

RESEARCH ARTICLE

Open Access



The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation

Davide Chicco^{1,2*} and Giuseppe Jurman³

Abstract

Background: To evaluate binary classifications and their confusion matrices, scientific researchers can employ several statistical rates, accordingly to the goal of the experiment they are investigating. Despite being a crucial issue in machine learning, no widespread consensus has been reached on a unified elective chosen measure yet. Accuracy and F₁ score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification tasks. However, these statistical measures can dangerously show overoptimistic inflated results, especially on imbalanced datasets.

Results: The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

Conclusions: In this article, we show how MCC produces a more informative and truthful score in evaluating binary classifications than accuracy and F₁ score, by first explaining the mathematical properties, and then the asset of MCC in six synthetic use cases and in a real genomics scenario. We believe that the Matthews correlation coefficient should be preferred to accuracy and F₁ score in evaluating binary classification tasks by all scientific communities.

Keywords: Matthews correlation coefficient, Binary classification, F₁ score, Confusion matrices, Machine learning, Biostatistics, Accuracy, Dataset imbalance, Genomics

Background

Given a clinical feature dataset of patients with cancer traits [1, 2], which patients will develop the tumor, and which will not? Considering the gene expression of neuroblastoma patients [3], can we identify which patients are going to survive, and which will not? Evaluating the metagenomic profiles of patients [4], is it possible to discriminate different phenotypes of a complex disease? Answering these questions is the aim of machine learning and computational statistics, nowadays pervasive in analysis of biological and health care datasets, and

many other scientific fields. In particular, these binary classification tasks can be efficiently addressed by supervised machine learning techniques, such as artificial neural networks [5], *k*-nearest neighbors [6], support vector machines [7], random forest [8], gradient boosting [9], or other methods. Here the word *binary* means that the data element statuses and prediction outcomes (class labels) can be twofold: in the example of patients, it can mean healthy/sick, or low/high grade tumor. Usually scientists indicate the two classes as the negative and the positive class. The term *classification* means that the goal of the process is to attribute the correct label to each data instance (sample); the process itself is known as the classifier, or classification algorithm.

*Correspondence: davidechicco@davidechicco.it

¹Krembil Research Institute, Toronto, Ontario, Canada

²Peter Munk Cardiac Centre, Toronto, Ontario, Canada

Full list of author information is available at the end of the article



Scientists have used binary classification to address several questions in genomics in the past, too. Typical cases include the application of machine learning methods to microarray gene expressions [10] or to single-nucleotide polymorphisms (SNPs) [11] to classify particular conditions of patients. Binary classification can also be used to infer knowledge about biology: for example, computational intelligence applications to ChIP-seq can predict transcription factors [12], applications to epigenomics data can predict enhancer-promoter interactions [13], and applications to microRNA can predict genomic inverted repeats (pseudo-hairpins) [14].

A crucial issue naturally arises, concerning the outcome of a classification process: how to evaluate the classifier performance? A relevant corpus of published works has stemmed until today throughout the last decades for possible alternative answers to this inquiry, by either proposing a novel measure or comparing a subset of existing ones on a suite of benchmark tasks to highlight pros and cons [15–28], also providing off-the-shelf software packages [29, 30]. Despite the amount of literature dealing with this problem, this question is still an open issue. However, there are several consolidated and well known facts driving the choice of evaluating measures in the current practice.

Accuracy, MCC, F_1 score. Many researchers think the most reasonable performance metric is the ratio between the number of correctly classified samples and the overall number of samples (for example, [31]). This measure is called *accuracy* and, by definition, it also works when labels are more than two (multiclass case). However, when the dataset is unbalanced (the number of samples in one class is much larger than the number of samples in the other classes), accuracy cannot be considered a reliable measure anymore, because it provides an overoptimistic estimation of the classifier ability on the majority class [32–35].

An effective solution overcoming the class imbalance issue comes from the *Matthews correlation coefficient* (MCC), a special case of the ϕ phi coefficient [36].

Stemming from the definition of the phi coefficient, a number of metrics have been defined and mainly used for purposes other than classification, for instance as association measures between (discrete) variables, with the *Cramér's V* (or *Cramér's ϕ*) being one of the most common rates [37].

Originally developed by Matthews in 1975 for comparison of chemical structures [38], MCC was re-proposed by Baldi and colleagues [39] in 2000 as a standard performance metric for machine learning with a natural extension to the multiclass case [40]. MCC soon started imposing as a successful indicator: for instance, the Food and Drug Administration (FDA) agency of the USA employed the MCC as the main evaluation

measure in the MicroArray II / Sequencing Quality Control (MAQC/SEQC) projects [41, 42]. The effectiveness of MCC has been shown in other scientific fields as well [43, 44].

Although being widely acknowledged as a reliable metric, there are situations - albeit extreme - where either MCC cannot be defined or it displays large fluctuations [45], due to imbalanced outcomes in the classification. Even if mathematical workarounds and Bayes-based improvements [46] are available for these cases, they have not been adopted widely yet.

Shifting context from machine learning to information retrieval, and thus interpreting positive and negative class as relevant and irrelevant samples respectively, the recall (that is the accuracy on the positive class) can be seen as the fraction of relevant samples that are correctly retrieved. Then its dual metric, the precision, can be defined as the fraction of retrieved documents that are relevant. In the learning setup, the pair precision/recall provides useful insights on the classifier's behaviour [47], and can be more informative than the pair specificity/sensitivity [48]. Meaningfully combining precision and recall generates alternative performance evaluation measures. In particular, their harmonic mean has been originally introduced in statistical ecology by Dice [49] and Sørensen [50] independently in 1948, then rediscovered in the 1970s in information theory by van Rijsbergen [51, 52] and finally adopting the current notation of F_1 measure in 1992 [53]. In the 1990s, in fact, F_1 gained popularity in the machine learning community, to the point that it was also re-introduced later in the literature as a novel measure [54].

Nowadays, the F_1 measure is widely used in most application areas of machine learning, not only in the binary scenario, but also in multiclass cases. In multiclass cases, researchers can employ the F_1 micro/macro averaging procedure [55–60], which can be even targeted for *ad-hoc* optimization [61].

The distinctive features of F_1 score have been discussed in the literature [62–64]. Two main properties characterize F_1 from MCC. First, F_1 varies for class swapping, while MCC is invariant if the positive class is renamed negative and vice versa. This issue can be overcome by extending the macro/micro averaging procedure to the binary case itself [17], by defining the F_1 score both on the positive and negative classes and then average the two values (macro), and using the average sensitivity and average precision values (micro). The micro/macro averaged F_1 is invariant for class swapping and its behaviour is more similar to MCC. However, this procedure is biased [65], and it is still far from being accepted as a standard practice by the community. Second, F_1 score is independent from the number of samples correctly classified as negative. Recently, several scientists highlighted drawbacks of the

F_1 measure [66, 67]: in fact, Hand and Peter [68] claim that alternative measures should be used instead, due to its major conceptual flaws. Despite the criticism, F_1 remains one of the most widespread metrics among researchers. For example, when Whalen and colleagues released *TargetFinder*, a tool to predict enhancer-promoters interactions in genomics, they showed its results measured only by F_1 score [13], making it impossible to detect the actual true positive rate and true negative rate of their tests [69].

Alternative metrics. The current most popular and widespread metrics include *Cohen's kappa* [70–72]: originally developed to test inter-rater reliability, in the last decades Cohen's kappa entered the machine learning community for comparing classifiers' performances. Despite its popularity, in the learning context there are a number of issues causing the kappa measure to produce unreliable results (for instance, its high sensitivity to the distribution of the marginal totals [73–75]), stimulating research for more reliable alternatives [76]. Due to these issues, we chose not to include Cohen's kappa in the present comparison study.

In the 2010s, several alternative novel measures have been proposed, either to tackle a particular issue such as imbalance [34, 77], or with a broader purpose. Among them, we mention the *confusion entropy* [78, 79], a statistical score comparable with MCC [80], and the *K measure* [81], a theoretically grounded measure that relies on a strong axiomatic base.

In the same period, Powers proposed *informedness* and *markedness* to evaluate binary classification confusion matrices [22]. Powers defines informedness as *true positive rate* – *true negative rate*, to express how the predictor is informed in relation to the opposite condition [22]. And Powers defines markedness as *precision* – *negative predictive value*, meaning the probability that the predictor correctly marks a specific condition [22].

Other previously introduced rates for confusion matrix evaluations are *macro average arithmetic* (MAvA) [18], *geometric mean* (Gmean or G-mean) [82], and *balanced accuracy* [83], which all represent classwise weighted accuracy rates.

Notwithstanding their effectiveness, all the aforementioned measures have not yet achieved such a diffusion level in the literature to be considered solid alternatives to MCC and F_1 score. Regarding MCC and F_1 , in fact, Dubey and Tatar [84] state that these two measure “provide more realistic estimates of real-world model performance”.

However, there are many instances where MCC and F_1 score disagree, making it difficult for researchers to draw correct deductions on the behaviour of the investigated classifier.

MCC, F_1 score, and accuracy can be computed when a specific statistical threshold τ for the confusion matrix is set. When the confusion matrix threshold is not unique,

researchers can instead take advantage of classwise rates: *true positive rate* (or *sensitivity*, or *recall*) and *true negative rate* (or *specificity*), for example, computed for all the possible confusion matrix thresholds. Different combinations of these two metrics give rise to alternative measures: among them, the area under the *receiver operating characteristic curve* (AUROC or ROC AUC) [85–91] plays a major role, being a popular performance measure when a singular threshold for the confusion matrix is unavailable. However, ROC AUC presents several flaws [92], and it is sensitive to class imbalance [93]. Hand and colleagues proposed improvements to address these issues [94], that were partially rebutted by Ferri and colleagues [95] some years later.

Similar to ROC curve, the *precision-recall (PR) curve* can be used to test all the possible positive predictive values and sensitivities obtained through a binary classification [96]. Even if less common than the ROC curve, several scientists consider the PR curve more informative than the ROC curve, especially on imbalanced biological and medical datasets [48, 97, 98].

If no confusion matrix threshold is applicable, we suggest the readers to evaluate their binary evaluations by checking both the PR AUC and the ROC AUC, focusing on the former [48, 97]. If a confusion matrix threshold is at disposal, instead, we recommend the usage of the Matthews correlation coefficient over F_1 score, and accuracy.

In this manuscript, we outline the advantages of the Matthews correlation coefficient by first describing its mathematical foundations and its competitors accuracy and F_1 score (“[Notation and mathematical foundations](#)” section), and by exploring their relationships afterwards (Relationships between rates). We decided to focus on accuracy and F_1 score because they are the most common metrics used for binary classification in machine learning. We then show some examples to illustrate why the MCC is more robust and reliable than F_1 score, on six synthetic scenarios (“[Use cases](#)” section) and a real genomics application (“[Genomics scenario: colon cancer gene expression](#)” section). Finally, we conclude the manuscript with some take-home messages (“[Conclusions](#)” section).

Methods

Notation and mathematical foundations

Setup. The framework where we set our investigation is a machine learning task requiring the solution of binary classification problem. The dataset describing the task is composed by n^+ examples in one class, labeled *positive*, and n^- examples in the other class, called *negative*. For instance, in a biomedical case control study, the healthy individuals are usually labelled negative, while the positive label is usually attributed to the sick patients. As a

general practice, given two phenotypes, the positive class corresponds to the *abnormal* phenotype. This ranking is meaningful for example, in different stages of a tumor.

The classification model forecasts the class of each data instance, attributing to each sample its predicted label (positive or negative): thus, at the end of the classification procedure, every sample falls in one of the following four cases:

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

This partition can be presented in a 2×2 table called *confusion matrix* $\mathbf{M} = \begin{pmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{pmatrix}$ (expanded in Table 1), which completely describes the outcome of the classification task.

Clearly $\text{TP} + \text{FN} = n^+$ and $\text{TN} + \text{FP} = n^-$. When one performs a machine learning binary classification, she/he hopes to see a high number of true positives (TP) and true negatives (TN), and less false negatives (FN) and false positives (FP). When $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ 0 & n^- \end{pmatrix}$ the classification is perfect.

Since analyzing all the four categories of the confusion matrix separately would be time-consuming, statisticians introduced some useful statistical rates able to immediately describe the quality of a prediction [22], aimed at conveying into a single figure the structure of \mathbf{M} . A set of these functions act classwise (either actual or predicted), that is, they involve only the two entries of \mathbf{M} belonging to the same row or column (Table 2). We cannot consider such measures fully informative because they use only two categories of the confusion matrix [39].

Accuracy. Moving to global metrics having three or more entries of \mathbf{M} as input, many researchers consider computing the accuracy as the standard way to go. Accuracy, in fact, represents the ratio between the correctly predicted instances and all the instances in the dataset:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{n^+ + n^-} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Table 1 The standard confusion matrix \mathbf{M}

	Predicted positive	Predicted negative
Actual positive	True positives TP	False negatives FN
Actual negative	False positives FP	True negatives TN

True positives (TP) and true negatives (TN) are the correct predictions, while false negatives (FN) and false positives (FP) are the incorrect predictions

Table 2 Classwise performance measures

Sensitivity, recall, true positive rate	$= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{n^+}$	Specificity, true negative rate	$= \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{n^-}$
Positive predictive value, precision	$= \frac{\text{TP}}{\text{TP} + \text{FP}}$	Negative predictive value	$= \frac{\text{TN}}{\text{TN} + \text{FN}}$
False positive rate, fallout	$= \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{n^-}$	False discovery rate	$= \frac{\text{FP}}{\text{FP} + \text{TP}}$

TP: true positives. TN: true negatives. FP: false positives. FN: false negatives

(worst value: 0; best value: 1)

By definition, the accuracy is defined for every confusion matrix \mathbf{M} and ranges in the real unit interval $[0, 1]$; the best value 1.00 corresponds to perfect classification

$\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ 0 & n^- \end{pmatrix}$ and the worst value 0.00 corresponds to perfect misclassification $\mathbf{M} = \begin{pmatrix} 0 & n^+ \\ n^- & 0 \end{pmatrix}$.

As anticipated (Background), accuracy fails in providing a fair estimate of the classifier performance in the class-unbalanced datasets. For any dataset, the proportion of samples belonging to the largest class is called the *no-information error rate* $ni = \frac{\max(n^+, n^-)}{n^+ + n^-}$; a binary dataset is (perfectly) balanced if the two classes have the same size, that is, $ni = \frac{1}{2}$, and it is unbalanced if one class is much larger than the other, that is $ni \gg \frac{1}{2}$. Suppose now that $ni \neq \frac{1}{2}$, and apply the trivial majority classifier: this algorithm learns only which is the largest class in the training set, and attributes this label to all instances. If the largest class is the positive class, the resulting confusion matrix is $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ n^- & 0 \end{pmatrix}$, and thus accuracy = ni . If the dataset is highly unbalanced, $ni \approx 1$, and thus the accuracy measure gives an unreliable estimation of the goodness of the classifier. Note that, although we achieved this result by mean of the trivial classifier, this is quite a common effect: as stated by Blagus and Lusa [99], several classifiers are biased towards the largest class in unbalanced studies.

Finally, consider another trivial algorithm, the coin tossing classifier: this classifier randomly attributes to each sample, the label positive or negative with probability $\frac{1}{2}$. Applying the coin tossing classifier to any binary dataset gives an accuracy with expected value $\frac{1}{2}$, since $\langle \mathbf{M} \rangle = \begin{pmatrix} n^+/2 & n^+/2 \\ n^-/2 & n^-/2 \end{pmatrix}$.

Matthews correlation coefficient (MCC). As an alternative measure unaffected by the unbalanced datasets issue, the Matthews correlation coefficient is a contingency matrix method of calculating the *Pearson product-moment correlation coefficient* [22] between actual and predicted values. In terms of the entries of \mathbf{M} , MCC reads as follows:

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

(worst value: -1; best value: +1)

MCC is the only binary classification rate that generates a high score only if the binary predictor was able to correctly predict the majority of positive data instances and the majority of negative data instances [80, 97].

It ranges in the interval $[-1, +1]$, with extreme values -1 and +1 reached in case of perfect misclassification and perfect classification, respectively, while $\text{MCC} = 0$ is the expected value for the coin tossing classifier.

A potential problem with MCC lies in the fact that MCC is undefined when a whole row or column of \mathbf{M} is zero, as it happens in the previously cited case of the trivial majority classifier. However, some mathematical considerations can help meaningfully fill in the gaps for these cases. If \mathbf{M} has only one non-zero entry, this means that all samples in the dataset belong to one class, and they are either all correctly (for $\text{TP} \neq 0$ or $\text{TN} \neq 0$) or incorrectly (for $\text{FP} \neq 0$ or $\text{FN} \neq 0$) classified. In this situations, $\text{MCC} = 1$ for the former case and $\text{MCC} = -1$ for the latter case. We are then left with the four cases where a row or a column of \mathbf{M} are zero, while the other two entries are non zero. That is, when \mathbf{M} is one of $\begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix}$, $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 \\ b & a \end{pmatrix}$ or $\begin{pmatrix} 0 & b \\ 0 & a \end{pmatrix}$, with $a, b \geq 1$: n in all four cases, MCC takes the indefinite form $\frac{0}{0}$. To detect a meaningful value of MCC for these four cases, we proceed through a simple approximation via a calculus technique. If we substitute the zero entries in the above matrices with the arbitrarily small value ϵ , in all four cases, we obtain

$$\begin{aligned} \text{MCC} &= \frac{a\epsilon - b\epsilon}{\sqrt{(a+b)(a+\epsilon)(b+\epsilon)(\epsilon+\epsilon)}} \\ &= \frac{\epsilon}{\sqrt{\epsilon}} \frac{a-b}{\sqrt{2(a+b)(a+\epsilon)(b+\epsilon)}} \\ &\approx \sqrt{\epsilon} \frac{a-b}{\sqrt{2ab(a-b)}} \rightarrow 0 \quad \text{for } \epsilon \rightarrow 0 \end{aligned}$$

With these positions MCC is now defined for all confusion matrices \mathbf{M} . As a consequences, $\text{MCC} = 0$ for the trivial majority classifier, and 0 is also the expected value for the coin tossing classifier.

Finally, in some cases it might be useful to consider the *normalized* MCC, defined as $\text{nMCC} = \frac{\text{MCC}+1}{2}$, and linearly projecting the original range into the interval $[0,1]$, with $\text{nMCC} = \frac{1}{2}$ as the average value for the coin tossing classifier.

F_1 score. This metric is the most used member of the parametric family of the F -measures, named after the parameter value $\beta = 1$. F_1 score is defined as the

harmonic mean of precision and recall (Table 2) and as a function of \mathbf{M} , has the following shape:

$$F_1 \text{ score} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

(worst value: 0; best value: 1)

F_1 ranges in $[0, 1]$, where the minimum is reached for $\text{TP} = 0$, that is, when all the positive samples are misclassified, and the maximum for $\text{FN} = \text{FP} = 0$, that is for perfect classification. Two main features differentiate F_1 from MCC and accuracy: F_1 is independent from TN, and it is not symmetric for class swapping.

F_1 is not defined for confusion matrices $\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & n^- \end{pmatrix}$: we can set $F_1 = 1$ for these cases. It is also worth mentioning that, when defining the F_1 score as the harmonic mean of precision and recall, the cases $\text{TP} = 0$, $\text{FP} > 0$, and $\text{FN} > 0$ remain undefined, but using the expression $\frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$, the F_1 score is defined even for these confusion matrices and its value is zero.

When a trivial majority classifier is used, due to the asymmetry of the measure, there are two different cases: if $n^+ > n^-$, then $\mathbf{M} = \begin{pmatrix} n^+ & 0 \\ n^- & 0 \end{pmatrix}$ and $F_1 = \frac{2n^+}{2n^+ + n^-}$, while if $n^- > n^+$ then $\mathbf{M} = \begin{pmatrix} 0 & n^+ \\ 0 & n^- \end{pmatrix}$, so that $F_1 = 0$. Further, for the coin tossing algorithm, the expected value is $F_1 = \frac{2n^+}{3n^+ + n^-}$.

Relationship between measures

After having introduced the statistical background of Matthews correlation coefficient and the other two measures to which we compare it (accuracy and F_1 score), we explore here the correlation between these three rates. To explore these statistical correlations, we take advantage of the Pearson correlation coefficient (PCC) [100], which is a rate particularly suitable to evaluate the linear relationship between two continuous variables [101]. We avoid the usage of rank correlation coefficients (such as Spearman's ρ and Kendall's τ [102]) because we are not focusing on the ranks for the two lists.

For a given positive integer $N \geq 10$, we consider all the possible $\binom{N+3}{3}$ confusion matrices for a dataset with N samples and, for each matrix, compute the accuracy, MCC and F_1 score and then the Pearson correlation coefficient for the three set of values. MCC and accuracy resulted strongly correlated, while the Pearson coefficient is less than 0.8 for the correlation of F_1 with the other two measures (Table 3). Interestingly, the correlation grows with N , but the increments are limited.

Similar to what Flach and colleagues did for their iso-metrics strategy [66], we depict a scatterplot of the MCCs and F_1 scores for all the 21 084 251 possible confusion matrices for a toy dataset with 500 samples (Fig. 1). We

Table 3 Correlation between MCC, accuracy, and F_1 score values

N	PCC (MCC, F_1 score)	PCC (MCC, accuracy)	PCC (accuracy, F_1 score)
10	0.742162	0.869778	0.744323
25	0.757044	0.893572	0.760708
50	0.766501	0.907654	0.769752
75	0.769883	0.912530	0.772917
100	0.771571	0.914926	0.774495
200	0.774060	0.918401	0.776830
300	0.774870	0.919515	0.777595
400	0.775270	0.920063	0.777976
500	0.775509	0.920388	0.778201
1 000	0.775982	0.921030	0.778652

Pearson correlation coefficient (PCC) between accuracy, MCC and F_1 score computed on all confusion matrices with given number of samples N

take advantage of this scatterplot to overview the mutual relations between MCC and F_1 score.

The two measures are reasonably concordant, but the scatterplot cloud is wide, implying that for each value of F_1 score there is a corresponding range of values of MCC and vice versa, although with different width. In fact, for any value $F_1 = \phi$, the MCC varies approximately between $[\phi - 1, \phi]$, so that the width of the variability range is

1, independent from the value of ϕ . On the other hand, for a given value $MCC = \mu$, the F_1 score can range in $[0, \mu + 1]$ if $\mu \leq 0$ and in $[\mu, 1]$ if $\mu > 0$, so that the width of the range is $1 - |\mu|$, that is, it depends on the MCC value μ .

Note that a large portion of the above variability is due to the fact that F_1 is independent from TN: in general, all matrices $\mathbf{M} = \begin{pmatrix} \alpha & \beta \\ \gamma & x \end{pmatrix}$ have the same value $F_1 = \frac{2\alpha}{2\alpha + \beta + \gamma}$ regardless of the value of x , while the corresponding MCC values range from $-\sqrt{\frac{\beta\gamma}{(\alpha + \beta)(\alpha + \gamma)}}$ for $x = 0$ to the asymptotic $\frac{\alpha}{\sqrt{(\alpha + \beta)(\alpha + \gamma)}}$ for $x \rightarrow \infty$. For example, if we consider only the 63 001 confusion matrices of datasets of size 500 where $TP = TN$, the Pearson correlation coefficient between F_1 and MCC increases to 0.9542254.

Overall, accuracy, F_1 , and MCC show reliable concordant scores for predictions that correctly classify both positives and negatives (having therefore many TP and TN), and for predictions that incorrectly classify both positives and negatives (having therefore few TP and TN); however, these measures show discordant behaviors when the prediction performs well just with one of the two binary classes. In fact, when a prediction displays many true positives but few true negatives (or many true negatives but few true positives) we will show that F_1 and accuracy can provide misleading information, while MCC

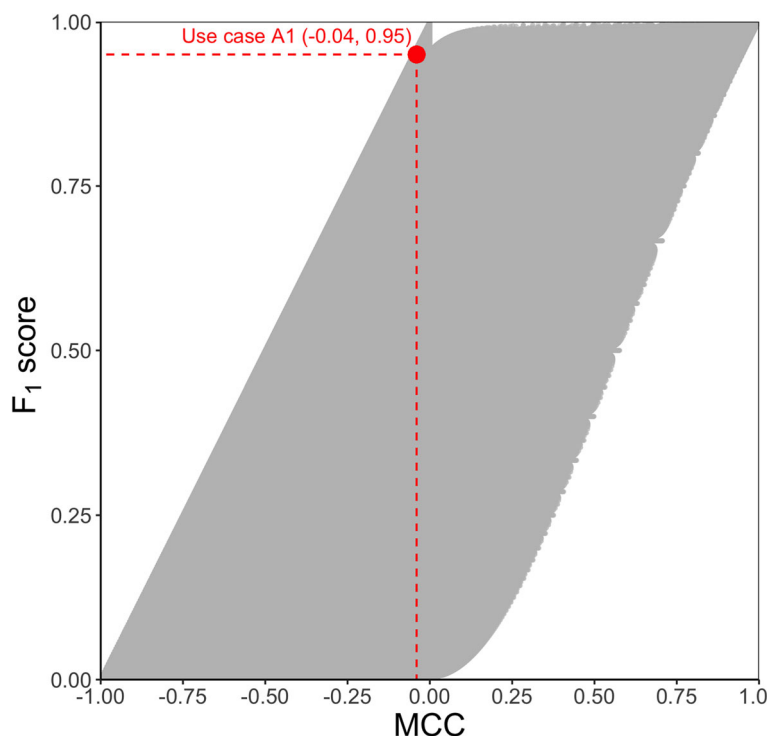


Fig. 1 Relationship between MCC and F_1 score. Scatterplot of all the 21 084 251 possible confusion matrices for a dataset with 500 samples on the MCC/ F_1 plane. In red, the $(-0.04, 0.95)$ point corresponding to use case A1

always generates results that reflect the overall prediction issues.

Results and discussion

Use cases

After having introduced the mathematical foundations of MCC, accuracy, and F_1 score, and having explored their relationships, here we describe some synthetic, realistic scenarios where MCC results are more informative and truthful than the other two measures analyzed.

Positively imbalanced dataset — Use case A1. Consider, for a clinical example, a positively imbalanced dataset made of 9 healthy individuals (negatives = 9%) and 91 sick patients (positives = 91%) (Fig. 2c). Suppose the machine learning classifier generated the following confusion matrix: TP=90, FN=1, TN=0, FP=9 (Fig. 2b).

In this case, the algorithm showed its ability to predict the positive data instances (90 sick patients out of 91 were correctly predicted), but it also displayed its lack of talent in identifying healthy controls (only 1 healthy individual out of 9 was correctly recognized) (Fig. 2b). Therefore,

the overall performance should be judged poor. However, accuracy and of F_1 showed high values in this case: $accuracy = 0.90$ and $F_1 \text{ score} = 0.95$, both close to the best possible value 1.00 in the $[0, 1]$ interval (Fig. 2a). At this point, if one decided to evaluate the performance of this classifier by considering only accuracy and F_1 score, he/she would overoptimistically think that the computational method generated excellent predictions.

Instead, if one decided to take advantage of the Matthews correlation coefficient in the Use case A1, he/she would notice the resulting $MCC = -0.03$ (Fig. 2a). By seeing a value close to zero in the $[-1, +1]$ interval, he/she would be able to understand that the machine learning method has performed poorly.

Positively imbalanced dataset — Use case A2. Suppose the prediction generated this other confusion matrix: TP = 5, FN = 70, TN = 19, FP = 6 (Additional file 1b).

Here the classifier was able to correctly predict negatives (19 healthy individuals out of 25), but was unable to correctly identify positives (only 5 sick patients out of 70). In this case, all three statistical rates showed a low score

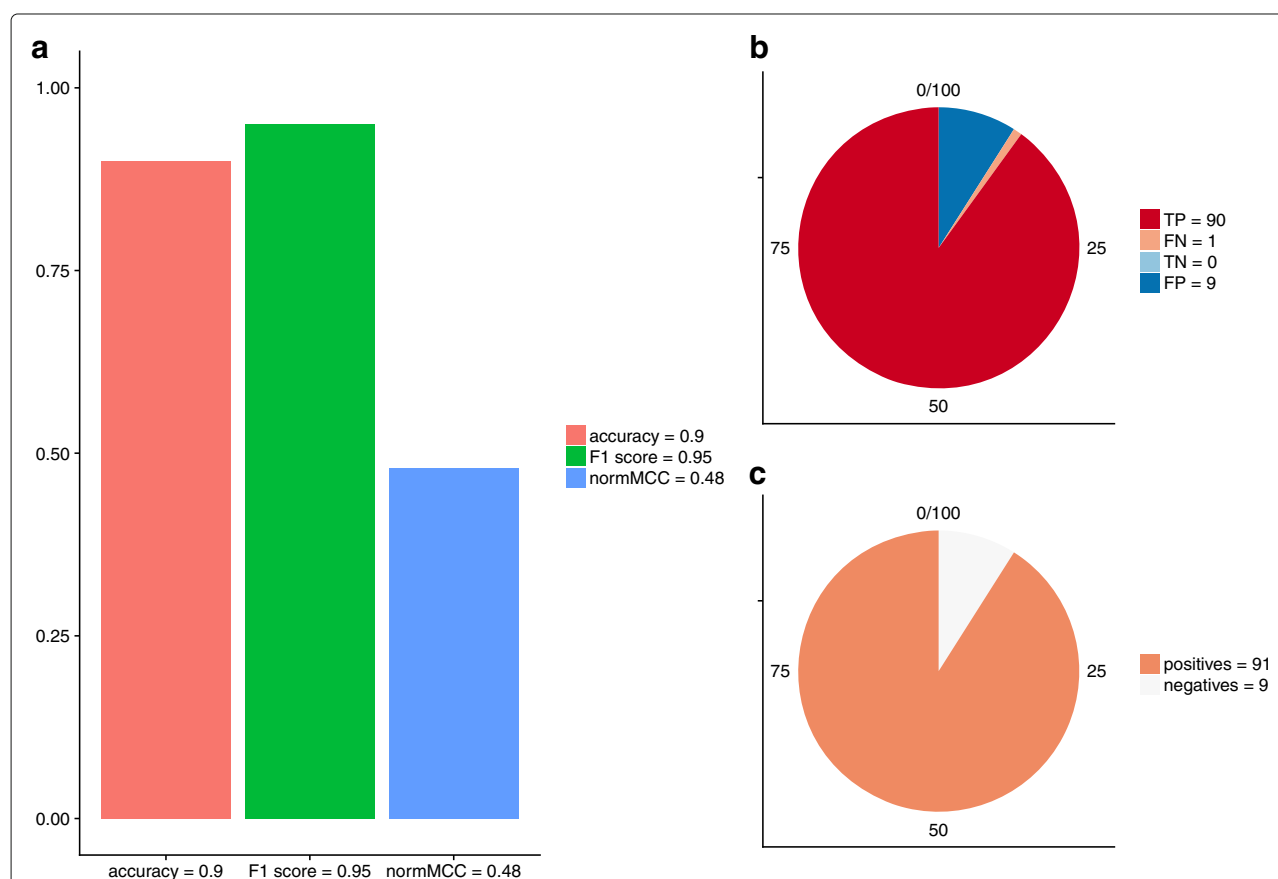


Fig. 2 Use case A1 — Positively imbalanced dataset. **a** Barplot representing accuracy, F_1 , and normalized Matthews correlation coefficient ($normMCC = (MCC + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case A1 positively imbalanced dataset. **b** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **c** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances

which emphasized the deficiency in the prediction process ($accuracy = 0.24$, $F_1 \text{ score} = 0.12$, and $MCC = -0.24$).

Balanced dataset — Use case B1. Consider now, as another example, a balanced dataset made of 50 healthy controls (negatives = 50%) and 50 sick patients (positives = 50%) (Additional file 2c). Imagine that the machine learning prediction generated the following confusion matrix: TP=47, FN=3, TN=5, FP=45 (Additional file 2b).

Once again, the algorithm exhibited its ability to predict the positive data instances (47 sick patients out of 50 were correctly predicted), but it also demonstrated its lack of talent in identifying healthy individuals (only 5 healthy controls of 50 were correctly recognized) (Additional file 2b). Again, the overall performance should be considered mediocre.

Checking only F_1 , one would read a good value (0.66 in the $[0, 1]$ interval), and would be overall satisfied about the prediction (Additional file 2a). Once again, this score would hide the truth: the classification algorithm has performed poorly on the negative subset. The Matthews correlation coefficient, instead, by showing a score close to random guessing (+0.07 in the $[-1, +1]$ interval) would be able to inform that the machine learning method has been on the wrong track. Also, it is worth noticing that accuracy would provide with an informative result in this case (0.52 in the $[0, 1]$ interval).

Balanced dataset — Use case B2. As another example, imagine the classifier produced the following confusion matrix: TP = 10, FN = 40, TN = 46, FP = 4 (Additional file 3b).

Similar to what happened for the Use case A2, the method was able to correctly predict many negative cases (46 healthy individuals out of 50), but failed in predicting most of positive data instances (only 10 sick patients were correctly predicted out of 50). Like for the Use case A2, accuracy, F_1 and MCC show average or low result scores ($accuracy = 0.56$, $F_1 \text{ score} = 0.31$, and $MCC = +0.17$), correctly informing you about the non-optimal performance of the prediction method (Additional file 3a).

Negatively imbalanced dataset — Use case C1. As another example, analyze now this imbalanced dataset made of 90 healthy controls (negatives = 90%) and 10 sick patients (positives = 10%) (Additional file 4c).

Assume the classifier prediction produced this confusion matrix: TP = 9, FN = 1, TN = 1, FP = 89 (Additional file 4b).

In this case, the method revealed its ability to predict positive data instances (9 sick patients out of 10 were correctly predicted), but it also has shown its lack of skill in identifying negative cases (only 1 healthy individual out of 90 was correctly recognized) (Additional file 4c). Again, the overall performance should be judged modest.

Similar to the Use case A2 and B2, all three statistical scores generated low results that reflect the mediocre

quality of the prediction: $F_1 \text{ score} = 0.17$ and $accuracy = 0.10$ in the $[0, 1]$ interval, and $MCC = -0.19$ in the $[-1, +1]$ interval (Additional file 4a).

Negatively imbalanced dataset — Use case C2. As a last example, suppose you obtained this alternative confusion matrix, through another prediction: TP = 2, FN = 9, TN = 88, FP = 1 (Additional file 5b).

Similar to the Use case A1 and B1, the method was able to correctly identify multiple negative data instances (88 healthy patients out of 89), but unable to correctly predict most of sick patients (only 2 true positives out of 11 possible elements).

Here, accuracy showed a high value: 0.90 in the $[0, 1]$ interval.

On the contrary, if one decided to take a look at F_1 and at the Matthews correlation coefficient, by noticing low values value ($F_1 \text{ score} = 0.29$ in the $[0, 1]$ interval and $MCC = +0.31$ in the $[-1, +1]$ interval), she/he would be correctly informed about the low quality of the prediction (Additional file 5a).

As we explained earlier, the key advantage of the Matthews correlation coefficient is that it generates a high quality score *only* if the prediction correctly classified a high percentage of negative data instances *and* a high percentage of positive data instances, with any class balance or imbalance.

Recap. We recap here the results obtained for the six use cases (Table 4). For the Use case A1 (negatively imbalanced dataset), the machine learning classifier was unable to correctly predict negative data instances, and it therefore produced confusion matrices featuring few true negatives (TN). There, accuracy and F_1 generated overoptimistic and inflated results, while the Matthews correlation coefficient was the only statistical rate which identified the aforementioned prediction problem, and therefore to provide a low truthful quality score.

In the Use case A2 (positively imbalanced dataset), instead, the method did not predict correctly enough positive data instances, and therefore showed few true positives. Even if accuracy showed an excessively high result score, the values of F_1 and MCC correctly reflected the low quality of the prediction.

In the Use case B1 (balanced dataset), the machine learning method was unable to correctly predict negative data instances, and therefore produced a confusion matrix featuring few true negatives (TN). In this case, F_1 generated an overoptimistic result, while accuracy and the MCC correctly produced low results that highlight an issue in the prediction.

The classifier did not find enough true positives for the Use case B2 (balanced dataset), too. In this case, all the analyzed rates (accuracy, F_1 , and MCC) produced average or low results which correctly represented the prediction issue.

Table 4 Recap of the six use cases results

	Balance		Confusion matrix				Accuracy [0, 1]	F ₁ score [0, 1]	MCC [-1, +1]	Figure	Informative Response
	Pos	Neg	TP	FN	TN	FP					
Use case A1 Positively imbalanced dataset	91	9	90	1	0	9	0.90	0.95	-0.03	Figure 2	MCC
Use case A2 Positively imbalanced dataset	75	25	5	70	19	6	0.24	0.12	-0.24	Suppl. Additional file 1	Accuracy, F₁ score, MCC
Use case B1 Balanced dataset	50	50	47	3	5	45	0.52	0.66	+0.07	Suppl. Additional file 2	Accuracy, MCC
Use case B2 Balanced dataset	50	50	10	40	46	4	0.56	0.31	+0.17	Suppl. Additional file 3	accuracy, F₁ score, MCC
Use case C1 Negatively imbalanced dataset	10	90	9	1	1	89	0.10	0.17	-0.19	Suppl. Additional file 4	accuracy, F₁ score, MCC
Use case C2 Negatively imbalanced dataset	11	89	2	9	88	1	0.90	0.29	+0.31	Suppl. Additional file 5	F₁ score, MCC

For the Use case A1, MCC is the only statistical rate able to truthfully inform the readership about the poor performance of the classifier. For the Use case B1, MCC and accuracy are able to inform about the poor performance of the classifier in the prediction of negative data instances, while for the Use case A2, B2, C1, all the three rates (accuracy, F₁, and MCC) are able to show this information. For the Use case C2, the MCC and F₁ are able to recognize the weak performance of the algorithm in predicting one of the two original dataset classes. pos: number of positives. neg: number of negatives. TP: true positives. FN: false negatives. TN: true negatives. FP: false positives. Informative response: list of confusion matrix rates able to reflect the poor performance of the classifier in the prediction task. We highlighted in bold the informative response of each use case

Also in the Use case C1 (positively imbalanced dataset), the machine learning method was unable to correctly recognize negative data instances, and therefore produced a confusion matrix with a low number of true negative (TN). Here, accuracy again generated an overoptimistic inflated score, while F₁ and the MCC correctly produced low results that indicated a problem in the prediction process.

Finally, in the last Use case C2 (positively imbalanced dataset), the prediction technique failed in predicting negative elements, and therefore its confusion matrix showed a low percentage of true negatives. Here accuracy again generated overoptimistic, misleading, and inflated high results, while F₁ and MCC were able to produce a low score that correctly reflected the prediction issue.

In summary, even if F₁ and accuracy results were able to reflect the prediction issue in some of the six analyzed use cases, the Matthews correlation coefficient was the only score which correctly indicated the prediction problem in all six examples (Table 4).

Particularly, in the Use case A1 (a prediction which generated many true positives and few true negatives on a positively imbalanced dataset), the MCC was the only statistical rate able to truthfully highlight the classification problem, while the other two rates showed misleading results (Fig. 2).

These results show that, while accuracy and F₁ score often generate high scores that do not inform the user about ongoing prediction issues, the MCC is a robust, useful, reliable, truthful statistical measure able to correctly reflect the deficiency of any prediction in any dataset.

Genomics scenario: colon cancer gene expression

In this section, we show a real genomics scenario where the Matthews correlation coefficient result being more informative than accuracy and F₁ score.

Dataset. We trained and applied several machine learning classifiers to gene expression data from the microarray experiments of colon tissue released by Alon et al. [103] and made it publically available within the Partial Least Squares Analyses for Genomics (*pls-genomics*) R package [104, 105]. The dataset contains 2,000 gene probsets for 62 patients, of which 22 are healthy controls and 40 have colon cancer (35.48% negatives and 64.52% positives) [106].

Experiment design. We employed machine learning binary classifiers to predict patients and healthy controls in this dataset: gradient boosting [107], decision tree [108], *k*-nearest neighbors (*k*-NN) [109], support vector machine (SVM) with linear kernel [7], and support vector machine with radial Gaussian kernel [7].

For gradient boosting and decision tree, we trained the classifiers on a training set containing 80% of randomly selected data instances, and test them on the test set containing the remaining 20% data instances. For k -NN and SVMs, we split the dataset into training set (60% data instances, randomly selected), validation set (20% data instances, randomly selected), and the test set (remaining 20% data instances). We used the validation set for the hyper-parameter optimization grid search [97]: number k of neighbors for k -NN, and cost C hyper-parameter for the SVMs. We trained each model having a different hyper-parameter on the training set, applied it to the validation set, and then picked the one obtaining the highest MCC as final model to be applied to the test set. For all the classifiers, we repeated the experiment execution ten times and recorded the average results for MCC, F_1 score, accuracy, true positive (TP) rate, and true negative (TN) rate.

We then ranked the results obtained on the test sets or the validation sets first based on the MCC, then based on the F_1 score, and finally based on the accuracy (Table 5).

Results: different metric, different ranking. The three rankings we employed to report the same results (Table 5) show two interesting aspects. First, the top classifier changes when we consider the ranking based on MCC, F_1

score, or accuracy. In the MCC ranking, in fact, the top performing method is gradient boosting (MCC = +0.55), while in the F_1 score ranking and in the accuracy ranking the best classifier resulted being k -NN (F_1 score = 0.87 and accuracy = 0.81). The ranks of the other methods change, too: linear SVM is ranked forth in the MCC ranking and in the accuracy ranking, but ranked second in the F_1 score ranking. Decision tree changes its position from one ranking to another, too.

As mentioned earlier, for binary classifications like this, we prefer to focus on the ranking obtained by the MCC, because this rate generates a high score only if the classifier was able to correctly predict the majority of the positive data instances and the majority of the negative data instances. In our example, in fact, the top MCC ranking classifier gradient boosting did quite well both on the recall (TP rate = 0.85) and on the specificity (TN rate = 0.69). k -NN, that is the top performing method both in the F_1 score ranking and in the accuracy ranking, instead, obtained an excellent score for recall (TP rate = 0.92) but just sufficient on the specificity (TN rate = 0.52).

The F_1 score ranking and the accuracy ranking, in conclusion, are hiding this important flaw of the top classifier: k -NN was unable to correctly predict a high percentage of patients. The MCC ranking, instead, takes into account this information.

Results: F_1 score and accuracy can mislead, but MCC does not. The second interesting aspect of the results we obtained relates to the radial SVM (Table 5). If a researcher decided to evaluate the performance of this method by observing only the F_1 score and the accuracy, she/he would notice good results (F_1 score = 0.75 and accuracy = 0.67) and might be satisfied about them. These results, in fact, mean 3/4 correct F_1 score and 2/3 correct accuracy.

However, these values of F_1 score and accuracy would mislead the researcher once again: with a closer look to the results, one can notice that the radial SVM has performed poorly on the true negatives (TN rate = 0.40), by correctly predicting less than half patients. Similar to the synthetic Use case A1 previously described (Fig. 2 and Table 4), the Matthews correlation coefficient is the only aggregate rate highlighting the weak performance of the classifier here. With its low value (MCC = +0.29), the MCC informs the readers about the poor general outcome of the radial SVM, while the accuracy and F_1 score show misleading values.

Conclusions

Scientists use confusion matrices to evaluate binary classification problems; therefore, the availability of a unified statistical rate that is able to correctly represent the quality of a binary prediction is essential. Accuracy and F_1

Table 5 Colon cancer prediction rankings

Classifier	MCC	F_1 score	Accuracy	TP rate	TN rate
MCC ranking:					
Gradient boosting	+0.55	0.81	0.78	0.85	0.69
Decision tree	+0.53	0.82	0.77	0.88	0.58
k -nearest neighbors	+0.48	0.87	0.80	0.92	0.52
Linear SVM	+0.41	0.82	0.76	0.86	0.53
Radial SVM	+0.29	0.75	0.67	0.86	0.40
F_1 score ranking:					
k -nearest neighbors	+0.48	0.87	0.80	0.92	0.52
Linear SVM	+0.41	0.82	0.76	0.86	0.53
Decision tree	+0.53	0.82	0.77	0.88	0.58
Gradient boosting	+0.55	0.81	0.78	0.85	0.69
Radial SVM	+0.29	0.75	0.67	0.86	0.40
Accuracy ranking:					
k -nearest neighbors	+0.48	0.87	0.80	0.92	0.52
Gradient boosting	+0.55	0.81	0.78	0.85	0.69
Decision tree	+0.53	0.82	0.77	0.88	0.58
Linear SVM	+0.41	0.82	0.76	0.86	0.53
Radial SVM	+0.29	0.75	0.67	0.86	0.40

Prediction results on colon cancer gene expression dataset, based on MCC, F_1 score, and accuracy. linear SVM: support vector machines with linear kernel. MCC: worst value -1 and best value +1. F_1 score, accuracy, TP rate, and TN rate: worst value 0 and best value 1. To avoid additional complexity and keep this table simple to read, we preferred to exclude the standard deviation of each result metric. We highlighted in bold the ranking of each rate

score, although popular, can generate misleading results on imbalanced datasets, because they fail to consider the ratio between positive and negative elements. In this manuscript, we explained the reasons why Matthews correlation coefficient (MCC) can solve this issue, through its mathematical properties that incorporate the dataset imbalance and its invariance for class swapping. The criterion of MCC is intuitive and straightforward: to get a high quality score, the classifier has to make correct predictions both on the majority of the negative cases, and on the majority of the positive cases, independently of their ratios in the overall dataset. F_1 and accuracy, instead, generate reliable results only when applied to balanced datasets, and produce misleading results when applied to imbalanced cases. For these reasons, we suggest all the researchers working with confusion matrices to evaluate their binary classification predictions through the MCC, instead of using F_1 score or accuracy.

Regarding the limitations of this comparative article, we recognize that additional comparisons with other rates (such as Cohen's Kappa [70], Cramér's V [37], and K measure [81]) would have provided further information about the role of MCC in binary classification evaluation. We preferred to focus on accuracy and F_1 score, instead, because accuracy and F_1 score are more commonly used in machine learning studies related to biomedical applications.

In the future, we plan to investigate further the relationship between MCC and Cohen's Kappa, Cramér's V, K measure, balanced accuracy, F macro average, and F micro average.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-019-6413-7>.

Additional file 1: Use case A2 — Positively imbalanced dataset. **(a)** Barplot representing accuracy, F_1 score, and normalized Matthews correlation coefficient ($\text{normMCC} = (\text{MCC} + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case A2 positively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 2: Use case B1 — Balanced dataset. **(a)** Barplot representing accuracy, F_1 score, and normalized Matthews correlation coefficient ($\text{normMCC} = (\text{MCC} + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case B1 balanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 3: Use case B2 — Balanced dataset. **(a)** Barplot representing accuracy, F_1 score, and normalized Matthews correlation coefficient ($\text{normMCC} = (\text{MCC} + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case B2 balanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP).

(c) Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 4: Use case C1 — Negatively imbalanced dataset.

(a) Barplot representing accuracy, F_1 score, and normalized Matthews correlation coefficient ($\text{normMCC} = (\text{MCC} + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case C1 negatively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Additional file 5: Use case C2 — Negatively imbalanced dataset.

(a) Barplot representing accuracy, F_1 score, and normalized Matthews correlation coefficient ($\text{normMCC} = (\text{MCC} + 1) / 2$), all in the $[0, 1]$ interval, where 0 is the worst possible score and 1 is the best possible score, applied to the Use case C2 negatively imbalanced dataset. **(b)** Pie chart representing the amounts of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). **(c)** Pie chart representing the dataset balance, as the amounts of positive data instances and negative data instances.

Abbreviations

k-NN: k-nearest neighbors; AUC: Area under the curve; FDA: Food and drug administration; FN: False negatives; FP: false positives. MAQC/SEQC: MicroArray II / sequencing quality control; MCC: Matthews correlation coefficient; PCC: Pearson correlation coefficient; PLS: Partial least squares; PR: Precision-recall; ROC: Receiver operating characteristic; SVM: Support vector machine; TN: True negatives; TP: True positives

Acknowledgments

The authors thank Julia Lin (University of Toronto) and Samantha Lea Wilson (Princess Margaret Cancer Centre) for their help in the English proof-reading of this manuscript, and Bo Wang (Peter Munk Cardiac Centre) for his helpful suggestions.

Authors' contributions

DC conceived the study, designed and wrote the "Use cases" section, designed and wrote the "Genomics scenario: colon cancer gene expression" section, reviewed and approved the complete manuscript. GJ designed and wrote the "Background" and the "Notation and mathematical foundations" sections, reviewed and approved the complete manuscript. Both the authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data and the R software code used in this study for the tests and the plots are publicly available at the following web URL: <https://github.com/davidechicco/MCC>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no competing interests.

Author details

¹Krembil Research Institute, Toronto, Ontario, Canada. ²Peter Munk Cardiac Centre, Toronto, Ontario, Canada. ³Fondazione Bruno Kessler, Trento, Italy.

Received: 24 May 2019 Accepted: 18 December 2019

Published online: 02 January 2020

References

- Chicco D, Rovelli C. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLoS ONE*. 2019;14(1):0208737.

2. Fernandes K, Chicco D, Cardoso JS, Fernandes J. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. *PeerJ Comput Sci.* 2018;4:154.
3. Maggio V, Chierici M, Jurman G, Furlanello C. Distillation of the clinical algorithm improves prognosis by multi-task deep learning in high-risk neuroblastoma. *PLoS ONE.* 2018;13(12):0208924.
4. Fioravanti D, Giarratano Y, Maggio V, Agostinelli C, Chierici M, Jurman G, Furlanello C. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics.* 2018;19(2):49.
5. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436.
6. Peterson LE. K-nearest neighbor. *Scholarpedia.* 2009;4(2):1883.
7. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intell Syst Appl.* 1998;13(4):18–28.
8. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of KDD 2016 – the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016. p. 785–94. <https://doi.org/10.1145/2939672.2939785>.
10. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. *Front Biosci.* 2008;13:691.
11. Nicodemus KK, Malley JD. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics.* 2009;25(15):1884–90.
12. Karimzadeh M, Hoffman MM. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *bioRxiv.* 2018;168419.
13. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet.* 2016;48(5):488.
14. Ng KLS, Mishra SK. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics.* 2007;23(11):1321–30.
15. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res.* 2006;7:1–30.
16. García S, Herrera F. An extension on “Statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons. *J Mach Learn Res.* 2008;9:2677–94.
17. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Informa Process Manag.* 2009;45:427–37.
18. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recogn Lett.* 2009;30:27–38.
19. Garcia V, Mollineda RA, Sanchez JS. Theoretical analysis of a performance measure for imbalanced data. In: *Proceedings of ICPR 2010 – the IAPR 20th International Conference on Pattern Recognition.* IEEE; 2010. p. 617–20. <https://doi.org/10.1109/icpr.2010.156>.
20. Choi S-S, Cha S-H. A survey of binary similarity and distance measures. *J Syst Cybernet Informa.* 2010;8(1):43–8.
21. Japkowicz N, Shah M. *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge: Cambridge University Press; 2011.
22. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol.* 2011;2(1):37–63.
23. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics.* 2012;13(4):2.
24. Shin SJ, Kim H, Han S-T. Comparison of the performance evaluations in classification. *Int J Adv Res Comput Commun Eng.* 2016;5(8):441–4.
25. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Comput Surv (CSUR).* 2016;49(2):31.
26. Ballabio D, Grisoni F, Todeschini R. Multivariate comparison of classification performance measures. *Chemom Intell Lab Syst.* 2018;174:33–44.
27. Tharwat A. Classification assessment methods. *Appl Comput Informa.* 2018;1–13. <https://doi.org/10.1016/j.aci.2018.08.003>.
28. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recogn.* 2019;91:216–31.
29. Anagnostopoulos C, Hand DJ, Adams NM. Measuring Classification Performance: the *hmeasure* Package. Technical report, CRAN. 2019;1–17.
30. Parker C. An analysis of performance measures for binary classifiers. In: *Proceedings of IEEE ICDM 2011 – the 11th IEEE International Conference on Data Mining.* IEEE; 2011. p. 517–26. <https://doi.org/10.1109/icdm.2011.21>.
31. Wang L, Chu F, Xie W. Accurate cancer classification using expressions of very few genes. *IEEE/ACM Trans Comput Biol Bioinforma.* 2007;4(1):40–53.
32. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: *Proceedings of Advances in Artificial Intelligence (AI 2006), Lecture Notes in Computer Science*, vol. 4304. Heidelberg: Springer; 2006. p. 1015–21.
33. Gu Q, Zhu L, Cai Z. Evaluation measures of the classification performance of imbalanced data sets. In: *Proceedings of ISICA 2009 – the 4th International Symposium on Computational Intelligence and Intelligent Systems, Communications in Computer and Information Science*, vol. 51. Heidelberg: Springer; 2009. p. 461–71.
34. Bekkar M, Djemaa HK, Alitouch TA. Evaluation measures for models assessment over imbalanced data sets. *J Informa Eng Appl.* 2013;3(10):27–38.
35. Akosa JS. Predictive accuracy: a misleading performance measure for highly imbalanced data. In: *Proceedings of the SAS Global Forum 2017 Conference.* Cary, North Carolina: SAS Institute Inc.; 2017. p. 942–2017.
36. Guilford JP. *Psychometric Methods.* New York City: McGraw-Hill; 1954.
37. Cramér H. *Mathematical Methods of Statistics.* Princeton: Princeton University Press; 1946.
38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA) Protein Struct.* 1975;405(2):442–51.
39. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics.* 2000;16(5):412–24.
40. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput Biol Chem.* 2004;28(5–6):367–74.
41. The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: a comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* 2010;28(8):827–38.
42. The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. *Nat Biotechnol.* 2014;32:903–14.
43. Liu Y, Cheng J, Yan C, Wu X, Chen F. Research on the Matthews correlation coefficients metrics of personalized recommendation algorithm evaluation. *Int J Hybrid Informa Technol.* 2015;8(1):163–72.
44. Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget.* 2017;8(57):97025.
45. Brown JB. Classifiers and their metrics quantified. *Mol Inform.* 2018;37:1700127.
46. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLoS ONE.* 2017;12(6):0177678.
47. Buckland M, Gey F. The relationship between recall and precision. *J Am Soc Inform Sci.* 1994;45(1):12–9.
48. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE.* 2015;10(3):0118432.
49. Dice LR. Measures of the amount of ecologic association between species ecology. *Ecology.* 1945;26(3):297–302.
50. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *K Dan Vidensk Sels.* 1948;5(4):1–34.
51. van Rijsbergen CJ. *Foundations of evaluation.* J Doc. 1974;30:365–73.
52. van Rijsbergen CJ, Joost C. *Information Retrieval.* New York City: Butterworths; 1979.
53. Chinchor N. MUC-4 evaluation metrics. In: *Proceedings of MUC-4 – the 4th Conference on Message Understanding.* McLean: Association for Computational Linguistics; 1992. p. 22–9.
54. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging.* 1994;13(4):716–24.

55. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation. In: *Information Retrieval Experiment*, Chap. 5. Amsterdam: Butterworths; 1981.
56. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, revisited. *Informa Process Manag.* 1992;28:467–90.
57. Lewis DD. Evaluating text categorization. In: *Proceedings of HLT 1991 – Workshop on Speech and Natural Language*. p. 312–8. <https://doi.org/10.3115/112405.112471>.
58. Lewis DD, Yang Y, Rose TG, Li F. RCV1: a new benchmark collection for text categorization research. *J Mach Learn Res.* 2004;5:361–97.
59. Tsoumakas G, Katakis I, Vlahavas IP. Random k-labelsets for multilabel classification. *IEEE Trans Knowl Data Eng.* 2011;23(7):1079–89.
60. Pillai I, Fumera G, Roli F. Designing multi-label classifiers that maximize F measures: state of the art. *Pattern Recogn.* 2017;61:394–404.
61. Lipton ZC, Elkan C, Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. In: *Proceedings of ECML PKDD 2014 – the 2014 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*, vol. 8725. Heidelberg: Springer; 2014. p. 225–39.
62. Sasaki Y. The truth of the F-measure. *Teach Tutor Mater.* 2007;1(5):1–5.
63. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc.* 2005;12(3):296–8.
64. Powers DMW. What the F-measure doesn't measure...: features, flaws, fallacies and fixes. *arXiv:1503.06410*. 2015.
65. Van Asch V. Macro- and micro-averaged evaluation measures. Technical report. 2013:1–27.
66. Flach PA, Kull M. Precision-Recall-Gain curves: PR analysis done right. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015)*. Cambridge: MIT Press; 2015. p. 838–46.
67. Yedidia A. Against the F-score. 2016. Blogpost: https://adamyedidia.files.wordpress.com/2014/11/f_score.pdf. Accessed 10 Dec 2019.
68. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. *Stat Comput.* 2018;28:539–47.
69. Xi W, Beer MA. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput Biol.* 2018;14(12):1006625.
70. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.
71. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–74.
72. McHugh ML. Interrater reliability: the Kappa statistic. *Biochem Med.* 2012;22(3):276–82.
73. Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. *Pharm Stat.* 2015;14:74–8.
74. Powers DMW. The problem with Kappa. In: *Proceedings of EACL 2012 – the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon: ACL; 2012. p. 345–55.
75. Delgado R, Tibau X-A. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE.* 2019;14(9):0222916.
76. Ben-David A. Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Syst Appl.* 2008;34:825–32.
77. Barandela R, Sánchez JS, García V, Rangel E. Strategies for learning in class imbalance problems. *Pattern Recogn.* 2003;36(3):849–51.
78. Wei J-M, Yuan X-J, Hu Q-H, Wang S-Q. A novel measure for evaluating classifiers. *Expert Syst Appl.* 2010;37:3799–809.
79. Delgado R, Núñez González JD. Enhancing confusion entropy (CEN) for binary and multiclass classification. *PLoS ONE.* 2019;14(1):0210264.
80. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE.* 2012;7(8):41882.
81. Sebastiani F. An axiomatically derived measure for the evaluation of classification algorithms. In: *Proceedings of ICTIR 2015 – the ACM SIGIR 2015 International Conference on the Theory of Information Retrieval*. New York City: ACM; 2015. p. 11–20.
82. Espíndola R, Ebecken N. On extending F-measure and G-mean metrics to multi-class problems. *WIT Trans Inf Commun Technol.* 2005;35:25–34.
83. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: *Proceedings of IAPR 2010 – the 20th IAPR International Conference on Pattern Recognition*. IEEE; 2010. p. 3121–4. <https://doi.org/10.1109/icpr.2010.764>.
84. Dubey A, Tarar S. Evaluation of approximate rank-order clustering using Matthews correlation coefficient. *Int J Eng Adv Technol.* 2018;8(2):106–13.
85. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29–36.
86. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 1997;30:1145–59.
87. Flach PA. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: *Proceedings of ICML 2003 – the 20th International Conference on Machine Learning*. Palo Alto: AAAI Press; 2003. p. 194–201.
88. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng.* 2005;17(3):299–310.
89. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett.* 2006;27(8):861–74.
90. Hand DJ. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Stat Med.* 2010;29:1502–10.
91. Suresh Babu N. Various performance measures in binary classification – An overview of ROC study. *Int J Innov Sci Eng Technol.* 2015;2(9):596–605.
92. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr.* 2008;17(2):145–51.
93. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics.* 2010;26(6):822–30.
94. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn.* 2009;77(9):103–23.
95. Ferri C, Hernández-Orallo J, Flach PA. A coherent interpretation of AUC as a measure of aggregated classification performance. In: *Proceedings of ICML 2011 – the 28th International Conference on Machine Learning*. Norristown: Omnipress; 2011. p. 657–64.
96. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. *PLoS ONE.* 2014;9(3):92209.
97. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min.* 2017;10(35):1–17.
98. Ozenne B, Subtil F, Maucourt-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol.* 2015;68(8):855–9.
99. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics.* 2010;11:523.
100. Sedgwick P. Pearson's correlation coefficient. *Br Med J (BMJ).* 2012;345:4483.
101. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest Geographicae.* 2011;30(2):87–93.
102. Chicco D, Ciceri E, Masseroli M. Extended Spearman and Kendall coefficients for gene annotation list correlation. In: *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer; 2014. p. 19–32. https://doi.org/10.1007/978-3-319-24462-4_2.
103. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci (PNAS).* 1999;96(12):6745–50.
104. Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinforma.* 2006;8(1):32–44.
105. Boulesteix A-L, Durif G, Lambert-Lacroix S, Peyre J, Strimmer K. Package 'plsgenomics'. 2018. <https://cran.r-project.org/web/packages/plsgenomics/index.html>. Accessed 10 Dec 2019.
106. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Data pertaining to the article 'Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays'. 2000. <http://genomics-pubs.princeton.edu/oncology/affydata/index.html>. Accessed 10 Dec 2019.
107. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal.* 2002;38(4):367–78.
108. Timofeev R. Classification and regression trees (CART) theory and applications. Berlin: Humboldt University; 2004.
109. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? In: *International Conference on Database Theory*. Springer; 1999. p. 217–35. https://doi.org/10.1007/3-540-49257-7_15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.