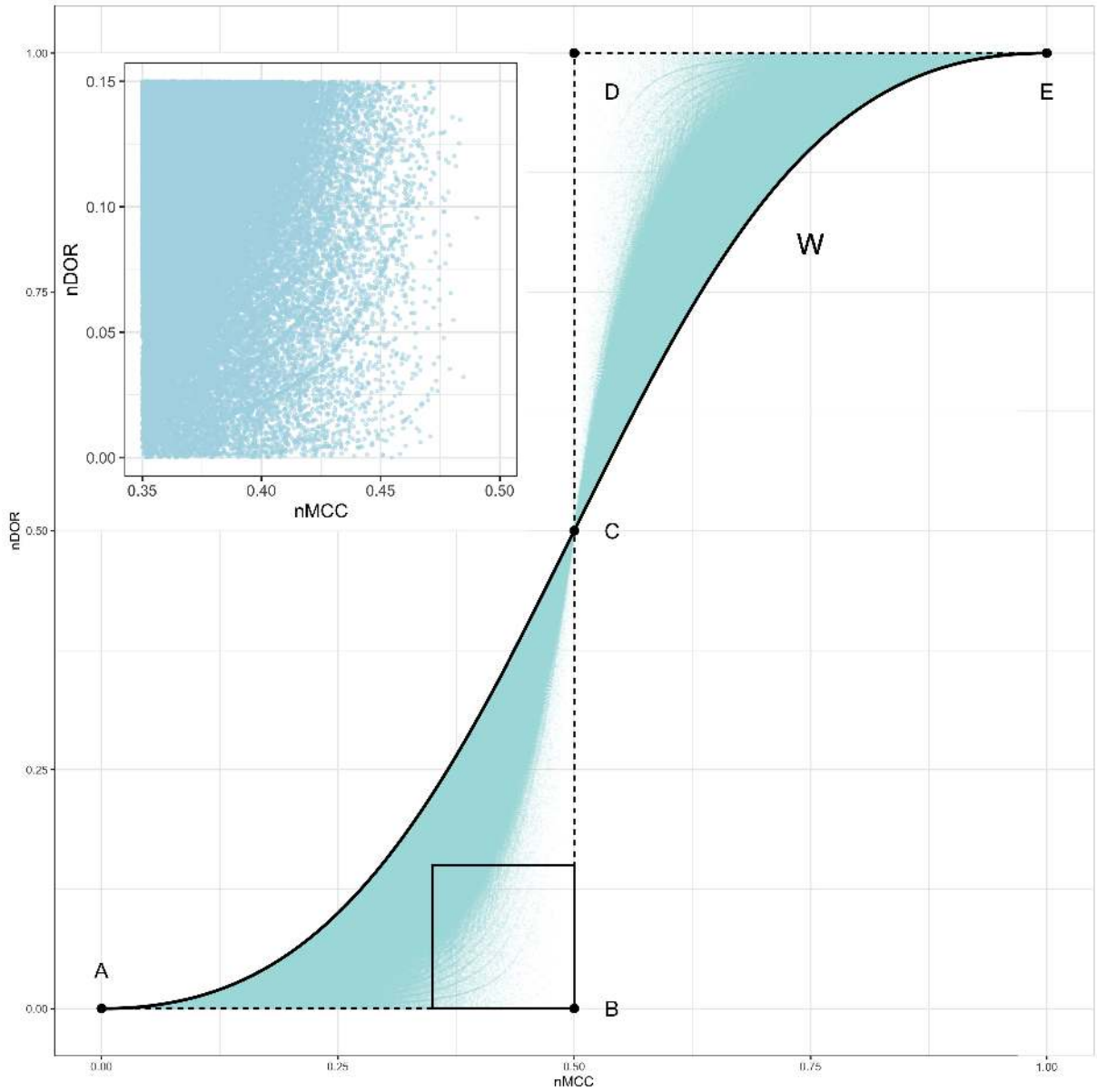


FIGURE 7. Scatterplot of (nMCC, nDOR) points corresponding to 9,249,560 confusion matrices with $n = \{5, 6, \ldots, 99, 10^2, 10^3, 10^4, 10^5, 10^6, 10^9\}$ samples with FP, FN > 0. In the inset, the zoom on the area marked by a square in the main plot. The red line marks the curve W, bounding the scattered areas. Note that in this scale, the use case UC1 is indistinguishable from point D, and the same happens with the use case UC2 and with point B (Figure 8).

- Both MCC and DOR (and thus also nMCC and nDOR) are invariant for dataset size scaling, that is, $\mathrm{MCC}\left(\begin{pmatrix}\mathrm{TP} & \mathrm{FN}\\ \mathrm{FP} & \mathrm{TN}\end{pmatrix}\right) = \mathrm{MCC}\left(\begin{pmatrix}\frac{k}{l}\cdot\mathrm{TP} & \frac{k}{l}\cdot\mathrm{FN}\\ \frac{k}{l}\cdot\mathrm{FP} & \frac{k}{l}\cdot\mathrm{TN}\end{pmatrix}\right)$ for $k \in \mathbb{N}$ and $l = \gcd(\mathrm{TP}, \mathrm{FN}, \mathrm{FP}, \mathrm{TN})$.

Interludio The invariance for size scaling highlights the relation of DOR with the classical odds ratio OR [41]. If we consider a scaled confusion matrix $nM = \begin{pmatrix}\frac{\mathrm{TP}}{n} & \frac{\mathrm{FN}}{n}\\ \frac{\mathrm{FP}}{n} & \frac{\mathrm{TN}}{n}\end{pmatrix}$ with entries in [0, 1] so that they can be viewed as probabilities, then $\mathrm{DOR}(M) = \mathrm{DOR}(nM) = \mathrm{OR}(nM)$, thus the two measures are mathematically equivalent, although their meaning is statistically different, being nM interpreted as a confusion matrix for DOR and as a contingency table for OR. Originally introduced in [42], OR compares the frequency of exposure to risk factors, mainly in survey research, in epidemiology [43] and in case-control studies to report clinical trials' results; statistically, OR can be viewed as a retrospective comparison of the impact of a given risk factor on two groups of individuals by estimating the chances of positive and negative outcomes. On the other hand, DOR is a measure of the effectiveness of a diagnostic test, that is a procedure performed to confirm or determine the presence of disease in an individual suspected of having a disease. A diagnostic test usually follows the

report of symptoms, or based on other medical test results. Thus, DOR is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease. This definition is the machine learning version of the evaluation of the frequency of exposure to risk factors in statistics, thus justifying the coinciding mathematical definition of DOR and OR.

2) VISUALIZING DOR AND MCC

Due to scaling invariance, it is possible to interpret both measures in term of the confusion tetrahedron (CT): CT is a geometric view of all possible (equivalence classes of) confusion matrices, normalized by the number of samples $\begin{pmatrix} \frac{TP}{n} & \frac{FN}{n} \\ \frac{FP}{n} & \frac{TN}{n} \end{pmatrix}$ so that all entries lie in $[0, 1]$ and their sum is one. The three axes represent TP, FP, TN while FN is not represented being dependent on the other three entries $FN = 1 - TP - TN - FN$. This is the equation of the plane in 3D space (Figure 1).

Due to the metric scaling invariance, each confusion matrix $M = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$ correspond to a normalized matrix $nM = \begin{pmatrix} \frac{TP}{n} & \frac{FN}{n} \\ \frac{FP}{n} & \frac{TN}{n} \end{pmatrix}$ in CT and the two matrices M, nM share the same MCC and the same DOR.

Coloring the points of CT according to the value of a scaling invariant performance measure μ (*e.g.*, as in the legend of Figure 2), we have a visual summary of the global behaviour of μ on the whole space of confusion matrices. Since a full 3D view of a μ-colored CT would hardly be human readable, we select some relevant 2D sections of CT. First we consider the four faces of CT, namely the triangles $\overset{\triangle}{ABC}$, $\overset{\triangle}{ABO}$, $\overset{\triangle}{AOC}$ and $\overset{\triangle}{CBO}$, corresponding to the sets of confusion matrices with $FN = 0$, $FP = 0$, $TN = 0$, and $TP = 0$, respectively. Equation 5, the value of nDOR is 1 on the whole triangles $\overset{\triangle}{ABC}$, $\overset{\triangle}{ABO}$ and 0 on the remaining faces $\overset{\triangle}{AOC}$ and $\overset{\triangle}{CBO}$, while nMCC has a more complex pattern (Figure 2). Same happens inside CT (Figure 3), although nMCC and nDOR are quite similar but not identical, as demonstrated by the boxplot in Figure 3(c,f).

3) RELATION WITH THE IMBALANCE RATIO

Consider the Imbalance Ratio (IR), defined as the ratio between the number of instances in the majority class and those in the minority class, originally introduced in [44]. Then, using the notation in [45], it follows that MCC and nMCC are dependent on IR, while DOR and nDOR are not.

4) CORRELATION

Consider now the confusion matrices $\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$ with $FP, FN > 0$, so that both MCC and DOR are defined. In particular, consider first the subset of all 4,249,560 matrices corresponding to classification task on small datasets, with $5 \leq n \leq 100$ and compute nMCC, nDOR and their Pearson correlation coefficient (PCC) for all these matrices. Correlation between the two metrics is very high (when $n > 53$ the PCC value is larger than 0.95) and increases with n, with speed decreasing with n (Figure 4). And $n > 53$ is a common situation in biomedical research studies.

However, despite the high correlation value (globally achieving PCC $= 0.9535$ over all the 4,249,560 pairs of values), the cloud of the (nMCC, nDOR) points is rather spread out, indicating the occurrence of several situations where MCC and DOR behave quite differently (Figure 6 and Figure 5). Such behaviour worsens if we consider even larger confusion matrices: as an example, sample 10^6 matrices for each value of $n = 10^t$, $t = 3, 4, 5, 6, 9$ for a total of five million matrices, compute nMCC and nDOR and add the obtained values to the previous scatterplot (Figure 7). As in the previous case, the Pearson correlation coefficient between the two measure increases (slowly) with n, from PCC $= 0.9726559$ for $n = 1000$ to PCC $= 0.9727163$. Over all the nearly ten million points, the average Pearson correlation coefficient is PCC $= 0.9622007$.

Again, despite the high correlation, the nMCC and nDOR corresponding values are even farther away from lying on a straight line: there are a large number of points in two curvilinear triangles (Figure 7), one bounded by the segments $\overline{AB}$, $\overline{BC}$ and the portion of the curve $\mathbf{W}$ between the points A and C, and the second, symmetric to the first one, bounded by the segments $\overline{CD}$, $\overline{DE}$ and the curve $\mathbf{W}$ between the points E and C, where $A = (0, 0)$, $B = (\frac{1}{2}, 0)$, $C = (\frac{1}{2}, \frac{1}{2})$, $D = (\frac{1}{2}, 1)$ and $E = (1, 1)$.

We conjecture that the curve $\mathbf{W}$, the curvilinear bound of the two triangles ABC and CDE consists of the points (nMCC, nDOR) corresponding to a set of symmetric confusion matrices

$$\mathbf{W}: \left\{ (\mathrm{nMCC}(M), \mathrm{nDOR}(M)) \in [0,1]^2 \, M = \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} : \alpha \in \mathbb{N}, \beta \in \mathbb{N}_0 \right\}. \quad (6)$$

Now, Equation 6 yields that

$$\begin{aligned} \mathrm{MCC}(M) &= \frac{\alpha^2 - \beta^2}{\sqrt{(\alpha+\beta)^4}}, \quad \mathrm{nMCC}(M) = \frac{\alpha}{\alpha+\beta}, \\ &= \frac{(\alpha+\beta)(\alpha-\beta)}{(\alpha+\beta)^2}, \quad = \frac{1}{1+\frac{\beta}{\alpha}}, \\ &= \frac{\alpha-\beta}{\alpha+\beta}. \\ \mathrm{DOR}(M) &= \frac{\alpha^2}{\beta^2}. \quad \mathrm{nDOR}(M) = \frac{\alpha^2}{\alpha^2+\beta^2}. \end{aligned}$$

and thus the curve $\mathbf{W}$ is defined by the rational function

$$\mathbf{W}: \quad \mathrm{nDOR} = \frac{\mathrm{nMCC}^2}{2 \cdot \mathrm{nMCC}^2 - 2 \cdot \mathrm{nMCC} + 1}.$$

Conjecture: Given a point $P(x_p, y_p)$ inside the curvilinear triangles ABC or CDE, there is a confusion matrix M such

TABLE 1. Use cases for MCC and DOR. MCC: Matthews correlation coefficient (Equation 1). DOR: diagnostic odds ratio (Equation 3). MCC has worst value equal to −1 and best value equal to +1. DOR has worst value 0 and best value +∞, theoretically. nMCC and nDOR have both worst value 0 and best value 1. Δ(nMCC, nDOR): absolute difference between nMCC and nDOR. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. TPR: true positive rate, sensitivity. TNR: true negative rate, specificity. PPV: positive predictive value, precision. NPV: negative predictive value. TPR, TNR, PPV, NPV, accuracy, and F_1 score have worst value 0 and best value 1. We reported the formulas of TPR, TNR, PPV, NPV, accuracy, and F_1 score in the Supplementary information. Threshold cut-off for predictions: $\tau = 0.5$.

case	TP	FN	FP	TN	MCC	DOR	nMCC	nDOR	Δ(nMCC, nDOR)
UC1	1,000,000	1,000	1	1	+0.022	1,000	0.511	0.999	0.488
UC2	100,000	1,000,000	10	1	–0.009	0.010	0.496	0.001	0.495
	positives	**negatives**	**TPR**	**TNR**	**PPV**	**NPV**	**accuracy**	**F_1 score**	
UC1	1,001,000	2	0.999	0.500	1.000	0.001	0.999	0.995	
UC2	1,100,000	11	0.091	0.091	0.999	0.000	0.091	0.167	

that $\text{nDOR}(M) = x_p$ and $\text{nMCC}(M) = y_p$. In fact, consider only the triangle CDE (situation for ABC is symmetrical), and, in particular, all points $(x, y) \in ABC$ whose coordinates are integer multiples of 0.5. Then, for each point, it is possible to find a confusion matrix M such that $\text{nMCC}(M) \approx x$ and $\text{nDOR}(M) \approx y$ (Table S1).

It is worth noticing that are no confusion matrices with MCC close to −1 and with DOR > 1. In fact, the first ($0 \leq \text{nMCC} < \frac{1}{2}, \frac{1}{2} \leq \text{nDOR} \leq 1$) and the fourth quadrant ($\frac{1}{2} < \text{nMCC} \leq 1, 0 \leq \text{nDOR} < \frac{1}{2}$) of the (nMCC,nDOR) cartesian plane are empty (Figure 7).

III. RESULTS

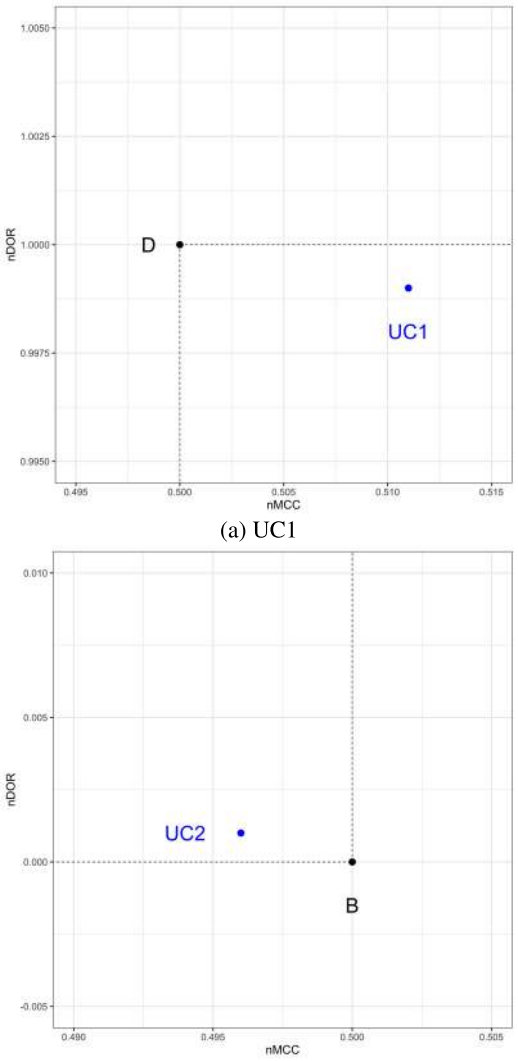
To better understand the different behavior of MCC and DOR, we decided to investigate two indicative, significant use cases where these two rates give discordant outcomes. We reported the results of the use cases UC1 and UC2 in Table 1 and in Figure 8.

A. USE CASE UC1

The UC1 confusion matrix refers to an extremely imbalanced dataset, having 1,001,000 positive data instances and only 2 negative data instances. The classifier there has correctly predicted one million of positives and half of the negatives (TN = 1). The UC1 confusion matrix can be represented as a point close to vertex D inside the triangle CDE (Figure 7).

Here, MCC has a value around 0 (and nMCC = 0.511), indicating that the prediction was similar to random guessing, while nDOR has a value of 0.999, indicating almost perfect prediction. It is clear to observe that these two outcomes are discordant: it would be difficult for a researcher to decide if the classifier performed well or not, but just looking at the values of these two rates.

If we examine the values of the four basic rates of this use case (TPR, TNR, PPV, and NPV in Table 1), we notice that the classifier performed well only on the sensitivity and precision, but just sufficiently on the specificity and badly on the negative predictive value. These four scores clearly state that the overall performance was far away from perfect, confirming the outcome of the Matthews correlation coefficient, and dismissing the outcome of the diagnostic odds ratio.



(a) UC1

(b) UC2

FIGURE 8. Zooming of the relative positions of the use cases UC1 and UC2 in the (nMCC,nDOR) cartesian plane of Figure 7.

B. USE CASE UC2

The dataset of the use case UC2 is extremely imbalanced, too, with 1,100,000 positive elements and 11 negative elements.

The UC2 confusion matrix was able to correctly classify 100,000 positives and only 1 negative. The UC1 confusion matrix can be represented as a point close to vertex B inside the triangle ABC in Figure 7, symmetric to the UC1 point.

Here, MCC has a value around 0 (and nMCC = 0.596) again, indicating that the prediction was similar to random guessing, while nDOR has a value of 0.001 (DOR = 0.010), indicating an extremely bad prediction. Again, here a practitioner would have difficulties in understanding if the classifier performed well or poorly, by just looking at the values of nMCC and nDOR. By considering the values of the four basic rates of this use case (TPR, TNR, PPV, and NPV in Table 1), we can notice that the classifier obtained extremely low sensitivity, specificity, and negative predictive value, but a very high precision.

Differently from the MCC, the value of the diagnostic odds ratio fails to communicate the high value of precision in UC2.

IV. DISCUSSION AND CONCLUSION

Even if four-category confusion matrices are a common tool to evaluate binary classifications in supervised machine learning and statistics, no general consensus has been reached on a unique statistical score able to informatively recap its key-message. In the past, several studies proposed the Matthews correlation coefficient as a statistical rate more informative and truthful than accuracy, F_1 score [5], cross-entropy error [3], bookmaker informedness, markedness, and balanced accuracy [8].

A recent article by Rácz *et al.* [10] presented the comparison of tens of statistical rates employed to represent confusion matrices, and indicated the diagnostic odds ratio and markedness as the two most informative scores. In the past, we already showed the advantages of the MCC over markedness [8].

Therefore, in the present study, we decided to compare the MCC with the DOR, by exploring their mathematical properties and relationships, and by examining some indicative use cases where these two rates generate discordant outcomes.

From our analyses, we deduced that the MCC is more informative and truthful than the DOR, because it produces a high score only if the values of all the four basic rates of a confusion matrix (sensitivity, specificity, precision, and negative predictive value) are high. In some cases, the diagnostic odds ratio, instead, can produce an inflated overoptimistic high value even if one of the basic rates has a low score, misleading the researcher.

Our discoveries about the Matthews correlation coefficient and the diagnostic odds ratio can also be interpreted with respect to previously published studies on this topic. In their highly-cited article, Glas *et al.* [9] stated that: ''The diagnostic odds ratio as a measure of test performance combines the strengths of sensitivity and specificity [. . .] with the advantage of accuracy as a single indicator''. We agree with this sentence, but we also reaffirm that DOR fails to take into account two other fundamental rates of confusion matrices: precision and negative predictive value. On the contrary, as already mentioned, the MCC combines all the four rates into a singular score; and when the MCC has a high value, it means that all the four basic rates have a high value, too.

The same article by Glas *et al.* [9] additionally states: ''These characteristics lend the DOR particularly useful for comparing tests whenever the balance between false negative and false positive rates is not of immediate importance''. We agree with this statement, that actually highlights an important limitation of this ratio: the diagnostic odds ratio fails to work effectively on imbalanced datasets. The Matthews correlation coefficient, instead, is particularly useful right on the imbalanced datasets, because it generates a high score only if the classifier was able to correctly recognize most of the positive elements and most of the negative elements, proportionally to their class size.

An article by Böhning and colleagues [30] advises against the use DOR to determine the optimal cut-off for diagnostic tests, and we broadly agree with this key-message. The same study also suggests to use the Youden index (also known as *bookmaker informedness*, and defined as MK = sensitivity + specificity − 1) for that goal, but we disagree with this indication: the Youden index, in fact, considers only two of the four basic rates of a confusion matrix (sensitivity and specificity), and does not consider the other two (precision and negative predictive value). We already proved that the MCC is more informative than bookmaker informedness, too, in another study [8].

In conclusion, we recommend the scientific community to use the Matthews correlation coefficient instead of the diagnostic odds ratio to assess binary classifications. In the future, we plan to investigate the relationships between the MCC and other rates, such as the Fowlkes-Mallows index [46] and the prevalence threshold [47].

LIST OF ABBREVIATIONS

AUC: area under the curve. CT: confusion tetrahedron. DOR: diagnostic odds ratio. MCC: Matthews correlation coefficient. NPV: negative predictive value. PPV: positive predictive value (precision). PR: precision-recall. ROC: receiver operating characteristic. SISCOM: subtraction ictal SPECT co-registered to MRI. TNR: true negative rate (specificity). TPR: true positive rate (sensitivity, recall).

ACKNOWLEDGMENT

The authors thank Dankmar Böhning (University of Southampton) for his suggestions, and Károly Héberger (Magyar Tudományos Akadémia) for his prompts.

REFERENCES

[1] B. W. Matthews, ''Comparison of the predicted and observed secondary structure of t4 phage lysozyme,'' *Biochimica et Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[2] R. Delgado and J. D. Nuñez-González, ''Enhancing confusion entropy (CEN) for binary and multiclass classification,'' *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0210264.
[3] G. Jurman, S. Riccadonna, and C. Furlanello, ''A comparison of MCC and CEN error measures in multi-class prediction,'' *PLoS ONE*, vol. 7, no. 8, Aug. 2012, Art. no. e41882.

[4] C. J. V. Rijsbergen, *Information Retrieval*. Oxford, U.K.: Butterworths, 1979.

[5] 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, Dec. 2020.

[6] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2017.

[7] D. M. Powers, "Evaluation evaluation," in *Proc. ECAI 18th Eur. Conf. Artif. Intell.*, 2008, pp. 843–844.

[8] 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, Dec. 2021.

[9] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: A single indicator of test performance," *J. Clin. Epidemiol.*, vol. 56, no. 11, pp. 1129–1135, Nov. 2003.

[10] A. Rácz, D. Bajusz, and K. Héberger, "Multi-level comparison of machine learning classifiers and their performance metrics," *Molecules*, vol. 24, no. 15, p. 2811, Aug. 2019.

[11] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, May 2000.

[12] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177678.

[13] K. Abhishek and G. Hamarneh, "Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation," 2020, *arXiv:2010.13454*. [Online]. Available: http://arxiv.org/abs/2010.13454

[14] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019.

[15] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0208737.

[16] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.

[17] J. Yao and M. Shepperd, "Assessing software defection prediction performance: Why using the Matthews correlation coefficient matters," in *Proc. Eval. Assessment Softw. Eng.*, Apr. 2020, pp. 120–129.

[18] J. A. Doust, P. P. Glasziou, E. Pietrzak, and A. J. Dobson, "A systematic review of the diagnostic accuracy of natriuretic peptides for heart failure," *Arch. Internal Med.*, vol. 164, no. 18, pp. 1978–1984, 2004.

[19] B. M. Tang, G. D. Eslick, J. C. Craig, and A. S. McLean, "Accuracy of procalcitonin for sepsis diagnosis in critically ill patients: Systematic review and meta-analysis," *Lancet Infectious Diseases*, vol. 7, no. 3, pp. 210–217, Mar. 2007.

[20] E. A. Tsochatzis, K. S. Gurusamy, S. Ntaoula, E. Cholongitas, B. R. Davidson, and A. K. Burroughs, "Elastography for the diagnosis of severity of fibrosis in chronic liver disease: A meta-analysis of diagnostic accuracy," *J. Hepatol.*, vol. 54, no. 4, pp. 650–659, Apr. 2011.

[21] S. B. Yoon, S.-H. Moon, J. H. Kim, T. J. Song, and M.-H. Kim, "The use of immunohistochemistry for IgG4 in the diagnosis of autoimmune pancreatitis: A systematic review and meta-analysis," *Pancreatology*, vol. 20, no. 8, pp. 1611–1619, Dec. 2020.

[22] S. Picot, M. Cucherat, and A.-L. Bienvenu, "Systematic review and meta-analysis of diagnostic accuracy of loop-mediated isothermal amplification (LAMP) methods compared with microscopy, polymerase chain reaction and rapid diagnostic tests for malaria diagnosis," *Int. J. Infectious Diseases*, vol. 98, pp. 408–419, Sep. 2020.

[23] E. Tedeschini, M. Fava, and G. I. Papakostas, "Placebo-controlled, antidepressant clinical trials cannot be shortened to less than 4 weeks' duration: A pooled analysis of randomized clinical trials employing a diagnostic odds ratio–based approach," *J. Clin. Psychiatry*, vol. 72, no. 1, pp. 98–103, 2010.

[24] M. Kokot, F. Petric, M. Cepanec, D. Miklic, I. Bejic, and Z. Kovacic, "Classification of child vocal behavior for a robot-assisted autism diagnostic protocol," in *Proc. 26th Medit. Conf. Control Autom. (MED)*, Jun. 2018, pp. 1–6.

[25] G. Aungaroon, A. Trout, R. Radhakrishnan, P. S. Horn, R. Arya, J. R. Tenney, T. M. Arthur, K. D. Holland, F. T. Mangano, J. L. Leach, L. Rozhkov, and H. M. Greiner, "Impact of radiotracer injection latency and seizure duration on subtraction ictal SPECT co-registered to MRI (SISCOM) performance in children," *Clin. Neurophysiol.*, vol. 129, no. 9, pp. 1842–1848, Sep. 2018.

[26] K. Kamyab-Hesari, N. Aghazadeh, P. Nourmohammad-pour, A. Ghanadan, A. Nikoo, F. Gholamali, M.-J. Nazemi, and Z. Rahbar, "Diagnostic accuracy measures for vertical and transverse scalp biopsies in cicatricial and non-cicatricial alopecias," *Dermatologica Sinica*, vol. 36, no. 1, pp. 30–35, Mar. 2018.

[27] E. Nwoye, W. L. Woo, O. Fidelis, C. Umeh, and B. Gao, "Development and investigation of cost-sensitive pruned decision tree model for improved schizophrenia diagnosis," *Int. J. Automat., Artif. Intell. Mach. Learn.*, vol. 1, no. 1, pp. 17–41, 2020.

[28] O. Rahmati, A. Kornejady, M. Samadi, R. C. Deo, C. Conoscenti, L. Lombardo, K. Dayal, R. Taghizadeh-Mehrjardi, H. R. Pourghasemi, S. Kumar, and D. T. Bui, "PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches," *Sci. Total Environ.*, vol. 664, pp. 296–311, May 2019.

[29] A.-M. Šimundić, "Measures of diagnostic accuracy: Basic definitions," *Electron. J. Int. Fed. Clin. Chem. Lab. Med.*, vol. 19, no. 4, p. 203, 2009.

[30] D. Böhning, H. Holling, and V. Patilea, "A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test," *Stat. Methods Med. Res.*, vol. 20, no. 5, pp. 541–550, Oct. 2011.

[31] J. Ekström, "The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate," Dept. Statistics, UCLA, Los Angeles, CA, USA, Tech. Rep. 7qp4604r, 2011.

[32] F. Boas, "Determination of the coefficient of correlation," *Science*, vol. 29, no. 751, pp. 823–824, May 1909.

[33] K. Pearson, "Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable," *Phil. Trans. Roy. Soc. A Math. Phys. Eng. Sci.*, vol. 195, nos. 262–273, pp. 1–47, 1900.

[34] G. U. Yule, "On the methods of measuring association between two attributes," *J. Roy. Stat. Soc.*, vol. 75, no. 6, pp. 579–652, 1912.

[35] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1946.

[36] T. R. Ostrowski and T. Ostrowski, "The basic four measures and their derivates in dichotomous diagnostic tests," *Int. J. Clin. Biostatistics Biometrics*, vol. 6, no. 1, p. 26, Jun. 2020.

[37] V. V. Starovoitov and Y. I. Golub, "Comparative study of quality estimation of binary classification," *Informatics*, vol. 17, no. 1, pp. 87–101, Mar. 2020.

[38] B. J. B. S. Haldane, "The estimation and significance of the logarithm of a ratio of frequencies," *Ann. Human Genet.*, vol. 20, no. 4, pp. 309–311, May 1956.

[39] B. Littenberg and L. E. Moses, "Estimating diagnostic accuracy from multiple conflicting reports," *Med. Decis. Making*, vol. 13, no. 4, pp. 313–321, Dec. 1993.

[40] L. E. Moses, D. Shapiro, and B. Littenberg, "Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations," *Statist. Med.*, vol. 12, no. 14, pp. 1293–1316, Jul. 1993.

[41] J. M. Bland and D. G. Altman, "The odds ratio," *BMJ*, vol. 320, no. 7247, p. 1468, 2000.

[42] J. Cornfield, "A method for estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix," *J. Nat. Cancer Inst.*, vol. 11, no. 6, pp. 1269–1275, 1951.

[43] M. Nurminen, "To use or not to use the odds ratio in epidemiologic analyses?" *Eur. J. Epidemiol.*, vol. 11, no. 4, pp. 365–371, Aug. 1995.

[44] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Comput.*, vol. 13, no. 3, pp. 213–225, Feb. 2009.

[45] A. Luque, A. Carrasco, A. Martín, and A. D. L. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019.

[46] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, Sep. 1983.

[47] J. Balayla, "Prevalence threshold (ϕe) and the geometry of screening curves," *PLoS ONE*, vol. 15, no. 10, Oct. 2020, Art. no. e0240215.

DAVIDE CHICCO received the Bachelor of Science and Master of Science degrees in computer science from the Università di Genova, Genoa, Italy, in 2007 and 2010, respectively, and the Ph.D. degree in computer engineering from the Politecnico di Milano University, Milan, Italy, in 2014. He also spent a semester as a Visiting Ph.D. Scholar with the University of California Irvine, USA. From September 2014 to September 2018, he has been a Postdoctoral Researcher with the Princess Margaret Cancer Centre and a Guest with the University of Toronto. From September 2018 to December 2019, he was a Scientific Associate Researcher with the Peter Munk Cardiac Centre, Toronto, ON, Canada. From January 2020 to January 2021, he has been a Scientific Associate Researcher with the Krembil Research Institute, Toronto. Since January 2021, he has been working as a Scientific Researcher with the Institute of Health Policy Management and Evaluation, University of Toronto.

VALERY STAROVOITOV received the Diploma degree in mathematics from Belarusian State University, in 1977, and the Ph.D. degree in computer science from the V. M. Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, in 1990. This Institute was the leading scientific centers of the USSR. He defended the Doctor of Sciences dissertation on the Application of Computing Technology and the Theoretical Foundations of Informatics from the Institute of Technical Cybernetics, National Academy of Sciences of Belarus, in 1999. He received the title of Full Professor, in 2003. He has been working all his life in one research organization - the United Institute of Informatics Problems (until 2002 the Institute of Technical Cybernetics), National Academy of Sciences of Belarus. Since 2000, he has taught as a Full Professor of computer graphics and image processing at a number of universities in Belarus, Poland, and Kazakhstan. He is a Laureate of the State Prize of the Republic of Belarus in the field of science and technology, in 2002. In the early 1990s, he was one of the organizers of the Belarusian branch of the International Association for Pattern Recognition (IAPR). His current research interests include image processing and analysis, quality assessment of digital images, and results of data classification. He was a member of the IEEE.

GIUSEPPE JURMAN received the Ph.D. degree in algebra from the Università di Trento, Italy, in 1998. After two years as a Postdoctoral Fellow with the Australian National University (ANU), Canberra, in 2002, he moved to the Fondazione Bruno Kessler (FBK), Trento, where he is currently the Head of the Data Science for Health Unit, coordinating a team of researchers working in computational biology. His main research interests include machine learning, mathematical modeling, and network analysis. He is also an expert in scientific programming with R/Python and other computing languages. He teaches data visualization in the Master of Science course in data science with the University of Trento. Since 2008, he has been the Co-director of WebValley, the FBK summer school for dissemination of interdisciplinary research for high school students.