A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes-Mallows index

Davide Chicco[*]
University of Toronto
ORCID: 0000-0001-9655-7142

Giuseppe Jurman
Fondazione Bruno Kessler
ORCID: 0000-0002-2705-5728

1st May, 2023

Abstract

Even if assessing binary classifications is a common task in scientific research, no consensus on a single statistic summarizing the confusion matrix has been reached so far. In recent studies, we demonstrated the advantages of the Matthews correlation coefficient (MCC) over other popular rates such as cross-entropy error, F_1 score, accuracy, balanced accuracy, bookmaker informedness, diagnostic odds ratio, Brier score, and Cohen's kappa. In this study, we compared the MCC to other two statistics: prevalence threshold (PT), frequently used in obstetrics and gynecology, and Fowlkes-Mallows index, a metric employed in fuzzy logic and drug discovery, Through the investigation of the mutual relations among three metrics and the study of some relevant use cases, we show that, when positive data elements and negative data elements have the same importance, the Matthews correlation coefficient can be more informative than its two competitors, even this time.

Keywords: Matthews correlation coefficient; prevalence threshold; Fowlkes-Mallows index; binary classification; confusion matrix; supervised machine learning.

[*]corresponding author: davidechicco@davidechicco.it

1 Introduction

Binary classification is a common scientific task where elements belonging to two categories are predicted by a classifier to be part of one of those two categories. Data elements are usually called *positives* and *negatives*, labelled with 1s and 0s, respectively, and can represent any dual entity such as deceased patients or survived patients, sick patients or healthy individuals, for example.

In supervised machine learning and computational statistics, if the predictive algorithm correctly predicts a positive element as positive, the element is added to the true positives (TP) tally. On the contrary, f the predictive algorithm wrongly predicts a positive element as negative, the element considered a false negative (FN). Similarly, a negative element correctly classified as negative is named true negative (TN), and a negative element mistakenly classified as positive is named false positive (FP).

All these four tallies (TP, FN, TN, and FP) are usually included in a 2 × 2 contingency table called *confusion matrix*. Since the understanding of the outcome of a binary classification test can difficultly be done by considering all these four values together. scientific researchers invented several rates to summarize the values of the confusion matrix, such as Matthews correlation coefficient, F_1 score, accuracy, markedness, informedness, and others. The goal of these statistics is to provide a recap of the confusion matrix in a single score at first glance: one real value that can inform analysts about the outcome of the binary classification. Even if several statistics for this goal have been designed, no common consensus for a standard metric has been found yet.

Among these rates, we believe the Matthews correlation coefficient (MCC) can be considered a good candidate for a standard metric role for assessing binary classifications, since it generates a high score only if most of the data elements where predicted correctly in both the data classes [1]. In the past few years, we explained the advantages of the MCC over other cross-entropy error [2], F_1 score and accuracy [3, 4], balanced accuracy, bookmaker informedness, and markedness [5], diagnostic odds ratio [6], Brier score and Cohen's kappa [7], and area under the receiver operating characteristic curve (ROC AUC) [1].

In this study, we focus on the comparison between the MCC and prevalence threshold (PT) [8, 9] and Fowlkes-Mallows index (FM) [10]. Prevalence threshold is a confusion matrix statistic commonly used in obstetrics and gynecology [11] but also in other biologically related fields [12, 13, 14] , while the Fowlkes-Mallows index had applications in fuzzy logic [15] and drug discovery [16]. The FM was generalized for topic model validation [17, 18], and also employed for clustering comparison [19, 20, 21, 22] and optimal decision processes [23], and recently compared to MCC and Youden index [24].

We included a thorough scientific literature review on the studies employing the MCC in our previous articles [2, 3, 4, 5, 6, 7, 1]. Regarding prevalence threshold and Fowlkes-Mallows index, instead, we noticed a small number of studies involving these rates in the scientific literature. A study by Jacques Balayla [8] explained the geometric aspects of prevalence threshold in the context of medical screening tests, while another preprint by the same author [9] investigated this rate among the conditions of maximum accuracy. An article of the same author Jacques Balayla, written with Liora Elfassy et al. [11], described the usage of prevalence thresholds of various screening tests in obstetrics and gynecology, concluding that many positive screening tests likely represent false positives, among the data considered. A similar study by Jacques Balayla and his team [25] was focused on prevalence threshold utilized for the interpretation of COVID-19 screening tests.

Regarding the studies involving the usage of the Fowlkes-Mallows index, we can report an article Silke Wagner and Dorothea Wagner [19] and a work by Marina Meila [20] that employ this metric for comparing clustering results. Being clustering an unsupervised machine learning family of methods, the authors do not refer to the traditional confusion matrix of supervised binary classification, of course, but rather to the relationships between clusters. The Fowlkes-Mallows index was employed to compare clusterings also by Pinar Yildirim and colleagues [16], which leverage it in a study analyzing drugs data to cluster antibiotics. Eduardo Ramirez et al. [17], in their study, proposed a generalized version of the Fowlkes-Mallows index for topic model validation.

Even if prevalence threshold and Fowlkes-Mallows index seem less employed than the Matthews correlation coefficient in the scientific community, we decided to analyze these metrics in detail, to reveal the differences in their outcomes and insights.

We organized the rest of this article as follows. After this Introduction, we describe the mathematical background of the three metrics considered in section 2 and the relationships between the rates in section 3. We then report some use cases and their results in section 4, and discuss some final remarks in section 5.

2 Mathematical background

Although quite different in their purpose, the three performance measures can being effectively compared on a common ground, by explicitly reading their original definition in terms of the four entries of the confusion matrix $\left(\begin{smallmatrix}\mathrm{TP} & \mathrm{FN}\\ \mathrm{FP} & \mathrm{TN}\end{smallmatrix}\right)$: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). We provide the definition of the true positive rate (TPR), true negative rate (TNR) and positive predictive value (PPV) in the Supplementary Information.

2.1 Matthews correlation coefficient (MCC)

Originally born in the context of molecular chemistry [26] and rediscovered only recently by the machine learning community [27], MCC has gained more and more interest due to its robustness and reliability as a classifier performance measure, especially in the binary setup, although its definition can be naturally extended to the multiclass case [28]. In particular, the invariance to the class imbalance and the property of being high if all the four basic rates of the confusion matrix are high characterize MCC as a good metric [5].

As a function of the confusion matrix entries, MCC reads as follows:

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

(minimum and worst value $= -1$; maximum and best value $= +1$)

naturally extended to the cases where the denominator is zero [4], so that MCC is defined for every confusion matrix. By definition, MCC uses all four entries of the confusion matrix, and it is invariant for class swapping (Positive versus Negative) and True versus False swapping. Finally, MCC ranges between its worst value -1 (full misclassification) and its best value 1 (perfect classification), with 0 corresponding to random classification (coin tossing). For comparison's purposes, MCC is usually linearly projected into the range $[0, 1]$ by the function

$$\mathrm{normMCC} = \frac{\mathrm{MCC}+1}{2} \tag{2}$$

(minimum and worst value $= 0$; maximum best value $= 1$)

2.2 Prevalence threshold (PT)

Prevalence threshold can be expressed as a function of true positive rate and true negative rate:

$$\begin{aligned} \text{PT} &= \frac{\sqrt{\text{TPR}\cdot(1-\text{TNR})}-(1-\text{TNR})}{\text{TPR}-(1-\text{TNR})} = \\ &= \frac{\sqrt{\text{FPR}}}{\sqrt{\text{FPR}}+\sqrt{\text{TPR}}} = \\ &= \left(1+\sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}}\cdot\frac{\text{TN}+\text{FP}}{\text{FP}}}\right)^{-1} \end{aligned} \tag{3}$$

(minimum and best value $= -1$; maximum and worst value $= 0$)

By definition, PT ranges between the worst value 1 (full misclassification) and 0 (perfect classification): to ease comparison with the other metrics, it is natural to consider its complement to one

$$\text{complPT} = 1 - \text{PT} \tag{4}$$

(minimum and worst value $= 0$; maximum and best value $= 1$)

Since FPR is defined when there is at least 1 negative sample and TPR is defined whenever at least 1 positive sample occurs, PT is defined for each binary classification task where the two classes have both at least 1 sample. Further, note that PT relies on all four entries of the confusion matrix $\left(\begin{smallmatrix}\text{TP} & \text{FN}\\ \text{FP} & \text{TN}\end{smallmatrix}\right)$, but it is not invariant for both swaps true versus false and positive versus negative.

2.3 Fowlkes–Mallows index (FM)

When translating the definition as a function of the confusion matrix entries, we obtain

$$\begin{aligned} \text{FM} &= \sqrt{\text{PPV}\cdot\text{TPR}} = \\ &= \frac{\text{TP}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})}} \end{aligned} \tag{5}$$

(minimum and worst value $= 0$; maximum and best value $= 1$)

ranging between the full misclassification value 0 to the perfect classification 1. Similarly to the definition of the F_1 score, the value for TN is not an input, so FM is not involving all the entries of the confusion matrix. As a direct consequence, also FM is not invariant for true versus false and positive versus negative swaps.

3 Relationships between rates

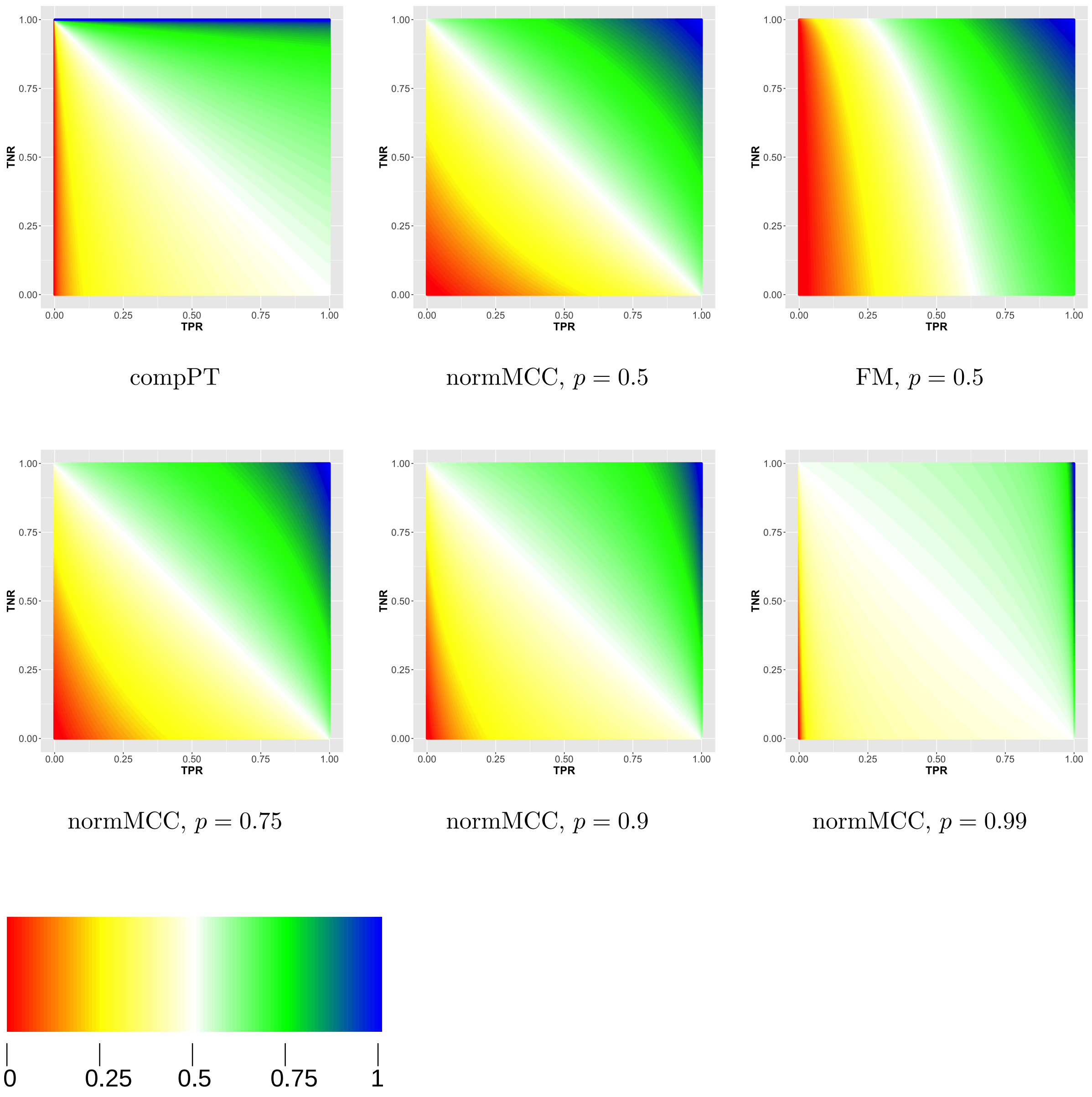
The different nature of the three metrics prevent to establish detailed and meaningful mathematical relationships between MCC, FM and PT. In fact, no straightforward connection emerges among the measures, regardless of the employed input rates, that are the entries TP, TN, FP, FN of the confusion matrix as in Equation 1, Equation 3, Equation 5, as well as using the basic rates true negative rate (TNR) and true positive rate (TPR), together with the prevalence $p = \frac{\text{TP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$:

$$\text{MCC} = \frac{\text{TNR}+\text{TPR}-1}{\sqrt{\left(1-\text{TNR}+\frac{p}{1-p}\ \text{TPR}\right)\left(1-\text{TPR}+\frac{1-p}{p}\ \text{TNR}\right)}}$$

$$\text{PT} = \frac{\sqrt{\text{TPR}\cdot(\text{1-TNR})}+\text{TNR}-1}{\text{TPR}+\text{TNR}-1}$$

$$\text{FM} = \text{TPR}\cdot\sqrt{\frac{p}{p\cdot\text{TPR}+(1-\text{TNR})(1-p)}}$$

Nonetheless, the expression of the three metrics in terms of TPR, TNR and p allows displaying the behaviour of the measures themselves on the Cartesian (TPR,TNR) plane, for different values of the prevalence, as shown in Figure 1 and Figure 2. The twelve heatmaps report in detail the overall behaviour of the metrics on the whole of the TPR, TNR plane, showing their very diverse structure. In particular, we show the comparison of the three metrics for the perfectly balanced dataset case $p = 0.5$, and in some specific unbalanced cases such as the clinically relevant values $p = 0.75, 0.9$ and the extreme $p = 0.99$, and the symmetric cases $p = 0.25, p = 0.1, p = 0.01$. These last three cases have not been considered for normMCC due to its invariance to class swapping. Finally, we recall that PT does not depend on the prevalence, so the corresponding plot is the same regardless of the value of p. As an immediate consideration that can be drawn from the plots, FM tends to move faster towards extreme values 0 and 1 in the highly unbalanced cases, while normMCC tends to keep a smoother trend.



compPT

normMCC, $p = 0.5$

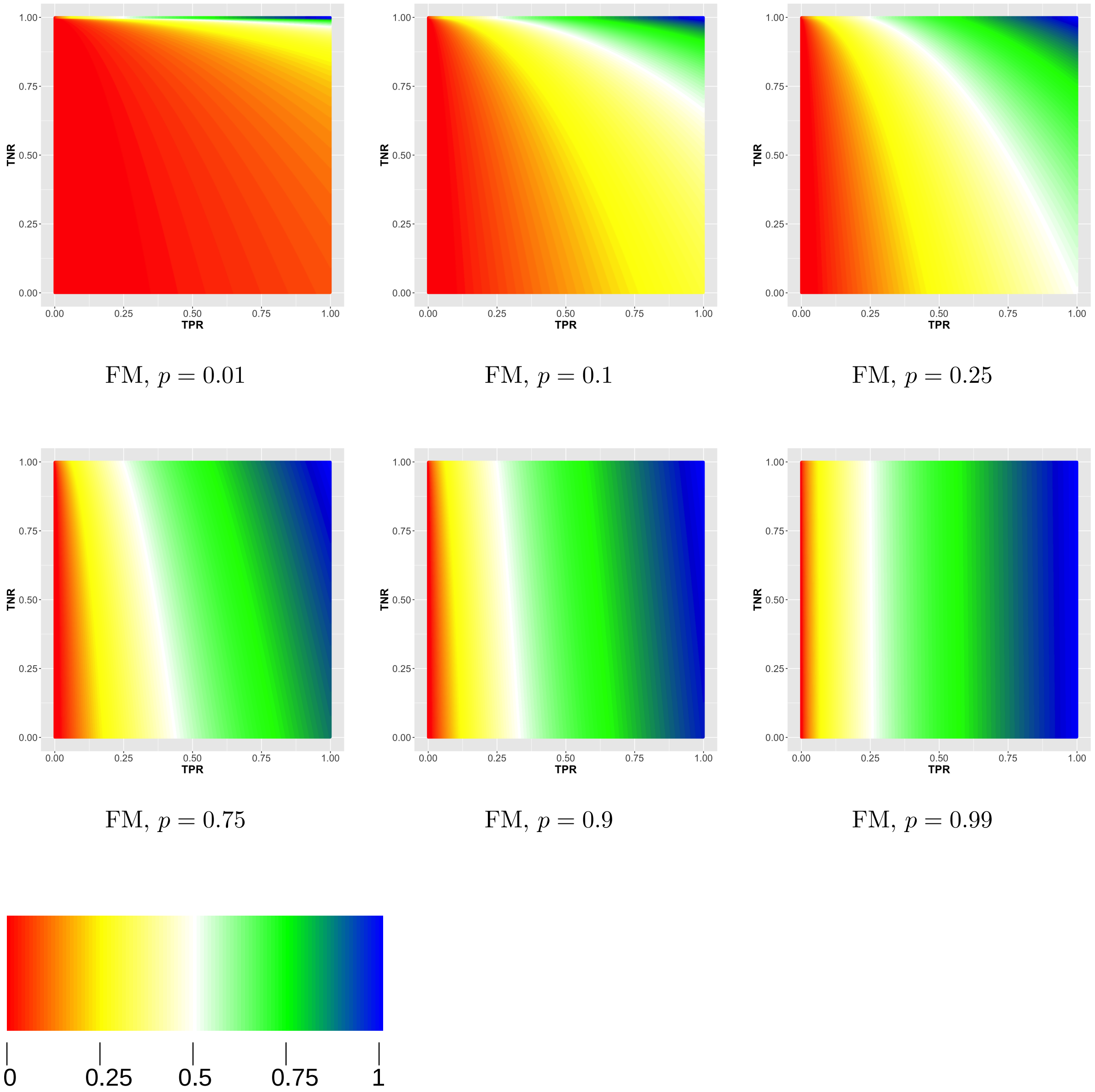
FM, $p = 0.5$

normMCC, $p = 0.75$

normMCC, $p = 0.9$

normMCC, $p = 0.99$

Figure 1: ***normMCC, compPT and FM values on the TPR, TNR Cartesian plane.*** *Comparison of the three metrics for the balanced prevalence $p = 0.5$ case (top row) and comparison of normMCC values for growing prevalence $p = 0.75, 0.9, 0.99$. Due to the invariance to class swapping for normMCC, only prevalence values larger than 0.5 are considered. No prevalence value is set for compPT since it is independent of such parameter.*



FM, $p = 0.01$

FM, $p = 0.1$

FM, $p = 0.25$

FM, $p = 0.75$

FM, $p = 0.9$

FM, $p = 0.99$

Figure 2: ***FM values on the TPR, TNR Cartesian plane in unbalanced prevalence cases.*** *Due the non-invariance of FM for class swapping, the set of values* $p = 0.01, 0.1, 0.25, 0.75, 0.9, 0.99$ *are considered.*

The three metrics are quite different in nature, both by definition and by the diverse tasks they were conceived for. However, some interesting considerations can be drawn through an extensive experimental mutual analysis, as shown in what follows. Note that some peculiar relations have been discussed in the paper [9]. In particular, first the author introduced the concept of positive prevalence threshold as the level in the precision-prevalence curve below which binary classification performances start to fail and show that, for the perfect accuracy cases, FM ranges in $[1, \sqrt{2}]$. Similarly, he also introduced the negative prevalence threshold as the level beyond which the NPV curve drops most significantly, and show that the area between both these two thresholds bounds MCC in the $[\frac{\sqrt{2}}{2}, \sqrt{2}]$ range.

Thus, to better sense the relationship between the Matthews correlation coefficient and the prevalence threshold, and between the Matthews correlation coefficient and the Fowlkes-Mallows index, we depicted two scatterplots having the Matthews correlation coefficient on the x axis and each of the the two rates on the y axis. As reasonably predictable, there is a basic coherence between MCC and the other two measures, but at different levels and with interesting twists due to the very diverse nature of the measures.

To properly compare the MCC and the PT, we employed the normalized MCC (Equation 2) and the complementary PT (Equation 4), which both range in the $[0, 1]$ interval and have 0 as worst possible score and 1 as best possible value. As shown by the scatterplot in Figure 3, the correlation between the two measures is confirmed by the point clouds of the plot lying in the first and third quadrant of the $[0, 1] \times [0, 1]$ Cartesian plane. Nonetheless, the width of the clouds stemming from the parting line $y = x$ indicates that a wide landscape of situations can occur. In particular, while such variability is quite limited around the coin tossing case, it tends to rapidly grow while moving towards better (or worse, since the plot is symmetric in the two quadrants) predictions. For instance, for a given value of complementary PT between 0.6 and 1, the corresponding normMCC can even reach values as low as 0.6 and as high as 0.9, thus covering situation ranging from just slightly better than coin tossing to almost perfect classification. The dual case, *i.e.*, the variability of PT for a given value of MCC, is not very different, indicating that the two measures, although roughly agreeing at a high level, they are indeed quantifying different aspects of the structure of a confusion matrix, namely the classification performance versus the decay of the positive predictive value. Quite interestingly and extreme, an optimal prevalence threshold can be matched by any (positive) value of MCC, as shown by the segment $\text{complPT} = 1,\ 0.5 < \text{normMCC} \leq 1$. Such behaviour is known and has been pinpointed in [8], where the author proposes the introduction of a novel measure, the negative prevalence threshold, to deal with this odd issue.

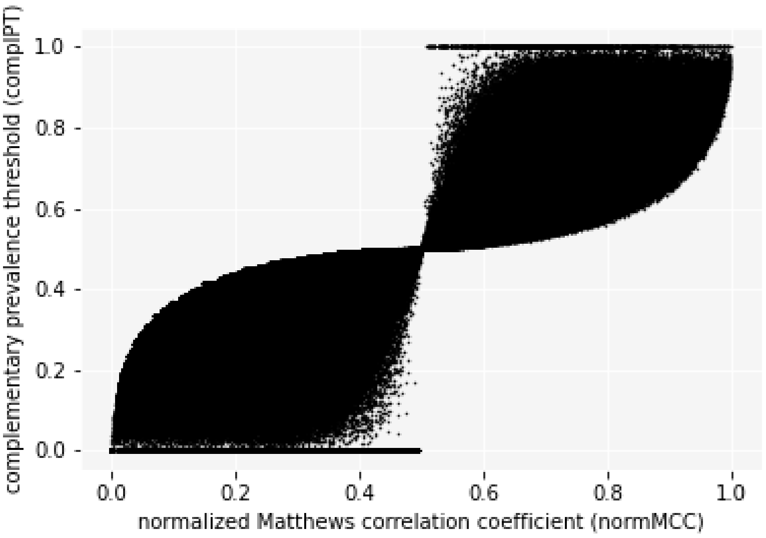


Figure 3: ***Relationship between MCC and prevalence threshold***. *We computed the normalized MCC and the prevalence threshold for* 10^7 *confusion matrices.*

Although with different mathematical insights, the relation between MCC and FM is quite similar. Again, for fair comparison purposes, we employed the normalized MCC (Equation 2) and the original FM (Equation 5), which both range in the $[0, 1]$ interval and have 0 as worst possible score and 1 as best possible value. The corresponding scatterplot is reported in Figure 4. Rough correlation between the two metrics is warranted here, too, but the shape of the point cloud suggests a number of significant differences with respect to the previous case involving PT. First, the variability of MCC for a given value of FM is approximately constant and thus independent of the value of FM: in particular, the point cloud is roughly a parallelogram, bounded by the sharp segments FM $= 2 \cdot$ normMCC and FM $= 0$ for $0 \leq$ normMCC ≤ 0.5 and the two coarser lines FM $= 2 \cdot$ (normMCC $- 0.5$) and FM $= 1$ for $0.5 \leq$ normMCC ≤ 1. This yields that, for a given value of FM, normMCC can vary across a range as wide as 0.5, thus covering a very large set of cases. On the other hand, variability of FM as a function of normMCC changes linearly, first widening from the minimum of no variation for normMCC $= 0$ to the maximum of 1 for normMCC $= 0.5$, and then symmetrically narrowing down again for larger values of normMCC. Thus FM variability is quite limited for the cases of very low or very high normMCC, but can be extremely large when normMCC is closer to the coin tossing value. Again, the two metrics are measuring different properties of a confusion matrix: this is particularly evident in this case, where FM has been tried as a classification performance evaluation measure mainly because of its mathematical definition of being the geometric mean between precision and true positive rate, warranting a kind of compromise between the two averaged rates.

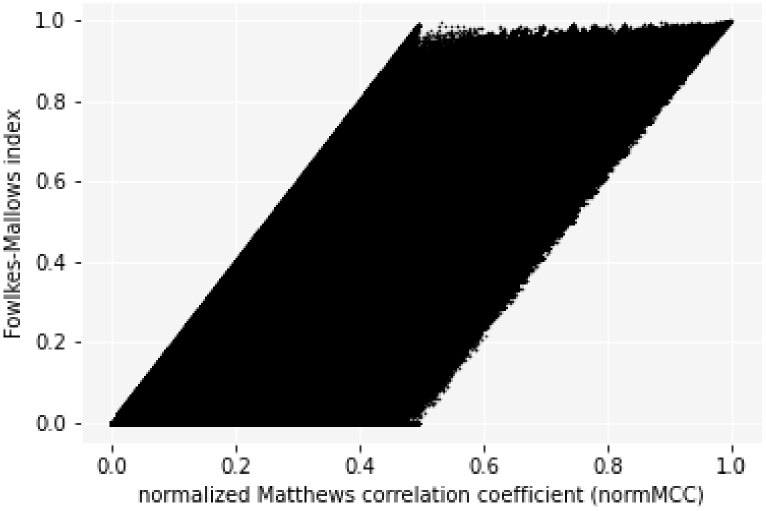


Figure 4: ***Relationship between MCC and Fowlkes-Mallows index.*** *We computed the normalized MCC and the Fowlkes-Mallows index for* 10^7 *confusion matrices.*

4 Use cases

To better understand the behavior of Matthews correlation coefficient, prevalence threshold, and Fowlkes-Mallows index, we designed several indicative use cases where each pair of rates have different outcomes. Moreover, for each use case, we report the numbers of its confusion matrix and the real values of its four *basic rates* (true positive rate, true negative rate, positive predictive value, and negative predictive value) [5]. All these four basic rates have minimum value equal to 0, meaning worst possible result, and maximum value equal to 1, meaning best possible result (Supplementary information).

It is important to mention that we interpreted the results of these use cases under the condition where positive elements and negative elements have the same importance, and therefore predicting correctly a positive data instance has the same relevance of predicting a negative data instance. We are aware that there are several scientific scenarios, especially in the biomedical sciences, where positives are more important than negatives (or vice versa). That happens, for example, when positive data instances represent patients diagnosed with a specific disease [29]. In those cases, all the considerations about MCC, PT, and FMI presented in this study no longer stand, and the involvement of other statistical rates, giving more importance to correctly predicted positive elements (or to the negatively ones). The discussion of the rates employed in those cases falls beyond the scope of this study; here, we consider the cases where positives and negatives are equally important.

4.1 MCC and PT

We reported six indicative use cases where the Matthews correlation coefficient and the prevalence threshold produce different key-messages in Table 1.

case	TP	TN	FP	FN	TPR	TNR	PPV	NPV	PT	MCC	complPT	normMCC	Δ(c,n)
UC01	1	43,001	99,001	1,001	0.001	0.303	0	0.977	0.964	−0.126	0.036	0.437	0.401
UC02	1	99,001	1,001	99,001	0	0.99	0.001	0.5	0.969	−0.071	0.031	0.465	0.434
UC03	1,001	97,001	1	99,001	0.01	1	0.999	0.495	0.031	+0.07	0.969	0.535	0.434
UC04	1,001	98,001	1	99,001	0.01	1	0.999	0.497	0.031	+0.07	0.969	0.535	0.434
UC05	99,001	1,001	1	99,001	0.5	0.999	1	0.01	0.043	+0.071	0.957	0.535	0.422
UC06	97,001	1,001	1	43,001	0.693	0.999	1	0.023	0.037	+0.125	0.963	0.563	0.4

Table 1: ***Use cases for comparisons between Matthews correlation coefficient and prevalence threshold***. *MCC: Matthews correlation coefficient (Equation 1). normMCC: normalized Matthews correlation coefficient (Equation 2). PT: prevalence threshold (Equation 3). complPT: complementary prevalence threshold (Equation 4). TPR, TNR, PPV, NPV, normMCC, and complPT have worst value equal to 0 and best value equal to 1. MCC has worst value equal to –1 and best value equal to +1. PT has worst value equal to 1 and best value equal to 0. Δ(c, n): absolute difference between normMCC and complPT. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. TPR: true positive rate, sensitivity, recall. TNR: true negative rate, specificity. PPV: positive predictive value, precision. NPV: negative predictive value. Threshold cut-off for predictions:* $\tau = 0.5$. *We reported the formulas of TPR, TNR, PPV, and NPV in the Supplementary information.*

The use cases UC01 and UC02 show confusion matrices which produce MCCs around 0, and prevalence thresholds close to 1 (Table 1). Namely, the MCC says the classifier behaved like a random guesser, while the prevalence threshold says the prediction was an almost complete failure. By observing the confusion matrices of these two use cases, a machine learning practitioner could wonder which of the two responses is more correct, if the one of the MCC or the one of the prevalence threshold. By checking the values of the four confusion matrix *basic rates* (TPR, TNR, PPV, and NPV), we can notice that UC01 has NPV = 0.977 and very low true positive rate, true negative rate, and precision. And the UC02 confusion matrix, instead, has TNR = 0.99, very low true positive rate and precision, and average NPV. This aspect show that both the use cases UC01 and UC02 have at least one very high score among the four basic rates, which mean that the predictions were not complete failures as the prevalence thresholds would suggest. The values of Matthews correlations coefficients seem more coherent with the results of the four basic rates of the two use cases.

Both the UC03 and UC04 use cases (Table 1) show a very low prevalence threshold (0.031) and an MCC close to zero (+0.07). These discordant values mean that the prevalence threshold deems excellent the binary classification made by the classifiers, while the Matthews correlation coefficient considers it similar to random guessing. Again, by checking the four basic rates of these two sue cases, we can observe that the negative predictive value has an average value, the true negative rate and precision are very high, but the true positive rate is very low, with TPR = 0.01. With its almost perfect score (PT = 0.031), the prevalence threshold hides the poor performance on the true positive rate, and the average outcome on the negative predictive value. The MCC, instead, generates a value related to random guessing, that is far from perfection. Again, the outcome of the Matthews correlation coefficient recaps the results of the four basic rates in a better way than prevalence threshold.

The basic rates of the two remaining use cases show an average true positive rate for UC05 and a slightly high true positive rate for UC06, and a high true negative rate, a high precision, a low negative predictive value for both use cases (Table 1). With these premises, we would expect the recapping scores to produce outcomes meaning poor result. The Matthews correlation coefficient confirms this key-message, by generating values close to zero for both the use cases. On the contrary, the prevalence threshold produces almost perfect outcomes for both the use cases (PT = 0.043 and 0.037): these outcomes of the prevalence threshold are clearly misleading, because they fail to communicate the poor performance of the two classifiers on the negative predictive values.

4.2 MCC and FM

We reported some indicative use cases where the Matthews correlation coefficient and the Fowlkes-Mallows index generate discordant outcomes in Table 2.

The UC07 and UC08 show two confusion matrices with a high number of true negatives and a very low number of true positives (Table 2). UC07 also has many false negatives and almost no false positives; while UC08 has a lot of false positives and almost no false negatives. Both the use cases have values for MCC and FM around zero. The two rates generate discordant outcomes: an MCC around zero in the $[-1, +1]$ interval means the prediction is an average correct classification, and a FM around zero in the $[0, 1]$ range indicates almost complete misclassification. A machine learning practitioner, at this point, might wonder which of the two outcomes is more informative and truthful. By looking at the four basic rates once again, we can notice that the UC07 classifier generates poor true positive rate, high true negative rate, and average precision and negative predictive rate (Table 2). Regarding UC08, its classifier obtains average true positive rate and true negative rate, scarce precision, and perfect negative predictive value. Therefore, the two use cases present a similar situation: one high basic rate, two average basic rates, and one low basic rate. The mean of these four basic rates suggest that the prediction can be considerate correct on

case	TP	TN	FP	FN	TPR	TNR	PPV	NPV	MCC	FM	normMCC	Δ(F,n)
UC07	1	99,001	1	99,001	0	1	0.5	0.5	0	0.002	0.5	0.498
UC08	1	99,001	99,001	1	0.5	0.5	0	1	0	0.002	0.5	0.498
UC09	1	99,001	1	2,001	0	1	0.5	0.98	+0.015	0.016	0.508	0.492
UC10	99,001	1	2,001	1,001	0.99	0	0.98	0.001	−0.013	0.985	0.493	0.492
UC11	99,001	1	1,001	2,001	0.98	0.001	0.99	0	−0.013	0.985	0.493	0.492
UC12	54,001	1	1,001	1,001	0.982	0.001	0.982	0.001	−0.017	0.982	0.491	0.491

Table 2: ***Use cases for comparisons between Matthews correlation coefficient and Fowlkes-Mallows index***. *MCC: Matthews correlation coefficient (Equation 1). normMCC: normalized Matthews correlation coefficient (Equation 2). FM: Fowlkes-Mallows index (Equation 5). TPR, TNR, PPV, NPV, normMCC, and FM have worst value equal to 0 and best value equal to 1. MCC has worst value equal to –1 and best value equal to +1. Δ(F,n): absolute difference between normMCC and FM. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. TPR: true positive rate, sensitivity, recall. TNR: true negative rate, specificity. PPV: positive predictive value, precision. NPV: negative predictive value. Threshold cut-off for predictions:* $\tau = 0.5$. *We reported the formulas of TPR, TNR, PPV, and NPV in the Supplementary information.*

average (like an MCC of 0 suggests), and not close to a complete failure (like a FM of 0.002 indicates).

The use case UC09 is similar to UC07, but with less false negatives and therefore a higher NPV (Table 2). The UC09 basic rates report true negative rate and high negative predictive value, average precision, and very low true positive rate. From these basic rates, it is clear that the classifier performance can be considered moderately good, higher than the average correct classification. By checking the values of the MCC and the FM, we notice that the former reflects the basic rates' outcome with a coefficient of +0.015 in the $[-1, +1]$ interval, while the FM judges the classification extremely bad, with an index of 0.016 in the $[0, 1]$ range.

The UC10, UC11, and UC12 differ from the previous ones because they do not have average values for their four basic rates: they all have high true positive rate and precision, but low true negative rate and negative predictive value. A reasonable consequence of having these four scores would be to have recap metrics indicating a prediction correct on average, not completely wrong and not even completely correct. The Fowlkes-Mallows index, instead, provides an very high score for all the three use cases: 0.985, 0.985, and 0.982. In the $[0, 1]$ FM interval, these three real values indicate almost perfect prediction. However, we know that the three confusion matrices of these three use cases have all poor true negative rate and NPV, so the almost perfect results of the FM looks misleading. The Matthews correlation coefficient, instead, produces values around 0 for all the three use cases, indicating a prediction half correct and half wrong, as the four basic rates suggest.

All the twelve use cases demonstrate that the Matthews correlation coefficient correctly recaps the results of the four basic rates, while the prevalence threshold and the Fowlkes-Mallows index can fail to communicate the poor performance of a classifier in one or more of the four basic rates. The prevalence threshold and the Fowlkes-Mallows index can therefore be misleading for a practitioner.

5 Conclusions

Which rate to employ for the assessment of binary classifications is an open debate in computational statistics and machine learning [30], and no consensus on a single metric has been reached yet, In previous studies [2, 4, 6, 7, 5], we showed the advantages of using the Matthews correlation coefficient (MCC) rather than other statistics, and here we compare this rate with two other popular metrics: the prevalence threshold (PT) and the Fowlkes-Mallows index (FM).

We described and explored the mathematical properties of these two metrics in relationship with the MCC, and tested them on some indicative use cases, where positive data instances and negative data instances have the same importance. From the results observed, we noticed that both prevalence threshold and Fowlkes-Mallows index produced misleading results, hiding low values of at least one basic rate. We therefore confirm, once again, the greater trustworthiness of the MCC, that we recommend to use in any binary classification study.

In the future, we plan to expand this study by considering the behavior of these three statistics in the multi-class classification scenario [31, 32, 33] or in the context of probability threshold reclassification [34].

Additional sections

List of abbreviations

AUC: area under the curve. complPT: complementary prevalence threshold. FM: Fowlkes–Mallows index. FN: false negatives. FP: false positives. MCC: Matthews correlation coefficient. normMCC: normalized Matthews correlation coefficient. NPV: negative predictive value. PPV: positive predictive value, precision. PR: precision-recall. PT: prevalence threshold. ROC: receiver operating characteristic. TN: true negatives. TNR: true negative rate, specificity. TP: true positives. TPR: true positive rate, sensitivity, recall.

Competing interests

The authors declare they have no competing interest.

Ethics approval and consent to participate

Not applicable.

Funding

The authors received no specific funding for this study.

Software availability

Our R and Python software code is publicly available under the GNU General Public License v3.0 at: `https://github.com/davidechicco/MCC_versus_PT_and_FM`

References

[1] Davide Chicco and Giuseppe Jurman. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1):1–23, 2023.

[2] Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. A comparison of MCC and CEN error measures in multi-class prediction. *PLOS One*, 7(8):e41882, 2012.

[3] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(35):1–17, 2017.

[4] Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, 2020.

[5] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):1–22, 2021.

[6] Davide Chicco, Valery Starovoitov, and Giuseppe Jurman. The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access*, 9:47112–47124, 2021.

[7] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, 9:78368–78381, 2021.

[8] Jacques Balayla. Prevalence threshold (ϕe) and the geometry of screening curves. *PLOS One*, 15(10):e0240215, 2020.

[9] Jacques Balayla. Prevalence threshold and bounds in the accuracy of binary classification systems. *arXiv*, preprint arXiv:2112.13289:15, 2021.

[10] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[11] Liora Elfassy, Ariane Lasry, Yaron Gil, and Jacques Balayla. Prevalence threshold of screening tests in obstetrics and gynecology. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 259:191–195, 2021.

[12] M. Krieger, S. Eisenberg, H. Köhler, F. Freise, and A. Campe. Within-herd prevalence threshold for the detection of Mycobacterium avium ssp. paratuberculosis antibody–positive dairy herds using pooled milk samples: A field study. *Journal of Dairy Science*, 105(1):585–594, 2022.

[13] Seungman Cha, Mousab Siddig Elhag, Young-Ha Lee, Dae-Seong Cho, Hassan Ahmed Hassan Ahmed Ismail, and Sung-Tae Hong. Epidemiological findings and policy implications from the nationwide schistosomiasis and intestinal helminthiasis survey in Sudan. *Parasites and Vectors*, 12(1):429, 2019.

[14] Tim Lobstein and Jo Jewell. What is a "high" prevalence of obesity? Two rapid reviews and a proposed set of thresholds for classifying prevalence levels. *Obesity Reviews*, 23(2):e13363, 2021.

[15] Ricardo Campello. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7):833–841, 2007.

[16] Pinar Yildirim, Ljiljana Majnarić, Ozgur Ilyas Ekmekci, and Andreas Holzinger. Knowledge discovery of drug data on the example of adverse reaction prediction. *BMC Bioinformatics*, 15(6):1–11, 2014.

[17] Eduardo H Ramirez, Ramon Brena, Davide Magatti, and Fabio Stella. Probabilistic metrics for soft-clustering and topic model validation. In *Proceedings of WI-IAT 2010 – the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 406–412. IEEE, 2010.

[18] Eduardo H Ramirez, Ramon Brena, Davide Magatti, and Fabio Stella. Topic model validation. *Neurocomputing*, 76(1):125–133, 2012.

[19] Silke Wagner and Dorothea Wagner. Comparing clusterings: an overview. Technical report, Universität Karlsruhe, 2007.

[20] Marina Meilă. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, pages 173–187. Springer, 2003.

[21] Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(1):8574, 2019.

[22] Alicja Rachwał, Emilia Popławska, Izolda Gorgol, Tomasz Cieplak, Damian Pliszczuk, Łukasz Skowron, and Tomasz Rymarczyk. Determining the Quality of a Dataset in Clustering Terms. *Applied Sciences*, 13(5):2942, 2023.

[23] Emir Demirović and Peter J. Stuckey. Optimal Decision Trees for Nonlinear Metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):3733–3741, 2021.

[24] Guanis de Barros Vilela Junior, Bráulio Nascimento Lima, Heros Ferreira Ribeiro, Marcelo Francisco Rodrigues, Adriano de Almeida Pereira, José Ricardo Lourenço de Oliveira, Luís Felipe Silio, and Ricardo Pablo Passos. Importância do índice Fowlkes-Mallows (FMI), do coeficiente de correlação de Matthews (MCC) e do índice Youden (IY) nos classificadores de inteligência artificial na área da saúde. *Centro de Pesquisas Avançadas em Qualidade de Vida*, 14(v14n2):1, 2022.

[25] Jacques Balayla, Ariane Lasry, Yaron Gil, and Alexander Volodarsky-Perel. Prevalence threshold and temporal interpretation of screening tests: the example of the SARS-CoV-2 (COVID-19) pandemic. *medRxiv*, page 2020.05.17.20104927, 2020.

[26] Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure*, 405(2):442–451, 1975.

[27] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[28] Jan Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5–6):367–374, 2004.

[29] Davide Chicco and Giuseppe Jurman. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Scientific Reports*, 10(1):1–12, 2020.

[30] Kjetil Dyrland, Alexander Selvikvåg Lundervold, and PierGianLuca Porta Mana. Does the evaluation stand up to evaluation? A first-principle approach to the evaluation of classifiers. Center for Open Science, 2022.

[31] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[32] Mahendra Sahare and Hitesh Gupta. A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(3):160, 2012.

[33] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. *arXiv preprint arXiv:1901.00544*, 2019.

[34] Yu-Heng Lai, Wei-Ning Chen, Te-Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports*, 10(1):4679, 2020.

Supplementary information

Binary statistical rates

List of statistical rates to evaluate confusion matrices and their formulas:

$$\text{F}_1\ \text{score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{true positive rate, TPR, recall, sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{true negative rate, TNR, specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{positive predictive value, PPV, precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{negative predictive value, NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{11}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{balanced accuracy, BA} = \frac{\text{TPR + TNR}}{2} \tag{12}$$

(minimum and worst value = 0; maximum best value = 1)

$$\text{bookmaker informedness, BM} = \text{TPR + TNR − 1} \tag{13}$$

(minimum and worst value = -1; maximum best value = $+1$)

$$\text{markedness, MK} = \text{PPV + NPV − 1} \tag{14}$$

(minimum and worst value = -1; maximum best value = $+1$)

$$\text{diagnostic odds ratio, DOR} = \frac{\text{TP} \cdot \text{TN}}{\text{FP} \cdot \text{FN}} \tag{15}$$

(minimum and worst value = 0; maximum and best value = ∞)

$$\text{Cohen's } \kappa \ = \frac{2 \cdot (\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP} + \text{FP}) \cdot (\text{FP} + \text{TN}) + (\text{TP} + \text{FN}) \cdot (\text{FN} + \text{TN})} \tag{16}$$

(minimum and worst value $= -1$; maximum and best value $= +1$)

$$\text{Precision-Recall (PR) curve} = \begin{cases} true\ positive\ rate & on\ the\ x\ axis \\ positive\ predictive\ value & on\ the\ y\ axis \end{cases} \tag{17}$$

(minimum and worst value $= 0$; maximum best value $= 1$)

$$\text{ROC curve} = \begin{cases} false\ positive\ rate & on\ the\ x\ axis \\ true\ positive\ rate & on\ the\ y\ axis \end{cases} \tag{18}$$

(minimum and worst value $= 0$; maximum best value $= 1$)