IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.DOI

The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment

DAVIDE CHICCO*[1], MATTHIJS J. WARRENS[2], and GIUSEPPE JURMAN[3]
[1]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[2]Groningen Institute for Educational Research, University of Groningen, Groningen, Netherlands
[3]Data Science for Health Unit, Fondazione Bruno Kessler, Trento, Italy

*Corresponding author: Davide Chicco (e-mail: davidechicco@davidechicco.it).

ABSTRACT Even if measuring the outcome of binary classifications is a pivotal task in machine learning and statistics, no consensus has been reached yet about which statistical rate to employ to this end. In the last century, the computer science and statistics communities have introduced several scores summing up the correctness of the predictions with respect to the ground truth values. Among these scores, the Matthews correlation coefficient (MCC) was shown to have several advantages over confusion entropy, accuracy, F_1 score, balanced accuracy, bookmaker informedness, markedness, and diagnostic odds ratio: MCC, in fact, produces a high score only if the majority of the predicted negative data instances and the majority of the positive data instances are correct, and therefore it results being very trustworthy on imbalanced datasets. In this study, we compare MCC with two other popular scores: Cohen's Kappa, a metric that originated in social sciences, and the Brier score, a strictly proper scoring function which emerged in weather forecasting studies. After explaining the mathematical properties and the relationships between MCC and each of these two rates, we report some use cases where these scores generate different values, which lead to discordant outcomes, where MCC provides a more truthful and informative result. We highlight the reasons why it is more advisable to use MCC rather that Cohen's Kappa and the Brier score to evaluate binary classifications.

INDEX TERMS Matthews correlation coefficient; Cohen's Kappa; binary classification; confusion matrix; supervised machine learning; Brier score; confusion matrix; applied machine learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

I. INTRODUCTION

Two-class binary classification is a popular task in machine learning and computational statistics. When the goal of the study is to classify or predict elements in groups, usually the practitioner assigns labels 0 and 1 to them in the original ground truth dataset. The data instances with label 0 are usually called *negatives*, while the data instances labeled 1 are usually called *positives*.

A trained classifier then makes a prediction by associating a real or binary value to each element of the ground truth dataset. If the values are real, they are often made binary by assigning the value 0 to the predictions that are below a specific cut-off threshold τ (usually equal to 0.5) and by assigning the value 1 to the predictions that are greater than or equal to that threshold (prediction $\geq \tau$). This way, both the ground truth elements and the predictions can be split into positives and negatives. At this point, a two-class confusion matrix can be created:

- The actual positives that are correctly predicted positives are called true positives (TP);
- The actual positives that are wrongly predicted negatives are called false negatives (FN);
- The actual negatives that are correctly predicted negatives are called true negatives (TN);
- The actual negatives that are wrongly predicted positives are called false positives (FP).

Each of these four categories contains a quantitative number that can be important for the study carried on; considering all the four tallies together, however, can be complicated and uneasy. For this reason, scientific researchers have invented several metrics able to recap the quantitative information of a confusion matrix or of the original predictions themselves.

The Matthews correlation coefficient [1], in particular, is a rate that resulted being more informative than confusion entropy (CEN) [2], accuracy and F_1 score [3], balanced accuracy, bookmaker informedness, and markedness [4], and diagnostic odds ratio [5] in the past (Supplementary Information). In this study, we decided to continue this series of comparisons by confronting MCC with another two-class confusion matrix rate (Cohen's Kappa), and with a strictly proper score function representing the original predictions of a classifier (Brier score).

Matthews correlation coefficient (MCC). The Matthews correlation coefficient has been introduced by Brian W. Matthews to evaluate the predicted structure of an enzyme, in a biochemical study in 1975 [1]. Since then, it has been used in several studies, but has never become as popular as accuracy and F_1 score in the mathematics and computer science communities [3]. The situation changed after 2000, when MCC was reproposed as a standard metric for binary classification by Baldi and colleagues [6] and its spread started to grow.

Since then, for example, MCC has been used as a standard metric in several scientific competitions, such as the Kaggle competition to detect power line fault detection [7] and the DataDriven challenge to identify clogged blood vessels in the brain of mice with Alzheimer's dementia [8]. Additionally, MCC has been included in DREAMTools [9], a Python package to assess results of collaborative DREAM challenges [10], and can be found on several software packages of free open source programming languages such as `Python`, `R`, and `TensorFlow`.

The Matthews correlation coefficient gained popularity when the US Food and Drug Administration (FDA) agency employed it as the main evaluation metric in the MicroArray / Sequencing Quality Control (MAQC/SEQC) comprehensive analyses in 2010 and 2014 [11], [12].

Recently, Boughorbel and colleagues [13] described an enhanced classifier based on the Matthews correlation coefficient, while Zhu [14] investigated the behavior of MCC on several imbalanced cases.

With the growing spread of the Matthews correlation coefficient [15], [16], specialized blogs about machine learning and technology started to discuss this rate, too. For example, articles on MCC appeared on the blog of Towards Data Science [17] and on the blog of the graphic designer David Lettier [18].

For 2×2 confusion matrices MCC is identical to the *phi* (ϕ) coefficient [19]–[21]. Other generalizations of the *phi* coefficient were proposed in Janson and Vegelius [22] and Gorodkin [23]. As *phi* coefficient, the Matthews correlation coefficient is employed often in psychometrics [24].

Cohen's Kappa. The Kappa coefficient is a metric for summarizing the agreement between two nominal classifications, based on the same categories. It is extensively used in social, behavioral and medical sciences, as a measure of agreement between two raters [25]–[28]. It was first introduced by Jacob Cohen in 1960 as an alternative metric to accuracy that considers agreement due to chance [29]. The Kappa coefficient can be interpreted as a measure of agreement beyond chance compared to the maximum possible beyond chance agreement [30], [31].

Originally, Kappa was designed for classifications with more than two classes [29], [32]–[35]. Nevertheless, it is commonly applied to two-class classification problems too [36], [37]. Similar to MCC, Cohen's Kappa considers all the four categories of the binary classification confusion matrix: true positives, true negatives, false positives, and false negatives. Furthermore, both metrics are balanced measures that summarize the classification problem in one value [38] and have value equal to $+1$ in the case of perfect prediction (except for indeterminate cases) and 0 if the prediction is random.

It can be shown that Cohen's Kappa is equivalent to the Hubert-Arabie adjusted Rand index [39], that has been employed in cluster analysis for quantifying agreement between two partitions [40]. Furthermore, the relationship between Cohen's Kappa and operating characteristic curves (ROC) has been explored by Ben-David [41].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

Several authors have presented population models for Cohen's Kappa [42], [43]. Under several of these models, Kappa can be interpreted as an association coefficient. However, Kappa is also commonly used as a sample statistic or performance measure, for example, when calculating Kappa for a sample of subjects is one step in a series of research steps, or when Kappa is used for analyzing a binary classification. In these cases, researchers can usually be interested in the agreement in the sample, not in the agreement of a population. In the case of 2×2 confusion tables, the test statistic for Cohen's Kappa is the same as Pearson's chi-squared (χ^2) test [44]. Tables for sample size determination for a variety of common study designs involving Cohen's Kappa can be found in a study of Cantor [45], and standard errors for Cohen's Kappa can be found in works of Garner [46] and Shan and Wang [47].

As a sample statistic, Cohen's Kappa is known to be marginal or prevalence dependent since it takes the class sizes into account [48]–[52]. In social sciences, it is well known that the value of Kappa depends on the prevalence of the class being diagnosed. In the 2×2 case values of Kappa can be quite low if one class is quite common or very rare [53], [54]. Various authors have shown that if two pairs of binary classifications have the same accuracy, the pair whose class distributions are more similar to each other may have a lower Kappa value than the pair with more divergent class distributions [53], [55]. Since binary classifications with similar class distributions usually have a higher amount of agreement expected to occur by chance, a fixed accuracy will lead to a lower Kappa value due to the definition of the statistic [56]. The dependence of Cohen's Kappa on the class distributions has been studied extensively by means of examples of 2×2 confusion tables in the literature [50], [51], [53], [54]. Warrens [57] presented exact formulations of many of these properties and observations. In general, the use of Kappa is accepted: its pitfalls can be overcome by considering the class distributions. Nevertheless, multiple researchers have proposed alternative metrics for 2×2 confusion tables [54], [55], [58].

The popularity of Cohen's Kappa has led to the development of various extensions, including weighted Kappa coefficients for classifications with three or more ordered classes [59]–[63], Kappa coefficients for three or more observers or classifications [64], and a Kappa coefficient that can handle missing data [65]. Inequalities between different weighted Kappa variants for ordered classes have been discussed in studies of Warrens [28], [34]. Furthermore, various authors have found applications of Cohen's Kappa that are different than the original context considered by Cohen. For example, Chang [66] used Cohen's Kappa to capture discrimination in the same way as the receiver operating characteristic curve. Holle and Rein [67] employed Cohen's Kappa to assess agreement for segmentation and annotation. Vieira and coauthors [68] used Cohen's Kappa as a performance measure for feature selection.

Other studies describe the drawbacks of Cohen's Kappa in remote sensing [69], [70]. Stein et al. [69] saw the Cohen's Kappa single-value as a flaw, incapable to express the overall assessment of the classification. Instead, they proposed the Bradley-Terry model, that gives information on the separate categories and not just a single number. The Bradley-Terry model could be useful for the multi-class predictions, but not for binary classifications.

Pontius and Millones [70] criticized the Kappa statistic because it can generate values that do not make sense in remote sensing, and stated that Kappa coefficient's statistically expected agreement can be irrelevant for the same domain. Instead, Pontius and Millones [70] proposed two alternative metrics (quantity disagreement and allocation disagreement) as an alternative to Cohen's Kappa that can be used complementary to accuracy in remote sensing applications [71].

Brier score. Unlike Cohen's Kappa and the Matthews correlation coefficient, the Brier score is a strictly proper scoring rule and hence favours probability forecasts that are well calibrated. Similarly to the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and of the precision-recall (PR) curve, the Brier score does not consider a specific cut-off threshold to split the predicted values into positives and negatives. The predicted values used for the Brier score are usually forecast probabilities, differently from AUC. For example, AUC is unchanged if the probabilities are transformed monotonically. We usually refer to AUC as measuring only discrimination whereas strictly proper scoring rules like the Brier score are influenced by both the discriminating ability of the forecasts and their calibration, where *calibration* here means the relative frequency of observed outcomes [72]. For example, a perfect calibration happens when a claim predicts an event to appear with a 70% likelihood, and that event actually occurs 70% of those times [72]. Calibration is important if the forecasts are going to be taken at face value by users.

With regard to classification, the Brier score can be interpreted as the loss expected for a uniform distribution of cost-loss ratios when the classification is made by applying the Bayes decision rule to the forecasts. Accuracy relates to the loss expected when classification is made using a fixed threshold, and ROC AUC relates to the loss expected for another method of choosing the threshold [73]. Thus the Brier score is a useful measure of the performance of the classifier that we would create if we were to trust the forecast probabilities (that is, if we were to assume that the forecasts are calibrated and so consider the Bayes rule optimal). If the forecasts are not calibrated, however, then it may be possible to achieve better classifier performance by using other decision rules.

The Brier score was originally introduced by Glenn W. Brier in 1950 for weather forecasting related to the probability of rain [74]. Several decades later, a few researchers investigated the mathematical details of this cost function: Blattenberger and Lad [75] presented a graphical description of the separation into distinct calibration and refinement components of the Brier score, while Murphy and colleagues [76] described a decomposition of the Brier score based on conditional distributions and mean errors.

Almost twenty years later, the Brier score came back to the attention of the statistics and weather community with several articles published in the same period. Ikeda et al. [77] studied the relationships between the Brier score and binormal receiver operating characteristics (ROC) area under the curve (AUC), while in his preprint Jewson [78] described some clear issues regarding the Brier score in weather forecasting.

Gerds and Schumacher [79] described their findings when employing the Brier score for survival analysis. Another meteorological application regards the study of Casati and colleagues [80], who employed the Brier score to forecast lightnings.

Roulston [81], Stephenson and colleagues [82], and Ferro et al. [83] investigated some mathematical properties of the Brier score. Bradley and colleagues [84] explored the sampling uncertainties of the Brier score and its variant Brier skill score [85].

Rufibach published a short report [86] where he described the advantages of the Brier score for binary predictions over Spiegelhalter's z-statistic [87], while Jachan and colleagues [88] described a biomedical case study where they used the Brier score to assess predictions of epileptic seizures.

Johansson and coauthors [89] investigated how to use the Brier score for existing rule extraction, and applied their methods on 26 datasets of the University of California Irvine Machine Learning Repository [90].

The theme of the Brier score decomposition was treated again in the correspondence article of Young [91], in an correspondence article by Ferro and Fricker [92], in a letter by Siegert [93], and in a study by Merkle and Hartman [94].

Hernandez-Orallo and colleagues [95] proposed a curve based on the Brier score as an alternative to traditional curves such as receiver operating characteristics (ROC) or precision-recall (PR) curve. Lesik and Leake [96] described an application of the Brier score to assess the placement of students among mathematics courses after Scholastic Assessment Test (SAT) examinations.

A recent article by Assel and coauthors [97] claims that the Brier score is incapable of predicting diagnostic tests or prediction models in clinical environments.

The application fields. Although the three metrics (MCC, κ, Brier score) share a common statistically grounded origin in their definition, they faced a different evolution in their usage in the following years. The κ statistic originated in the social sciences and then became of general purpose, being commonly used in all research fields whenever the level of agreement between two nominal classifications is investigated. The Brier score was originally introduced in weather forecasting studies, but its usage has become increasingly widespread as a risk score in survival and prediction models in medicine, being nowadays its elective application field. Oppositely, MCC was originally conceived as a performance metric for classifiers in biochemistry and as such it has been used in several biomedical domains in the following years, becoming quite common in bioinformatics and computational biology. In the last years, its popularity has overcome the life science limits, and its use is spreading across all scientific and technological disciplines.

To the best of our knowledge, no study comparing MCC, Cohen's Kappa, and the Brier score has been released in the scientific literature so far; we fill this gap by presenting the current study on these three statistical rates.

This study. We organized the rest of this article as follows. After this Introduction, we explain the mathematical background of MCC, Cohen's Kappa, and the Brier score (section II). Afterwards, we describe the relationship between MCC and Cohen's Kappa and the relationship between MCC and the Brier score (section III), and discuss some use cases where these pairs of rates give discordant messages (section IV). At the end of the article, we outline some conclusions and future developments (section V).

II. MATHEMATICAL BACKGROUND

Matthews correlation coefficient. The Matthews correlation coefficient (MCC) [1] is a case of the Cramér's V [19] applied to a 2×2 traditional confusion matrix, having true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) (Equation 1). The metric is defined as:

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

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

MCC is class symmetric: switching positives and negatives would lead to the same result. The minimum value of MCC is –1, meaning perfectly wrong prediction, where a classifier labels all the positives as negatives and all the negatives as positives. The maximum value of MCC is +1, which means perfect classification. If the value of MCC is around 0, it means that the prediction made was similar to random guessing. The Matthews correlation coefficient can be undefined when a pair of confusion matrix values are both 0, but these cases can be handled with some mathematical steps [3].

Cohen's Kappa. Cohen's Kappa [29] was originally proposed for quantifying agreement between two observers that judged the same set of persons on a nominal scale, with two or more classes. The metric is also commonly used for two-class classification problems. Using the cells of a 2×2 traditional confusion matrix Cohen's Kappa [27], [40], [42] is defined as:

$$\kappa = \frac{2\cdot(TP\cdot TN - FP\cdot FN)}{(TP+FP)\cdot(FP+TN)+(TP+FN)\cdot(FN+TN)} \quad (2)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

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

Cohen's Kappa shares various properties with MCC. Both these rates are class symmetric, their minimum value is –1 (perfectly wrong prediction) and their maximum value is +1 (perfect classification). Furthermore, if $\kappa \approx 0$, the prediction made was similar to random guessing. Finally, κ can be undefined in some cases, but these cases can be handled with mathematical operations similar to the ones needed when MCC is undefined [3].

In 1960, Cohen's Kappa was originally proposed as a chance-corrected measure, more precisely a chance-corrected version of accuracy. The metric in Equation 2 is equivalent to:

$$\kappa = \frac{\text{accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}} \tag{3}$$

where the formula of accuracy is given by:

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$\text{(worst value} = 0\text{; best value} = 1)$$

and where the formula of expected accuracy is given by:

$$\begin{aligned} \text{expected accuracy} &= \\ &= \left(\tfrac{TP+FP}{N} \cdot \tfrac{TP+FN}{N}\right) + \\ &+ \left(\tfrac{TN+FP}{N} \cdot \tfrac{TN+FN}{N}\right) \end{aligned} \tag{5}$$

where N is the number of samples in the dataset. The formula of expected accuracy (Equation 5) is the value of accuracy (Equation 4) under statistical independence of the observers (or two nominal variables). In inter-rater reliability studies, accuracy is generally considered artificially high since some agreement might be due to chance. Therefore, it makes sense to use a measure that takes this aspect into account.

Various authors later discovered that Cohen's Kappa may be interpreted as chance-corrected version of various measures other than accuracy in Equation 4 [33]. In fact, all special cases of:

$$M(\alpha) = \frac{\alpha \cdot TP + (2-\alpha) \cdot TN}{\alpha \cdot TP + FP + FN + (2-\alpha) \cdot TN} \tag{6}$$

$$\text{(worst value} = 0\text{; best value} = 1)$$

become Cohen's Kappa after correction for agreement due to chance [33]. Two examples are the F_1 score ($\alpha = 2$) and accuracy ($\alpha = 1$). The special case for $\alpha = 0$ was studied by Cicchetti and Feinstein [54].

Brier score. The Brier score [74] is a strictly proper scoring function that is equivalent to the mean squared error:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \tag{7}$$

$$\text{(worst value} = 1\text{; best value} = 0)$$

where N is the number of samples in the dataset, x_i is the predicted value for the ith element and y_i is the actual value of the ith element.

In the general case when x_i is an actual probability, a comparison to MCC and κ can be difficult to interpret, since the two aforementioned measures are applicable only in the hard classification cases when x_i is binarized to correspond to one of the two class labels.

In particular, reducing to the case where the ground truth values are zeros and ones, since the prediction probability range in the [0, 1] interval, by setting the confusion matrix threshold τ is set to 0.5, the Brier score can be expressed through traditional two-class confusion matrix classes. We call this Brier score binary variant binaryBS:

$$\text{binary}BS = \frac{FP + FN}{TP + FP + FN + TN} = 1 - \text{accuracy} \tag{8}$$

$$\text{(worst value} = 1\text{; best value} = 0)$$

binaryBS is the complementary value of accuracy and, like the original Brier score, has its best value equal to 0 (perfect prediction) and its worse value equal to 1 (prediction with maximum errors possible).

III. RELATIONSHIPS BETWEEN RATES

In this section, we first study the mathematical relationships and correlations between the Matthews correlation coefficient and Cohen's Kappa, and then between the Matthews correlation coefficient and the Brier score.

A. MCC AND COHEN'S KAPPA

The formulas of MCC in Equation 1 and Cohen's Kappa in Equation 2 have a number of features in common. We have MCC $= \kappa$ if and only if $FP = FN$, that is, the metrics coincide when the 2×2 confusion matrix is symmetric. Furthermore, MCC and Kappa are, respectively, the geometric mean and harmonic mean of the following quantities:

$$\frac{TP \cdot TN - FP \cdot FN}{(TP+FP) \cdot (FP+TN)} \quad \text{and} \quad \frac{TP \cdot TN - FP \cdot FN}{(TP+FN) \cdot (FN+TN)}. \tag{9}$$

From the geometric-harmonic-means inequality we obtain the inequality $\|\text{MCC}\| \geq \|\kappa\|$ [37], [38]. From this inequality it follows that the Kappa value will always be closer to 0 than the MCC value: the Kappa value will always be equal or less extreme. In turn, this implies that, in the case of positive association (that is: $TP \cdot TN \geq FP \cdot FN$), it is impossible that Kappa produces a higher value than MCC in the case of a binary classification [37], [38].

Since MCC $= \kappa$ if and only if $FP = FN$, the largest differences between MCC and Kappa are quite likely to be found when FP and FN are very different, which is more likely when the metrics produces negative values. To highlight this aspect, we depicted a scatterplot with all the possible values of the Matthews correlation coefficient on the x axis and all the possible values of Cohen's Kappa on the y axis (Figure 1), both in the $[-1, +1]$ interval.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

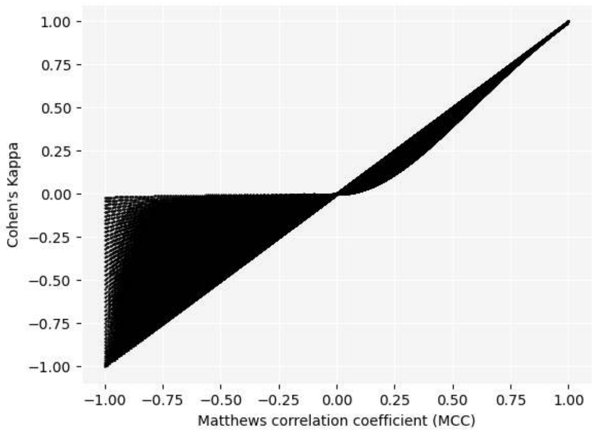

FIGURE 1: **Relationship between MCC and Cohen's Kappa**. We computed MCC and Cohen's Kappa for 10^3 possible confusion matrices.

As one can notice, MCC and κ have almost identical values in the top-right quarter, that is where the values of both MCC and κ are positive (Figure 1). In the $[0, +1]$ interval, in fact, the two rates are generally concordant, showing the same trend and minimal differences between values. The top difference of 0.11 can be noticed when MCC equals to +0.339 and κ equals to +0.229, as we discuss later (section IV). A difference of 0.11 between MCC and κ means a 5% difference in the total range of 2, so we can consider that minimal.

On the contrary, MCC and Cohen's Kappa show very different behavior on the bottom-left quarter, that corresponds to the values in the $[-1, 0]$ interval (Figure 1). To a MCC of –1, for example, can correspond any negative value of κ. This ambiguity results being very strong, because both these rates have different meanings for 0 and for –1: a value close to zero, in fact, means that the prediction is similar to random guessing, while a value close to –1 means perfect opposite prediction. Note that these values can happen when the predictor generated no true positive and no true negative. We discuss this scenario later in several use cases (section IV).

Finally, the inequality $\|\text{MCC}\| \geq \|\kappa\|$ does not hold for the case of multi-class classification. Delgado and Tibau [38] presented various cases in which a worse classifier gets a higher Kappa value, differing qualitatively from the MCC value, although in most cases the two metrics produce similar values.

B. MCC AND BRIER SCORE

The Brier score has a huge difference from MCC and Cohen's Kappa: it is a strictly proper score function with values ranging from 0 (perfect prediction) to 1 (worst prediction). Therefore, the Brier score is not generated by the two-class confusion matrix categories, but rather as the cumulative sum of the squared mean error computed between the predicted values and the ground truth values (Equation 7).

If one wanted to investigate the relationship between MCC and the Brier score through FP, FP, TN, and TP, she/he would therefore need to use binaryBS (Equation 8) instead of the original Brier score. As we mentioned earlier, binaryBS is a variant of accuracy, and therefore has the same properties. The relationships between MCC and accuracy have been already investigated in previous study [3].

For this reason, to investigate the relationship between MCC and the Brier score, we decided to focus on scatterplots having these two rates on the x axis and y axis. To generate proper scatterplots, we first had to find a way to generate a reasonable set of predictions. Following the example of Cao and colleagues [98] for the MCC-F1 curve, we used Beta distributions [99], that are probability distributions controlled by two shape parameters. Beta distributions generate real values in the $[0, 1]$, like a traditional machine learning classifier. By changing the two shape parameters, we simulated various different classifiers.

Figure 2 presents three example classifiers based on the Beta distributions. When the two shape parameters have identical values, for example Beta(4, 4), the beta distribution is symmetric and a majority of simulated prediction scores will be scattered around 0.5. If the shape parameters are quite distinct, the majority of simulated scores will be closer to 0 (for example, Beta(9, 15)) or 1 (for example, Beta(15, 8)).

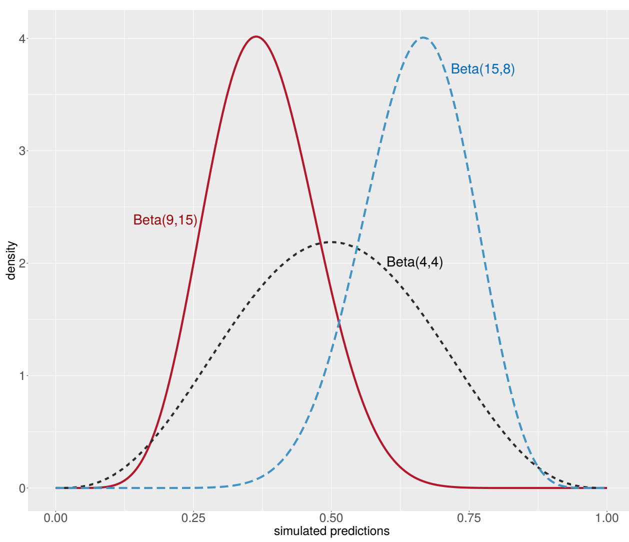

FIGURE 2: **Beta distributions plot**. Three example simulated classifiers based on Beta distributions [99].

Regarding the ground truth, we employed three synthetic datasets: a balanced dataset with 5,000 positives and 5,000 negatives; a negatively imbalanced dataset with 1,000 positives and 9,000 negatives; and a positively imbalanced dataset: 9,000 positives and 1,000 negatives. Regarding the simulated classifiers, we generated two groups of predictions: in the first case (symmetric simulated predictions), we asso-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

ciated a particular Beta distribution to the positives, and a particular Beta distribution to the negatives; in the second case (asymmetric simulated predictions), we associated a particular Beta distribution to the positives, a particular Beta distribution to the first 70% of the negatives, and a particular Beta distribution to the last 30% of the negatives.

Symmetric simulated predictions. In this case, we associated to the positive data instances the values $Beta(a, b)$ distribution and associated to the negative data instances the values $Beta(c, d)$ distribution with a, b, c, d ranging from 1 to 15. Since the worst value of MCC is –1 and the best value of MCC is +1, while the Brier score is best when its value is 0 and worst if the value is +1, we prefered to employ the normalized MCC and the complementary Brier score for these plots. Both the normalized MCC ($normMCC = (MCC + 1)/2$) and the complementary Brier score ($complBS = 1 - BS$) range in the $[0, 1]$ interval, and have 0 as worst possible score and 1 as best possible score.

We computed all the possible classifiers varying a, b, c, d, and depicted the values of MCC and the Brier score in a scatterplot (Figure 3).

As one can notice, both normMCC and complBS have different behaviors in the three plots (Figure 3).

In the balanced dataset plot (Figure 3A), the two measures are fairly concordant, generating a thin plot that behaves like a $x = y$ function scaled-up on the y axis. This plot shows also that complBS is always higher than normMCC in this case. Regarding the association between scores, one can notice that multiple values of normMCC correspond to few values of complBS: when complBS is around 0.6, all the points having normMCC in the $[0.1, 0.5]$ range are associated to it. Some values of normMCC relate to multiple values of complBS, too, but in a smaller interval: when normMCC is around 0.48, the complBS values range in the $[0.45, 0.7]$ interval. This trend means that: multiple values of the Brier score correspond to many values of the Matthews correlation coefficient; few values of the Matthews correlation coefficient correspond to many values of the Brier score. Both these behaviors can generate discordant or ambiguous messages about the binary classification assessment, especially regarding the Brier scores that could mean both excellent MCC and poor MCC in the same time. We will deal with this issue more in detail in the use cases section (section IV).

The negatively imbalanced dataset plot (Figure 3B) results being identical to the positively imbalanced dataset plot (Figure 3C), and this aspect comes with no surprise since both the Brier score and the Matthews correlation coefficient are class-invariant: differently from F_1 score, inverting positives with negatives in the original datasets would not change the scores for MCC and the Brier score.

These two plots show several differences from the balanced dataset plot. Their points occupy almost completely the lower-left quadrant, precisely the area where complBS is in the $[0.2, 0.5]$ range and normMCC is in the $[0.2, 0.5]$ interval. Another area dense of points can be observed where normMCC equals to 0: for this normMCC value, complBS can have values that go from 0.8 to 0.2. This aspect means that there is an large multiplicity of normMCC-complBS associations in that area, which can lead again to ambiguous and discordant messages.

Asymmetric simulated predictions. The previously described scatterplots between MCC and Brier score (Figure 3) have a symmetry between the positives and the negatives: we associated a particular Beta distribution to all the ground truth positive data instances, and another particular Beta distribution to the ground truth negative data instances.

To investigate a different case, similarly to what [98] did, we generated additional simulated classifiers with a change compared to before: we associated the values of a Beta distribution to the first 70% of the negative elements, and the values of a different Beta distribution to the last 30% of the negative elements. While we kept the values of $Beta(a, b)$ associated to the positive data instances, we used the values of $Beta(c, d)$ for the first 70% of the negatives and $Beta(e, f)$ for the last 30% of the negatives, with a, b, c, d, e, f ranging from 1 to 15.

We computed all the possible classifiers varying a, b, c, d, e, and f, and depicted the values of MCC and Brier score in a scatterplot (Figure 4).

As one can notice, the balanced dataset plot (Figure 4A) looks similar to its corresponding plot in the symmetric case (Figure 3A): a concordant trend scaled up from the $x = y$ line. The negatively imbalanced dataset plot (Figure 4B), also, shows a trend similar to the trend of the symmetric case (Figure 3B).

The MCC-Brier score plot of the positively imbalanced dataset has some significant differences from the previous ones (Figure 4C). As one can notice, the scatterplot cloud is wider: that means that a specific value of complBS corresponds to many values of normMCC, although with different widths. When complBS is approximately 0.3, for example, normMCC can range between 0.1 and 0.6. This scatterplot cloud is also longer than the other plots around normMCC = 0.6: this specific value corresponds to all the complBS between 0.625 and 0.8, approximately.

To conclude, the plots on the negatively imbalanced dataset (Figure 3B and Figure 4B) and the plots on the positively imbalance datasets (Figure 3C and Figure 4C) show clearly that:

- Several values of the Brier score correspond to a huge number of the Matthews correlation coefficients, generating ambiguous messages: cases where the Brier score indicates very good prediction, and MCC indicates poor prediction, and vice versa;
- Several values of the Matthews correlation coefficient correspond to many Brier scores, generating ambiguous messages, too: cases where the Brier score indicates very good prediction, and MCC indicates poor prediction, and vice versa.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

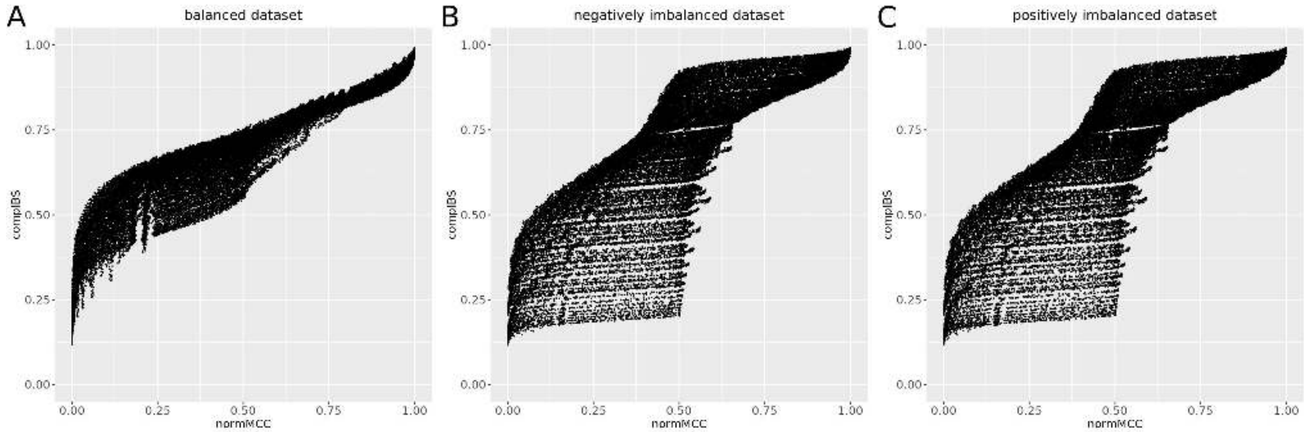

FIGURE 3: **Relationship between MCC and the Brier score, with simulated classifiers using same distributions on positives and negatives**. We report all the 50,625 points representing the complementary Brier score the normalized MCC generated by Beta distribution simulated classifiers on simulated datasets. (A) Balanced dataset: 5,000 positives and 5,000 negatives. (B) Negatively imbalanced dataset: 1,000 positives and 9,000 negatives. (C) Positively imbalanced dataset: 9,000 positives and 1,000 negatives. Simulated classification points associated to the positives: Beta(a, b) with a and b ranging from 1 to 15. Simulated classification points associated to the negatives: Beta(c, d) with c and d and f ranging from 1 to 15. normMCC = (MCC + 1)/2. complBS = 1 – BS. The values of both normMCC and complBS lay in the $[0, 1]$ interval, with worst value equal to 0 and best value equal to 1.

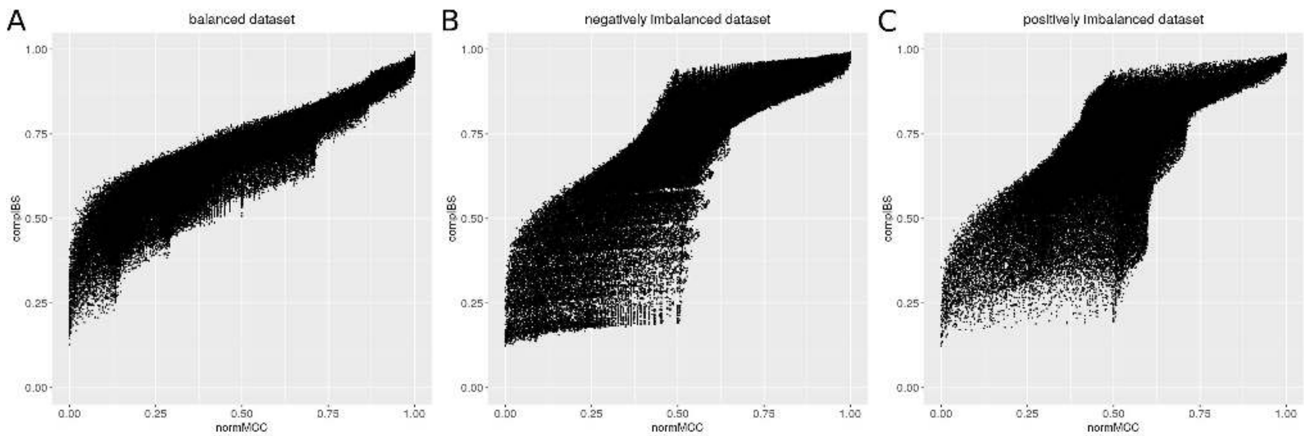

FIGURE 4: **Relationship between MCC and Brier score, with simulated classifiers using the different distributions on positives and negatives**. We report 100,000 randomly selected points representing the complementary Brier score the normalized MCC generated by Beta distribution simulated classifiers on simulated datasets. (A) Balanced dataset: 50 positives and 50 negatives. (B) Negatively imbalanced dataset: 10 positives and 90 negatives. (C) Positively imbalanced dataset: 90 positives and 10 negatives. Simulated classification points associated to the positives: Beta(a, b) with a and b ranging from 1 to 15. Simulated classification points associated to the negatives: Beta(c, d) for the first 70% and Beta(e, f) for the last 30%, with c, d, e, and f ranging from 1 to 15. normMCC = (MCC + 1)/2. complBS = 1 – BS. The values of both normMCC and complBS lay in the $[0, 1]$ interval, with worst value equal to 0 and best value equal to 1.

In the balanced dataset (Figure 3A and Figure 4A), instead, both Brier score and MCC show concordant trends, with much smaller ambiguity. To each value of the Matthews correlation coefficient, in fact, correspond a few values of the Brier score.

1) The ambiguity when the Brier score ≈ 0.25

There is a special case of the Brier score where the ambiguity of its message, compared with MCC, is at its maximum: when the Brier score is approximately 0.25. Consider a binary classification tasks on a dataset with n_+ positive samples and n_- negative samples. To simplify notation when using the Brier score, label the positive class as 1 and the negative class as 0. Let ε be a real number in the interval

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

$[0, 0.5)$ and suppose the output of a probabilistic classifier is $1-\varepsilon$ for the samples of the positive class, and $0+\varepsilon$ for each negative sample. Then, by binarizing the output on the two classes 0 and 1, classification is perfect, thus MCC $= +1$ regardless the value of $0 \leq \varepsilon < 0.5$, while BS $= \varepsilon^2$. Thus, MCC is always one, while the Brier score can range between 0 and 0.25 (excluded).

Symmetrically, suppose that another classifier gives $1-\varepsilon$ as the prediction for each negative sample, and $0+\varepsilon$ for each positive sample. Then, in this case, MCC is always –1, while BS $= (1-\varepsilon)^2$ and thus it can range between 0.25 (excluded) and 1.

It follows that values of the Brier score very close to 0.25 can correspond to either perfect binary classification or full misclassification, as we will show later for the use cases BS7 and BS8.

IV. USE CASES

After having investigated the relationships between MCC and Cohen's Kappa and between MCC and Brier score, here we analyze some concrete use cases where each pair of scores generates a discordant outcome.

In these use cases, we consider the values of TP, TN, FP, and FN resulting from binary classifications when the threshold τ that discriminates between positive predictions and negative predictions equals 0.5, which is a cut-off commonly employed in machine learning and computational statistics. Some studies use alternative cut-off thresholds, through a phase called *reclassification* [100]; although interesting, the analysis of this topic goes beyond the scope of the present study.

A. MCC AND COHEN'S KAPPA USE CASES

As mentioned earlier, MCC and κ generate a concordant response in the $[0, +1]$ quarter, while they might have discordant values in the $[-1, 0]$ area of the plot of all the possible values.

To this end, we found six use cases where the classifier had no true positive and no true negative, and the value of MCC was –1 (K1, K2, K3, K5, and K6 in Table 1).

In K1, for example, MCC equals to –1, while κ equals to 0. In this case, the two rates generate a discordant message: the Matthews correlation coefficient states that the classifier made a prediction that is the opposite of the ground truth, while Cohen's κ states it was similar to random guessing. Checking the confusion matrix, we can see that TP, FP, and TN are all zero, and therefore we can confirm that the classification was perfectly wrong. In this case, MCC gave a more informative and truthful response than Cohen's Kappa.

The use cases K2 and K3 show a trend similar to K1: MCC is still –1, but κ equals to –0.22 and –0.471, respectively. Again, MCC suggests perfect wrong prediction, while κ suggests a prediction similar to random guessing. In these two use cases, there are many FN and FP, but true negatives and true positives are zero, so we can conclude that this prediction was totally wrong, and not similar to random guessing. Also in these two cases, we can state that MCC gave a more informative response than Cohen's Kappa.

In the use cases K5 and K6, instead, we can observe concordant values for MCC and κ, both at –1 or close to it. Cohen's Kappa "reaches" MCC, by confirming its message of perfect wrong classification. The absence of true positives and true negatives, also in these cases, suggests that the prediction was wrongly trained to recognize data instances, rather than behave like random guessing.

As previously observed, the largest differences between MCC and Kappa are quite likely to be found when FP and FN are very different, as for instance in the cases K12 and K15 (Table 1). If both MCC and κ are positive, the difference Δ(MCC, κ) is smaller than 0.12 (for example, in the use case K17).

We have MCC = –1.0 if TP = 0 and TN = 0, regardless of the values of FP and FN (for example, the K1 and K6 cases Table 1). But if TP = 0 and TN = 0, Kappa may produce values between 0.0 and –1.0. For example, we have Kappa = 0 if either FP = 0 or FN = 0 as in case K1, and we have Kappa = –1.0 if and only if FP = FN as in case K6.

Finally, consider occurring whenever a low value for Kappa and MCC is matched by an high agreement (accuracy) [53]–[55], as in the use cases K11 and K16: in these cases the low values of MCC and Kappa are welcomed, since the binary classification is far from being perfect. Formal proofs of these properties can be found in a study by Warrens [57].

We can therefore conclude the analysis of these use cases stating that MCC and κ generate similar and concordant positive scores, but they can generate discordant negative scores, on the same confusion matrices. When MCC and Cohen's Kappa generate negative discordant scores, the value produced by MCC is more reliable and informative of the real status of the corresponding confusion matrix.

B. MCC AND BRIER SCORE USE CASES

As mentioned earlier, we took advantage of Beta distributions to produce simulated classifiers to use to generate values of MCC and Brier score.

From all the possible classifiers generated earlier for the scatterplots (Figure 3 and Figure 4), we selected the ones with the highest difference between normMCC and complBS as use cases to analyze here. We reported the parameters and quantitative characteristics of these use use cases in Table 2 and Table 3.

We reported these differences as $\Delta(c, n)$ in Table 4. As one can notice, the Brier score (BS) generate discordant values from MCC for six presented use cases BS1, BS2, BS3, BS4, BS5, and BS6. The Matthews correlation coefficient ranges from –0.843 to –0.73, indicating a poor prediction performance close to a perfectly wrong prediction, where the classifier almost completely confused positives with negatives. On the contrary, the values of the Brier score range from 0.414 to 0.486 interval, indicating quite a slightly good prediction. The perfect value for the Brier score would be

use case	TP	FN	FP	TN	MCC	κ	Δ(MCC, κ)
K1	0	100	0	0	–1.000	0.000	1.000
K2	0	90	10	0	–1.000	–0.220	0.780
K3	0	80	20	0	–1.000	–0.471	0.529
K4	0	70	30	0	–1.000	–0.724	0.276
K5	0	60	40	0	–1.000	–0.923	0.077
K6	0	50	50	0	–1.000	–1.000	0.000
K7	27	45	1	27	+0.339	+0.229	0.110
K8	40	45	1	14	+0.293	+0.183	0.110
K9	20	59	1	20	+0.206	+0.102	0.103
K10	15	69	1	15	+0.116	+0.043	0.073
K11	90	1	9	0	–0.031	–0.018	0.013
K12	5	70	6	19	–0.240	–0.094	0.146
K13	47	3	45	5	+0.074	+0.040	0.034
K14	10	40	4	46	+0.173	+0.120	0.053
K15	9	1	89	1	–0.190	–0.018	0.172
K16	2	9	1	88	+0.313	+0.250	0.063
K17	30	40	0	30	+0.429	+0.310	0.118

TABLE 1: **Use cases for MCC and Cohen's Kappa**. MCC: Matthews correlation coefficient (Equation 1). κ: Cohen's Kappa (Equation 2). MCC and κ have worst value equal to –1 and best value equal to +1. Δ(MCC, κ): absolute difference between MCC and κ. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. Threshold cut-off for predictions: $\tau = 0.5$.

ground truth	BS1 balanced	BS2 negatively imbalanced	BS3 positively imbalanced
positives	$\mathrm{Beta}(9, 15)$	$\mathrm{Beta}(6, 15)$	$\mathrm{Beta}(7, 15)$
negatives	$\mathrm{Beta}(15, 8)$	$\mathrm{Beta}(15, 8)$	$\mathrm{Beta}(15, 7)$
# positives	5,000	9,000	1,000
# negatives	5,000	1,000	9,000
% positives	50%	90%	10%
% negatives	50%	10%	90%

TABLE 2: **Use cases BS1, BS2, and BS3: score distributions used for the three simulated classifiers and summary statistics for the datasets.** We listed the Beta distributions generated for the ground truth positives and negatives, in the three use cases BS1, BS2, and BS3. For example, we associated the real values generated by Beta(9, 15) to the BS1 positive data instances.

ground truth	BS4 balanced	BS5 negatively imbalanced	BS6 positively imbalanced
positives	$\mathrm{Beta}(9, 15)$	$\mathrm{Beta}(7, 15)$	$\mathrm{Beta}(7, 15)$
negatives	$\mathrm{Beta}(9, 15)$ for first 70% $\mathrm{Beta}(12, 7)$ for last 30%	$\mathrm{Beta}(8, 14)$ for first 70% $\mathrm{Beta}(15, 8)$ for last 30%	$\mathrm{Beta}(7, 15)$ for first 70% $\mathrm{Beta}(14, 6)$ for last 30%
# positives	50	90	10
# negatives	50	10	90
% positives	50%	90%	10%
% negatives	50%	10%	90%

TABLE 3: **Use cases BS4, BS5, and BS6: score distributions used for the three simulated classifiers and summary statistics for the datasets.** We listed the Beta distributions generated for the ground truth positives and negatives, in the three use cases BS4, BS5, and BS6. For example, we associated the real values generated by Beta(9, 15) to the BS4 positive data instances.

zero. To highlight these differences, we represent them as barplots in Figure 5.

Another interesting aspect to notice is that the binary Brier score (binaryBS) results are concordant with MCC, having values very close to 1 that indicate poor performance, and in contrast with the original Brier score values.

By taking a closer look to the corresponding confusion matrices (Table 4), we can see that in all the six BS1, ..., BS6 use cases there is a large majority of false positives and false negatives over true positives and true negatives. In BS1, for example, the false negatives are almost 9 times the true positives, while the false positives are 16 times

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

case	TP	FN	FP	TN	binBS	BS	complBS	MCC	normMCC	Δ(c, n)
BS1	511	4,489	4,706	294	0.920	0.419	0.581	–0.840	0.080	0.501
BS2	18	982	8,455	545	0.944	0.442	0.558	–0.769	0.116	0.442
BS3	323	8,677	962	38	0.964	0.476	0.524	–0.830	0.085	0.439
BS4	2	48	44	6	0.920	0.414	0.586	–0.843	0.079	0.507
BS5	1	9	85	5	0.940	0.444	0.556	–0.730	0.135	0.421
BS6	3	87	10	0	0.970	0.486	0.500	–0.862	0.069	0.446
BS7	1	4	4	1	0.800	0.251	0.749	–0.600	0.200	0.549
BS8	4	1	1	4	0.200	0.249	0.751	+0.600	0.800	0.049

TABLE 4: **Use cases for MCC and Brier score**. BS: Brier score (Equation 7). binBS: binaryBS, binary Brier score (Equation 8). MCC: Matthews correlation coefficient (Equation 1). normMCC: normalizedMCC = (MCC + 1) / 2. complBS: complementaryBS = 1 – BS. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. Threshold cut-off for predictions: $\tau = 0.5$. Δ(c, n): absolute difference between complBS and normMCC. We described the details of the simulated datasets and the simulated classifications BS1, B2, B3, B4, B5, and BS6 in Table 2 and Table 3.

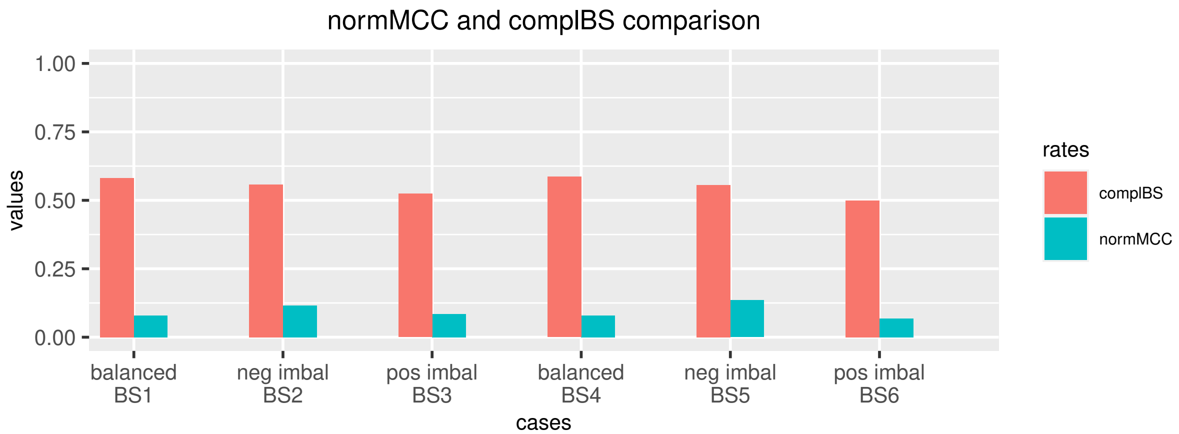

FIGURE 5: **Results of MCC and Brier score for the BS1, BS2, BS3, BS4, BS5, and BS6 use cases**. normMCC = (MCC + 1)/2. complBS = 1 – BS. The values of both normMCC and complBS lay in the [0, 1] interval, with worst value equal to 0 and best value equal to 1. We reported the details of these use cases in Table 4.

the true negatives. In this framework, it is clear that an informative rate would generate a negative response. MCC, in fact, produces a value of –0.84, confirming the poor ratio of positives with respect to negatives. On the contrary, the Brier score has a value of 0.419, which is closer to 0 (perfect prediction) than to 1 (worst prediction). Similar trends can be observed in the other use cases (BS2, BS3, BS4, BS5, and BS6).

We can therefore state that the Matthews correlation coefficient produces a more capable and informative outcome than the Brier score.

At this point, someone could rebut this statement by stating that the confusion matrix categories are not included in the Brier score computation, and therefore might be improper to use them here in this comparison. Even if we know that the Brier score does not produce and is not produced by two-class confusion matrices with a strict cut-off threshold, we believe that it is necessary to consider them for binary classification, because a clear distinction between positives and negatives is fundamental for experiment validation. In a clinical setting, for example, rates based on two-class confusion matrix scores must be employed when a clear distinction between healthy controls (negatives) and patients with disease (positives) need to be made.

The BS ≈ 0.25 ambiguity. As mentioned earlier (subsubsection III-B1), a strong discordance between the Brier score and MCC can happen when the Brier score has values around 0.25. This situation can happen especially when the classifier predicts values around the cut-off threshold for the confusion matrix, that traditionally is set to 0.5 in machine learning and statistics.

Let us consider now the use case BS7 with a dataset with 10 elements, having the following binary ground truth values:

```
ground truth values: (0, 0, 0, 0, 0, 1, 1, 1,
1, 1)
```

This dataset is perfectly balanced, with 5 negatives and 5 positives. And let us suppose that a classifier predicts the following values for them:

```
BS7 predictions: (0.501, 0.501, 0.501, 0.499, 0.501,
0.499, 0.501, 0.499, 0.499, 0.499)
```

This classifier would get Brier score = 0.251, meaning good outcome, and MCC = –0.6, meaning very bad performance (Table 4).

And let us consider now the use case BS8, with the same ground truth dataset of BS7, but with the following predictions:

```
BS8 predictions: (0.499, 0.499, 0.501, 0.499, 0.499,
0.499, 0.501, 0.501, 0.501, 0.501)
```

Regarding this performance, the value of the Brier score would be 0.249, meaning good prediction, and the coefficient of the Matthews correlation would be +0.6, meaning good prediction too (Table 4).

As one can notice and as we described earlier (subsubsection III-B1), a Brier score close to 0.25 has an ambiguous meaning: it could be associated to a prediction evaluated as poor like in the BS7 use case, or it could be associated to a prediction evaluated as good like in the BS8 use case.

V. CONCLUSIONS

Assessing binary evaluations is a key task in machine learning and computational statistics. The Matthews correlation coefficient (MCC), Cohen's Kappa, and the Brier score are three common rates employed to evaluate the predictions made by the classifier in relation to the corresponding dataset ground truth.

In our study, we showed that MCC is more informative, truthful, and reliable than Cohen's Kappa and the Brier score to this end. Cohen's Kappa, in fact, can provide misleading information in some particular cases, especially when true positives and true negatives are zero. On the other side, the Brier score can generate an ambiguous outcome when its value is close to 0.25, which can correspond both to a very good prediction and to a very bad prediction. The Matthews correlation coefficient, instead, does not have these flaws.

Although generally MCC is more informative than κ statistic and the Brier score, there are some cases where these rates are equally reliable. When the classifier is better than random (MCC and $\kappa > 0$) the correlation between the two metrics is very high; the difference when using MCC or κ is negligible (Figure 1). When the classifier is worse than random, the situation is quite symmetric. Given a specific MCC value, there is a wide range of different κ values that can be used to discriminate (Figure 1), and the same happens oppositely: for a given κ value, there are many MCC values (Figure 1). Thus, in this situation, using MCC or κ provides the same level of reliability.

Instead, the correlation between MCC and the Brier score is quite limited, so choosing one of the two heavily depends on their properties (Figure 3 and Figure 4). In fact, to a given value of MCC corresponds a quite broad range of BS values, and vice versa, thus there is no specific situation where MCC should not be preferred to BS. However, BS can be useful in discriminating situations sharing the same MCC. For instance, consider the use case with ground truth: `(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)`.

When the predicted values are `(0.499, 0.499, 0.501, 0.499, 0.499, 0.499, 0.501, 0.501, 0.501, 0.501)`, we have MCC = +0.6 and BS = 0.249.

If instead the predictions are `(0.001, 0.001, 0.501, 0.001, 0.001, 0.499, 0.999, 0.999, 0.999, 0.999)`, we obtain MCC = +0.6 again, but BS = 0.05, highlighting a different prediction with respect to the previous case. If a machine learning practitioner had to select a predictive algorithm by observing the predictions in the two cases, she/he could choose the first one, because it generated a higher Brier score than the second one.

Our results and statements about Cohen's Kappa confirm what was claimed by Delgado and Tibau [38] in their study: these authors showed that if marginal probabilities are really small, the distribution of a misclassification also affects κ. This way, worse classification results can achieve higher values of this score, which would therefore provide a misleading outcome. The authors claim that these drawbacks of Cohen's Kappa can be especially dramatic in clinical perspective, and we agree with them.

Our results and considerations regarding the Brier score are in line with what was highlighted by Assel and colleagues [97], who stated that the Brier score is unsuitable in clinical tests evaluation because it provides counter-intuitive results in several situations. As a major example, the Brier score will favor a test with high specificity if it is the case that prevalence is low even when the clinical context requires high sensitivity. Furthermore, the Brier score favours continuous models over binary tests even if the test is proven to be more effective. This is due to the fact that the Brier score measures the quality of prediction independently of the clinical scenario, thus issuing a caveat for its application [97].

For the reasons described in our article, we therefore suggest any machine learning practitioner to use the Matthews correlation coefficient rather than Cohen's Kappa or the Brier score to assess binary classification experiments.

In the future, we plan to make additional comparative analyses between the Matthews correlation coefficient and other rates, such as the Fowlkes-Mallows index [101], the prevalence threshold [102], and the Jaccard index [103], [104].

LIST OF ABBREVIATIONS

AUC: area under the curve. binaryBS: binary Brier score. BS: Brier score. complBS: complementary Brier score. DOR: diagnostic odds ratio. FDA: USA Food and Drug Administration (FDA) agency. FN: false negatives. FP: false positives. κ: Cohen's Kappa. MAQC/SEQC: MicroArray / Sequencing Quality Control. MCC: Matthews correlation coefficient. normMCC: normalized Matthews correlation coefficient. PR:

precision-recall. ROC: receiver operating characteristic. TN: true negatives. TP: true positives.

ACKNOWLEDGMENTS

The authors thank Christopher Ferro (University of Exeter) for his suggestions.

COMPETING INTERESTS

The authors declare they have no competing interest.

SOFTWARE AVAILABILITY

Our software code is publicly available at:
https://github.com/davidechicco/MCC_versus_BrierScore_and_CohensKappa

REFERENCES

[1] 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.

[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 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.

[4] 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.

[5] 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.

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

[7] Kaggle. Featured prediction competition: VSB power line fault detection. https://www.kaggle.com/c/vsb-power-line-fault-detection/overview/evaluation URL visited on 24th September, 2019.

[8] DataDriven.org. Clog loss: advance Alzheimer's research with stall catchers. https://www.drivendata.org/competitions/65/clog-loss-alzheimers-research/page/217/ URL visited on 9th October, 2020.

[9] Thomas Cokelaer, Mukesh Bansal, Christopher Bare, Erhan Bilal, Brian M Bot, Elias Chaibub Neto, Federica Eduati, Alberto de la Fuente, Mehmet Gönen, Steven M Hill, Bruce Hoff, Jonathan R Karr, Robert Küffner, Michael P Menden, Pablo Meyer, Raquel Norel, Abhishek Pratap, Robert J Prill, Matthew T Weirauch, James C Costello, Gustavo Stolovitzky, and Julio Saez-Rodriguez. DREAMTools: a Python package for scoring collaborative challenges. F1000Research, 4, 2015.

[10] Sage Bionetworks. DREAM Challenges. https://www.dreamchallenges.org/ URL visited on 24th September, 2020.

[11] MAQC Consortium. The MAQC-II project: a comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 28(8):827–838, 2010.

[12] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nature Biotechnology, 32(9):903–914, 2014.

[13] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS ONE, 12(6):e0177678, 2017.

[14] Qiuming Zhu. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognition Letters, 136:71–88, 2020.

[15] Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. A critical analysis of metrics used for measuring progress in artificial intelligence. arXiv preprint, 2008.02577, 2020.

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

[17] Boaz Shmueli. Matthews Correlation Coefficient is the best classification metric you've never heard of. https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a URL visited on 24th September, 2020.

[18] David Lettier. You need to know about the Matthews Correlation Coefficient. https://lettier.github.io/posts/2016-08-05-matthews-correlation-coefficient.html URL visited on 24th September, 2020.

[19] Harald Cramér. Mathematical Methods of Statistics, volume 43. Princeton University Press, Princeton, New Jersey, USA, 1999.

[20] Theresa Marchant-Shapiro. Chi-square and Cramer's V: what do you expect. Statistics for Political Analysis: Understanding the Numbers, pages 245–272, 2015.

[21] Bo Wu, Liangpei Zhang, and Yindi Zhao. Feature selection via Cramer's V-test discretization for remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 52(5):2593–2606, 2013.

[22] Svante Janson and Jan Vegelius. On generalizations of the G index and the phi coefficient to nominal scales. Multivariate Behavioral Research, 14(2):255–269, 1979.

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

[24] Peter V Zysno. The modification of the phi-coefficient reducing its dependence on the marginal distributions. Methods of Psychological Research Online, 2(1):41–52, 1997.

[25] Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy, 85(3):257–268, 2005.

[26] Shuyan Sun. Meta-analysis of Cohen's kappa. Health Services and Outcomes Research Methodology, 11(3-4):145–163, 2011.

[27] Matthijs J Warrens. Cohen's kappa can always be increased and decreased by combining categories. Statistical Methodology, 7(6):673–677, 2010.

[28] Matthijs J Warrens. Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. Statistical Methodology, 8(2):268–272, 2011.

[29] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.

[30] Matthijs J Warrens. New interpretations of Cohen's kappa. Journal of Mathematics, 2014, 2014.

[31] Matthijs J Warrens. Five ways to look at Cohen's kappa. Journal of Psychology & Psychotherapy, 5(4):1, 2015.

[32] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: a review of interrater agreement measures. Canadian Journal of Statistics, 27(1):3–23, 1999.

[33] Matthijs J Warrens. Cohen's kappa is a weighted average. Statistical Methodology, 8(6):473–484, 2011.

[34] Matthijs J Warrens. Conditional inequalities between Cohen's kappa and weighted kappas. Statistical Methodology, 10(1):14–22, 2013.

[35] Matthijs J Warrens. A comparison of Cohen's kappa and agreement coefficients by Corrado Gini. International Journal of Research and Reviews in Applied Sciences, 16:345–351, 2013.

[36] Frank Krummenauer, Peter Kalden, and Karl-Friedrich Kreitner. Cohen's kappa or McNemar's test? A comparison of binary repeated measurements. Rofo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin, 171(3):226–231, 1999.

[37] Matthijs J Warrens. Bounds of resemblance measures for binary (presence/absence) variables. Journal of Classification, 25(2):195–208, 2008.

[38] Rosario Delgado and Xavier-Andoni Tibau. Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE, 14(9):e0222916, 2019.

[39] Douglas Steinley. Properties of the Hubert-Arable Adjusted Rand Index. Psychological Methods, 9(3):386, 2004.

[40] Matthijs J Warrens. On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. Journal of Classification, 25(2):177–183, 2008.

[41] Arie Ben-David. About the relationship between ROC curves and Cohen's kappa. Engineering Applications of Artificial Intelligence, 21(6):874–882, 2008.

[42] Helena C Kraemer. Kappa coefficient. Wiley StatsRef: Statistics Reference Online, pages 1–4, 2014.

[43] Jingyun Yang and Vernon M Chinchilli. Fixed-effects modeling of Cohen's Kappa for bivariate multinomial data. Communications in Statistics—Theory and Methods, 38(20):3634–3653, 2009.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3084050, IEEE Access

[44] Marcia Feingold. The equivalence of Cohen's Kappa and Pearson's chi-square statistics in the 2 × 2 table. Educational and Psychological Measurement, 52(1):57–61, 1992.

[45] Alan B Cantor. Sample-size calculations for Cohen's kappa. Psychological Methods, 1(2):150, 1996.

[46] J Barry Garner. The standard error of Cohen's kappa. Statistics in Medicine, 10(5):767–775, 1991.

[47] Guogen Shan and Weizhen Wang. Exact one-sided confidence limits for Cohen's kappa as a measurement of agreement. Statistical Methods in Medical Research, 26(2):615–632, 2017.

[48] Bikas K Sinha, Pornpis Yimprayoon, and Montip Tiensuwan. Cohen's kappa statistic: a critical appraisal and some modifications. Calcutta Statistical Association Bulletin, 58(3-4):151–170, 2006.

[49] Volker W Steinijans, Edgar Diletti, B Bömches, Christian Greis, and Peter Solleder. Interobserver agreement: Cohen's kappa coefficient does not necessarily reflect the percentage of patients with congruent classifications. International Journal of Clinical Pharmacology and Therapeutics, 35(3):93–95, 1997.

[50] Werner Vach. The dependence of Cohen's kappa on the prevalence does not matter. Journal of Clinical Epidemiology, 58(7):655–661, 2005.

[51] Alexander von Eye and Maxine Von Eye. Can one use Cohen's kappa to examine disagreement? Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1(4):129, 2005.

[52] Shu Xu and Michael F Lorber. Interrater agreement statistics with skewed data: evaluation of alternatives to Cohen's kappa. Journal of Consulting and Clinical Psychology, 82(6):1219, 2014.

[53] Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. The problem of two paradoxes. Journal of Clinical Epidemiology, 43(6):543–549, 1990.

[54] Domenic V Cicchetti and Alvan R Feinstein. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology, 43(6):551–558, 1990.

[55] Ted Byrt, Janet Bishop, and John B Carlin. Bias, prevalence and kappa. Journal of Clinical Epidemiology, 46(5):423–429, 1993.

[56] Matthijs J Warrens. A formal proof of a paradox associated with Cohen's kappa. Journal of Classification, 27(3):322–332, 2010.

[57] Matthijs J Warrens. On marginal dependencies of the 2×2 Kappa. Advances in Statistics, 2014:1–6, 2014.

[58] Mikel Aickin. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics, pages 293–302, 1990.

[59] Domenic V Cicchetti, Ami Klin, and Fred R Volkmar. Assessing binary diagnoses of bio-behavioral disorders: the clinical relevance of Cohen's Kappa. Journal of Nervous and Mental Disease, 205(1):58–65, 2017.

[60] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4):213, 1968.

[61] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33(3):613–619, 1973.

[62] Tarald O Kvålseth. Note on Cohen's kappa. Psychological Reports, 65(1):223–226, 1989.

[63] Markus Wirtz and Marcus Kutschmann. Analyzing interrater agreement for categorical data using Cohen's kappa and alternative coefficients. Die Rehabilitation, 46(6):370–377, 2007.

[64] Kenneth J Berry and Paul W Mielke Jr. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educational and Psychological Measurement, 48(4):921–933, 1988.

[65] Patricia Simon. Including omission mistakes in the calculation of Cohen's Kappa and an analysis of the coefficient's paradox features. Educational and Psychological Measurement, 66(5):765–777, 2006.

[66] Chia-Hao Chang. Cohen's kappa for capturing discrimination. International Health, 6(2):125–129, 2014.

[67] Henning Holle and Robert Rein. The modified Cohen's kappa: calculating interrater agreement for segmentation and annotation. Understanding Body Movement, pages 261–277, 2013.

[68] Susana M Vieira, Uzay Kaymak, and João M Sousa. Cohen's kappa coefficient as a performance measure for feature selection. In Proceedings of FUZZ-IEEE 2010 – the 7th International Conference on Fuzzy Systems, pages 1–8. IEEE, 2010.

[69] Alfred Stein, Jagannath Aryal, and Gerrit Gort. Use of the Bradley-Terry model to quantify association in remotely sensed images. IEEE Transactions on Geoscience and Remote Sensing, 43(4):852–856, 2005.

[70] Robert G Pontius and Marco Millones. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32(15):4407–4429, 2011.

[71] Matthijs J Warrens. Properties of the quantity disagreement and the allocation disagreement. International Journal of Remote Sensing, 36(5):1439–1446, 2015.

[72] Philip E Tetlock and Dan Gardner. Superforecasting: the art and science of prediction. Random House, New York City, New York, USA, 2016.

[73] José Hernández-Orallo, Peter A Flach, and Cèsar Ferri Ramirez. A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal of Machine Learning Research, 13:2813–2869, 2012.

[74] Glenn W Brier. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1):1–3, 1950.

[75] Gail Blattenberger and Frank Lad. Separating the Brier score into calibration and refinement components: a graphical exposition. American Statistician, 39(1):26–32, 1985.

[76] Allan H Murphy. A new decomposition of the Brier score: formulation and interpretation. Monthly Weather Review, 114(12):2671–2673, 1986.

[77] Mitsuru Ikeda, Takeo Ishigaki, and Kazunobu Yamauchi. Relationship between Brier score and area under the binormal ROC curve. Computer Methods and Programs in Biomedicine, 67(3):187–194, 2002.

[78] Stephen Jewson. The problem with the Brier score. arXiv preprint, physics/0401046, 2004.

[79] Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biometrical Journal, 48(6):1029–1040, 2006.

[80] Barbara Casati and Lori J Wilson. A new spatial-scale decomposition of the Brier score: application to the verification of lightning probability forecasts. Monthly Weather Review, 135(9):3052–3069, 2007.

[81] Mark S Roulston. Performance targets and the Brier score. Meteorological Applications, 14(2):185–194, 2007.

[82] David B Stephenson, Caio A Coelho, and Ian T Jolliffe. Two extra components in the Brier score decomposition. Weather and Forecasting, 23(4):752–757, 2008.

[83] Christopher A Ferro. Comparing probabilistic forecasting systems with the Brier score. Weather and Forecasting, 22(5):1076–1088, 2007.

[84] A Allen Bradley, Stuart S Schwartz, and Tempei Hashino. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. Weather and Forecasting, 23(5):992–1006, 2008.

[85] Daniel S Wilks. Sampling distributions of the Brier score and Brier skill score under serial dependence. Quarterly Journal of the Royal Meteorological Society, 136(653):2109–2118, 2010.

[86] Kaspar Rufibach. Use of Brier score to assess binary predictions. Journal of Clinical Epidemiology, 63(8):938–939, 2010.

[87] David J Spiegelhalter. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine, 5(5):421–433, 1986.

[88] Michael Jachan, Hinnerk Feldwisch genannt Drentrup, Florian Posdziech, Armin Brandt, Dirk-Matthias Altenmüller, Andreas Schulze-Bonhage, Jens Timmer, and Björn Schelter. Probabilistic forecasts of epileptic seizures and evaluation by the Brier score. In Proceedings of ECIFMBE 2008 – the 4th European Conference of the International Federation for Medical and Biological Engineering, pages 1701–1705. Springer, 2009.

[89] Ulf Johansson, Rikard König, and Lars Niklasson. Genetic rule extraction optimizing Brier score. In Proceedings of GECCO 2010 – the 12th Annual Conference on Genetic and Evolutionary Computation, pages 1007–1014, 2010.

[90] University of California Irvine. Machine Learning Repository. https://archive.ics.uci.edu/ml/ URL visited on 28th September, 2021.

[91] Roland M Young. Decomposition of the Brier score for weighted forecast-verification pairs. Quarterly Journal of the Royal Meteorological Society, 136(650):1364–1370, 2010.

[92] Christopher A Ferro and Thomas E Fricker. A bias-corrected decomposition of the Brier score. Quarterly Journal of the Royal Meteorological Society, 138(668):1954–1960, 2012.

[93] Stefan Siegert. Variance estimation for Brier score decomposition. Quarterly Journal of the Royal Meteorological Society, 140(682):1771–1777, 2014.

[94] Edgar C Merkle. Weighted Brier score decompositions for topically heterogenous forecasting tournaments. Judgment and Decision Making, 13(2):185–201, 2018.

[95] José Hernández-Orallo, Peter A Flach, and Cèsar Ferri Ramirez. Brier curves: a new cost-based visualisation of classifier performance. In

Proceedings of ICML 2011 – the 28th International Conference on Machine Learning, pages 585–592, 2011.

[96] Sally A Lesik and Meg Leake. Using a Brier score analysis to assess the effectiveness of a mathematics placement policy. Journal of College Student Retention: Research, Theory & Practice, 14(2):209–225, 2012.

[97] Melissa Assel, Daniel D Sjoberg, and Andrew J Vickers. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and Prognostic Research, 1(1):1–7, 2017.

[98] Chang Cao, Davide Chicco, and Michael M Hoffman. The MCC-F1 curve: a performance evaluation technique for binary classification. arXiv preprint, 2006.11278, 2020.

[99] Simaan M AbouRizk, Daniel W Halpin, and James R Wilson. Fitting beta distributions based on sample data. Journal of Construction Engineering and Management, 120(2):288–305, 1994.

[100] 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):1–11, 2020.

[101] 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.

[102] Jacques Balayla. Prevalence threshold (ϕe) and the geometry of screening curves. PLoS ONE, 15(10):e0240215, 2020.

[103] Abdel A Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging, 15(1):1–28, 2015.

[104] Eduardo Fernandez-Moral, Renato Martins, Denis Wolf, and Patrick Rives. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. In Proceedings of IV'18 – the 2018 IEEE Intelligent Vehicles Symposium, pages 1051–1056. IEEE, 2018.

DAVIDE CHICCO (ORCID: 0000-0001-9655-7142) obtained his Bachelor of Science and Master of Science degrees in computer science at Università di Genova (Genoa, Italy) respectively in 2007 and 2010. He then started the PhD program in computer engineering at Politecnico di Milano university (Milan, Italy), where he graduated in spring 2014. He also spent a semester as visiting doctoral scholar at University of California Irvine (USA). From September 2014 to September 2018, he has been a post-doctoral researcher at the Princess Margaret Cancer Centre and a guest at University of Toronto. From September 2018 to December 2019, he was a scientific associate researcher at the Peter Munk Cardiac Centre (Toronto, Ontario, Canada). From January 2020 to January 2021, he has been a scientific associate researcher at the Krembil Research Institute (Toronto, Ontario, Canada). Since January 2021, he started to work as a scientific research associate at the Institute of Health Policy, Management and Evaluation of University of Toronto.

MATTHIJS J. WARRENS (ORCID: 0000-0002-7302-640X) studied mathematics and psychology at Leiden University (the Netherlands) and received his PhD Degree in 2008 at the same university. From 2017 onwards, he has been working as an associate professor at the GION Education/Research institute of the University of Groningen (the Netherlands). His scientific research interests include various statistical topics (reliability indices, inter-observer agreement, clustering methods, longitudinal latent variable modelling) as well as educational trajectories and the application of machine learning methods to gain insight into complex educational data. From 2011 to 2014, Matthijs J. Warrens worked on a VENI project on Kappa coefficients for measuring inter-observer agreement. He is an associate editor of the SSCI Journal of Classification, a board member of the Dutch/Flemish Classification Society, and a member of the Cluster Benchmarking Task Force of the International Federation of Classification Societies.

GIUSEPPE JURMAN (ORCID: 0000-0002-2705-5728) earned his PhD in algebra at the Università di Trento (Italy) in 1998. After two years as postdoctoral fellow at the Australian National University (ANU) Canberra, in 2002 he moved to the Fondazione Bruno Kessler (FBK) in Trento where he is now Senior Researcher in Data Science, working mainly on computational biology. His main research interests are 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 at the Università di Trento, and since 2008 he is co-director of WebValley, the FBK summer school for dissemination of interdisciplinary research for high school students.