A Related work

Properly choosing an evaluation measure is a significant problem that attracted much attention in recent and long-standing research. In this section, we cover some related papers. In summary, while there are many related studies, the field lacks systematic approaches. Some papers focus on particular advantages and flaws of particular measures, while others suggest some informal properties. In this paper, we suggest a unified analysis that generalizes and extends the existing research.

With some similarities to our research, the authors of [12] formulate a list of (informal) properties that are argued to be desirable for an evaluation measure. These properties include having a natural extension to the multiclass case, low complexity and computational cost, distinctiveness and discriminability, informativeness, and favoring the minority class. While informativeness seems to be an informal analog of our *constant baseline*, the properties are not formally defined, and thus no systematic analysis of measures with respect to the properties can be given.

Another work related to our research [23] defines a list of properties by describing several transformations of the confusion matrix that do not change measure value. As a result, the authors provide a table listing which measures are invariant under which transformations. This analysis includes our *symmetry* and also *scale invariance* which we discuss further in Appendix D. However, the discussed properties are quite simple, and the work does not cover the most important and complex ones like *constant baseline*, *monotonicity*, or *distance*.

A recent paper [3] advocates using the Matthews correlation instead of F_1 and accuracy. For that, the authors use several intuitive *use cases*, where it is clear that the performance is poor, but only CC can correctly detect that in all cases. We note that all the use cases are related to our *constant baseline* property. We also conclude that CC should be preferred over F_1 and accuracy because it satisfies the constant baseline property. Importantly, our conclusion is based on a rigorous analysis and formal definition of properties. The authors of [7] advocate using the Matthews correlation instead of Cohen's Kappa since the latter may have undesirable behavior in some scenarios. Essentially, the scenarios described in [7] are examples showing that Cohen's Kappa does not satisfy our strong monotonicity requirement.

A work conceptually related to ours is [21]. In this paper, the authors define a list of properties (they refer to them as *axioms*). Some properties are similar to ours: MON is our monotonicity, FIX is somewhat similar (but not the same) to our maximal and minimal agreement, CHA is the constant baseline, and SYM is our class symmetry. The properties CON and SDE/WDE are related to singularities. In the current paper, we do not focus on singularities since measures are naturally extended to such cases, as we discuss in Section B. Another property is called Robustness to Imbalance (IMB). This property requires a constant classifier that classifies all elements to either the positive or the negative class to get a constant similarity score k_1 or k_2, respectively. One can see that our constant baseline thus implies IMB with $k_1 = k_2$. On the other hand, having $k_1 \neq k_2$ may lead to bias towards a particular class, which does not seem to be desired. The authors show that several known measures do not satisfy some of the properties and propose *K measure*, which is shifted balanced accuracy with singularities properly resolved. Let us also note that the authors advocate against CC largely because of the fact that they do not use this same straightforward resolution to the singularities for this measure. Our work differs in the following aspects. First, we consider more comprehensive lists of measures and properties and check each property for each popular measure. In particular, our properties include symmetry (in terms of interchanging labelings), distance, and approximate constant baseline. We show that in terms of the extended list of properties, there are better variants than the K measure (which we refer to as balanced accuracy). We also provide a deep theoretical analysis of properties and propose a new family of 'good' measures. In addition, we rigorously analyze the multiclass scenario, including the properties of aggregation schemes. To sum up, while there are methodological similarities, there are significant differences in the analysis and outcomes.

There are also papers focusing on properties of a particular measure, for instance, Cohen's Kappa [7, 9, 17] Confusion Entropy [6], or Balanced Accuracy [2]. Some papers go beyond the threshold measures considered in our paper. For instance, [5] theoretically analyzes how the area under a ROC curve (AUC) relates to accuracy. Another paper focusing on AUC and accuracy is [15]. This paper formally defines two properties: *degree of consistency* and *degree of discriminancy*. The degree of consistency is not a property of a measure, but rather a property of a *pair* of measures. Basically, in

Table 7: Notation

Variable	Definition
n	number of elements
m	number of classes
c_{ij}	number of elements of class i that are predicted as j
A_i	elements with true label i
B_i	elements with predicted label i
$\mathcal{C} = (c_{ij})$	$m \times m$ confusion matrix
$a_i = \sum_{j=0}^{m-1} c_{ij}$	size of i-th class in the true labeling
$b_i = \sum_{j=0}^{m-1} c_{ji}$	size of i-th class in the predicted labeling
$p_A = \frac{a_1}{n}$, $p_B = \frac{b_1}{n}$	fraction of positive entries (for binary classification)
$p_{AB} = \frac{c_{11}}{n}$	fraction of agreeing positives (for binary classification)
$M(\mathcal{C})$, $M(A, B)$, $M(p_{AB}, p_A, p_B)$	classification validation measure

our experiments on synthetic and real data, we compute such degrees of (in)consistency. The degree of discriminancy, in turn, can be reformulated as the number of different values that a measure has (in a given domain).

To the best of our knowledge, our work is the first to give a comprehensive systematic analysis of many measures and many rigorously formalized properties. Thus, it differs from the previous research that either focuses on particular measures or particular (and often informal) properties.

Going beyond particular measures, some studies compare the properties of micro- and macro-averagings [24]. However, to the best of our knowledge, our work is the first one giving a formal approach to the problem. In particular, we show that: 1) macro averaging should be preferred as it preserves more properties, 2) measures having natural extensions to multiclass should be preferred over any averaging. For instance, the multiclass variant of SBA has the same desirable properties as the binary version, i.e., it satisfies all properties except distance.

Finally, as we discuss in the main text in more detail, our work is motivated by a recent study [11] that analyzes the properties of *cluster validation* measures. We refer to this paper for an overview of related work in cluster analysis.

B More on classification validation measures

Notation For convenience, Table 7 lists notation frequently used throughout the text.

Resolving singularities When some of the classes are not present in the predicted (or, more rarely, true) labelings, some measures from Table 1 may not be defined. Let us discuss how to appropriately resolve such singularities.

For some measures, singularities can only occur when the measures maximally or minimally agree with each other. For example, the denominator of Jaccard is only zero if $a_1 = b_1 = 0$, in which case $A = B$ must hold so that the singularity is easily resolved by maximal agreement, leading to $J(A, B) = 1$.

For measures such as Matthews Coefficient, singularities can be resolved using constant baseline. For CC, a singularity can only occur whenever either $n^2 = \sum_{m=1}^{n} a_i^2$ or $n^2 = \sum_{m=1}^{n} b_i^2$. This implies that either A or B classifies all elements to the same class. If both A and B classify all elements to the same class, then the singularity can be resolved by maximal agreement (if they classify to the same class) or minimal agreement (otherwise). If one of A and B classifies all elements to the same class, then the constant baseline tells us that $M(A, B) = 0$ should hold.

Table 8: Correspondence of binary classification measures and pair-counting clustering measures

Classification	Clustering
F_1	Dice
Jaccard	Jaccard
Matthews Correlation Coefficient	Pearson Correlation Coefficient
Accuracy	Rand
Cohen's Kappa	Adjusted Rand
Symmetric Balanced Accuracy	Sokal&Sneath
Correlation Distance	Correlation Distance

Similarly, some measures, e.g., BA and SBA, contain terms c_{ii}/a_i (or c_{ii}/b_i) that may have singularities. In cases where $a_i = 0$, these singularities can be algebraically resolved by $c_{ii} = 0 = \frac{a_i b_i}{n}$. This leads to $\frac{c_{ii}}{a_i} = \frac{b_i}{n}$ and ensures that such singularities will not lead to violations of constant baseline.

Correspondence with pair-counting cluster validation measures As discussed in the main text, there is a correspondence between pair-counting cluster validation measures and binary classification validation measures. We refer to Table 8 for some corresponding pairs.

C Checking the properties

Table 2 lists which measures satisfy the discussed properties and which averaging schemes preserve them. In this section, we formally prove all the results. Recall that if a measure does not have a natural extension to the multiclass case, then we analyze its binary variant. Additionally, if a property is violated in the binary case, then we do not check it in the multiclass case.

Using the results from [11] As discussed in the previous section, there is a correspondence between some pair-counting clustering evaluation measures and classification ones. Recall that a pair-counting clustering measure is a function of N_{11}, N_{10}, N_{01}, and N_{00}, where N_{11} is the number of element-pairs belonging to the same cluster in both partitions, N_{00} is the number of pairs belonging to different clusters in both partitions, N_{10} is the number of pairs belonging to the same cluster in the true partition but to different clusters in the predicted partition, and N_{01} is the number of pairs belonging to different clusters in the true partition but to the same cluster in the predicted partition. Thus, pair-counting clustering measures are functions of TP, TN, FP, and FN defined for *classifying element-pairs* into "intra-cluster" and "inter-cluster" pairs. So, replacing N_{ij} by c_{ij} we naturally get a binary classification measure.

Using Table 8, we can use the results of [11] and transfer them to the corresponding classification measures. One can easily check that if the corresponding clustering measure satisfies a given property (minimal/maximal agreements, symmetry, monotonicity, distance, etc.), then the binary classification measure also satisfies it. Similarly, all counterexamples from [11] mean that the corresponding classification measure also violates the same property. This approach allows us to transfer all properties of F_1, Jaccard and the following negative cases: distance of CC; exact and approximate constant baselines of Acc; minimal agreement, distance, strong monotonicity of κ; distance of SBA.

C.1 Maximal and minimal agreement

Accuracy clearly satisfies maximal agreement. Indeed, substituting a diagonal confusion matrix, we get $\sum_{i=0}^{m-1} c_{ii} / \sum_{i=0}^{m-1} c_{ii} = 1$. On the other hand, if a confusion matrix is non-diagonal, then the denominator is strictly greater than $\sum_{i=0}^{m-1} c_{ii}$ and so $\mathrm{Acc} < 1$, which proves the property. Regarding the minimal agreement, we have $\mathrm{Acc} = c_{\min} = 0$ if all diagonal elements are equal to zero, otherwise $\mathrm{Acc} > 0$.

Similarly, by substituting a diagonal confusion matrix into BA, we get that $c_{\max} = 1$. Otherwise, if $\mathcal{C}$ is non-diagonal, then $c_{ii} < a_i$ for some i and $\mathrm{BA} < 1$. Clearly, the minimal agreement is satisfied with $c_{\min} = 0$.

For SBA, if $\mathcal{C}$ is diagonal, then SBA $= 1/(2m)\sum_{i=0}^{m-1} 2 = 1$. If there is a positive non-diagonal value, then for some i we have $c_{ii} < \min(a_i, b_i)$, and so SBA < 1. The minimal agreement is obvious with $c_{\min} = 0$.

Note that CC is monotonically decreasing on a_i and b_j. Thus, we get that CC $\leq (n^2 - \sum_{i=0}^{m-1} c_{ii}^2)/(n^2 - \sum_{i=0}^{m-1} c_{ii}^2) = 1$, with equivalence iff $\mathcal{C}$ is diagonal, that proves maximal agreement property with $c_{\max} = 1$. In contrast to the binary case, CC does not have minimal agreement property: considering the confusion matrices $\mathcal{C}_1 = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ and $\mathcal{C}_2 = \left(\begin{smallmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right)$ we get CC($\mathcal{C}_1$) $\neq$ CC($\mathcal{C}_2$), while $\mathcal{C}_1, \mathcal{C}_2$ are zero-diagonal.

Similarly to CC, κ is strictly lower than $\left(n\sum_{i=0}^{m-1} c_{ii} - \sum_{i=0}^{m-1} c_{ii}^2\right) / \left(n^2 - \sum_{i=0}^{m-1} c_{ii}^2\right) = 1$ if the confusion matrix is non-diagonal. Also, $\kappa = 1$ if $\mathcal{C}$ is diagonal. So, κ satisfies maximal agreement with $c_{\max} = 1$.

Note that CE $= 0$ if a classification is perfect, i.e. $\mathcal{C}$ is diagonal. Otherwise, there exists a pair (i, j) such that $c_{ij} > 0, a_i > 0, b_j > 0$ and we get that $-$CE < 0. Thus, maximal agreement property holds for $-$CE with $c_{\max} = 0$. In turn, $-$CE $= -1$ on the confusion matrix $\left(\begin{smallmatrix} 3 & 3 \\ 3 & 3 \end{smallmatrix}\right)$. At the same time, -1 is the minimal value of $-$CE. So, this example disproves the minimal agreement property.

Talking about GM, we note that GM $\leq \left((c_{00} + c_{11})\, c_{11} - c_{11}^2\right) / \left(c_{11} c_{00}\right) = 1$ with equality iff the confusion matrix $\mathcal{C}$ is diagonal. So, GM has maximal agreement property with $c_{\max} = 1$. To prove that GM has minimal agreement property we substitute a zero-diagonal confusion matrix into the measure and get the value -1, which does not depend on c_{ij}. Using strong monotonicity of GM (Lemma 2) we prove that GM satisfies minimal agreement.

As discussed above, the proof for the maximum agreement of F_1 and Jaccard follows from [11], together with counterexamples for minimal agreement.

C.2 Symmetry

Class symmetry It is clear that almost all considered measures are class symmetric: they do not change after interchanging class labels. The following measures are an exception: F_1 and J.

For GM, it is less obvious that class symmetry is satisfied. This measure is class symmetric since it can be rewritten as $\left(c_{11}c_{00} - c_{01}c_{10}\right) / \left(\sqrt[r]{\frac{1}{2}\left(a_1^r a_0^r + b_1^r b_0^r\right)}\right)$.

Symmetry This property is easily verified by interchanging a_i and b_i. This shows that all measures except BA are symmetric.

C.3 Distance

We start by proving that CD is indeed a distance for any binary and multi-class classification:

Lemma 1. *The Correlation Distance CD* $= \frac{1}{\pi}\arccos(CC)$ *is a distance for any* $m \geq 2$.

Proof. Let us represent a classification by a matrix via one-hot-encoding, i.e. $A = (a_{ij})_{i\in[n], j\in[k]}$, where $a_{ij} = \mathbb{1}\{A(i) = k\}$, and define $a_j = \sum_i a_{ij}$. Note that for two labelings A and B, the Frobenius inner product is given by

$$\langle A, B\rangle = \sum_j c_{jj},$$

where c_{jj} is the j-th diagonal entry of the confusion matrix of A and B. Next, we define $\bar{A} = (a_{ij} - \frac{a_j}{n})_{i\in[n], j\in[k]}$. Then, for two labelings A and B, the Frobenius inner product of these mappings is given by

$$\langle \bar{A}, \bar{B}\rangle = \sum_j c_{jj} - \frac{a_j b_j}{n}.$$

And the squared length equals

$$\|\bar{A}\|^2 = n - \frac{\sum_j a_j^2}{n},$$

so that

$$\mathrm{CC}(\mathcal{C}) = \frac{\langle \bar{A}, \bar{B} \rangle}{\|\bar{A}\| \cdot \|\bar{B}\|},$$

so that its arccosine is indeed the angle between $\bar{A}$ and $\bar{B}$, which is indeed a metric distance. □

For the remainder of this section, we prove that the remaining measures are not distances.

The fact that GM is not a distance follows from Theorem 1 of the main text: this measure satisfies both monotonicity and has a constant baseline, so that it cannot be a distance.

Taking classifications $A = (1, 1, 0), B = (1, 1, 1), C = (1, 0, 1)$ we get $\mathrm{CE}(A, C) > \mathrm{CE}(A, B) + \mathrm{CE}(B, C)$ that disproves the distance property.

Finally, let us check the triangle inequality of Acc: we need to prove that $1 + \left(\sum_{i=0}^{m-1} AC_{ii}\right)/n \geq \left(\sum_{i=0}^{m-1} AB_{ii}\right)/n + \left(\sum_{i=0}^{m-1} BC_{ii}\right)/n$ where AB, BC and AC are confusion matrices corresponding to classifications A, B; B, C and A, C, respectively.

Note that we can bring the classification B to the form of A without decreasing the right side of the distance inequality: denote by A_i the elements classified by A as i. If x simultaneously belongs to A_i and B_j $(j \neq i)$ then we classify it as B_i. During such a modification $\left(\sum_{i=0}^{m-1} AB_{ii}\right)/n$ will increase by $1/n$ while $\left(\sum_{i=0}^{m-1} BC_{ii}\right)/n$ will decrease by $1/n$ at most.

So, we just need to check the distance inequality in the case of $A = B$, that clearly holds.

C.4 Monotonicity

Monotonicity For a start we introduce the following fact: if a classification measure is linear in all c_{ii} for the fixed a_i and b_j, then it is monotone. More precisely, let us assume that a measure $M(\mathcal{C}) = L_0 + \sum_{i=0}^{m-1} L_i c_{ii}$, where $\{L_t\}_{t=0}^{m}$ are functions of a_i, b_j and $L_t > 0$ for $t \geq 1$, then M is monotone. We will rely heavily on this fact in the further verification of monotonicity.

Monotonicty of GM, Acc, BA and SBA will follow from their strong monotonicity.

To prove that κ and CC are monotone we simply use its linearity.

Finally, we note that $-\mathrm{CE}$ is not monotone. Here we can use the similar counterexample as in the minimal agreement paragraph. Taking $\mathcal{C} = \left(\begin{smallmatrix} 3 & 3 \\ 3 & 3 \end{smallmatrix}\right)$ we get $-\mathrm{CE} = -1$, that is the minimal value of $-\mathrm{CE}$. But we can obtain the confusion matrix $\left(\begin{smallmatrix} 0 & 6 \\ 6 & 0 \end{smallmatrix}\right)$ by a monotone transformation leaving a_i and b_j unchanged. After this transformation, the value of $-\mathrm{CE}$ should decrease, but it is already minimal on the matrix $\mathcal{C}$.

Strong monotonicity

Lemma 2. *GM is strongly monotone (border cases SBA ($r = -1$) and CC ($r = 0$) we consider separately).*

Proof. We rewrite GM in terms of $p_{AB} = c_{11}/n, p_A = (c_{10} + c_{11})/n, p_B = (c_{01} + c_{11})/n$:

$$\mathrm{GM}(p_{AB}, p_A, p_B) = \frac{p_{AB} - p_A p_B}{\sqrt[r]{\frac{1}{2}\left(p_A^r \left(1 - p_A\right)^r + p_B^r \left(1 - p_B\right)^r\right)}},\ p_A, p_B \in (0, 1).$$

Then the strong-monotonicity condition on c_{11} can be rewritten as

$$(1 - p_{AB}) \frac{\partial \mathrm{GM}}{\partial p_{AB}} + (1 - p_A) \frac{\partial \mathrm{GM}}{\partial p_A} + (1 - p_B) \frac{\partial \mathrm{GM}}{\partial p_B} > 0.$$

Substituting GM in this inequality and simplifying it we just need to prove that

$$\begin{aligned} g(p_{AB}, p_A, p_B) := {} & p_A p_B \left(1 + p_{AB}\right) \left(p_A^r \left(1 - p_A\right)^r + p_B^r \left(1 - p_B\right)^r\right) \\ & - p_B \left(p_A^2 + p_{AB}\right) p_A^r \left(1 - p_A\right)^r - p_A \left(p_B^2 + p_{AB}\right) p_B^r \left(1 - p_B\right)^r > 0, \end{aligned}$$

for positive p_A, p_B. Note that $g(p_{AB}, p_A, p_B)$ is strictly decreasing on p_{AB} since

$$\frac{\partial g}{\partial p_{AB}} = p_A^r \left(1 - p_A\right)^r p_B(p_A - 1) + p_B^r \left(1 - p_B\right)^r p_A(p_B - 1) < 0.$$

Also, we know that p_{AB} is strictly upper bounded by $\min\{p_A, p_B\}$. Let us assume that $0 < p_A \leq p_B < 1$, then for all possible p_{AB}, p_A, p_B we get

$$g\left(p_{AB}, p_A, p_B\right) > g\left(p_A, p_A, p_B\right) = p_B^r \left(1 - p_B\right)^r \left(p_B - p_A\right)\left(1 - p_B\right) \geq 0.$$

To check that GM is monotone on c_{10} we should prove that

$$-p_{AB}\frac{\partial \text{GM}}{\partial p_{AB}} + \left(1 - p_A\right)\frac{\partial \text{GM}}{\partial p_A} - p_B\frac{\partial \text{GM}}{\partial p_B} < 0,$$

which can be rewritten as

$$\begin{aligned} f(p_{AB}, p_A, p_B) \coloneqq \left(p_A p_B p_{AB} - p_A p_{AB} - p_B p_{AB} + p_{AB}\right) p_A^r \left(1 - p_A\right)^r \\ + \left(p_A p_B p_{AB} - p_A p_B^2 + p_A p_B - p_A^2 p_B\right) p_B^r \left(1 - p_B\right)^r > 0. \end{aligned}$$

Similarly to the previous case $\partial f / \partial p_{AB} > 0$ for all $p_A, p_B \in (0, 1)$. Note that p_{AB} is strictly lower bounded by $\max\{p_A + p_B - 1, 0\}$. Firstly, we will assume that $p_A + p_B \leq 1$. Then for such p_A, p_B and all possible p_{AB} we get that

$$f\left(p_{AB}, p_A, p_B\right) > f\left(0, p_A, p_B\right) = p_B^r \left(1 - p_B\right)^r p_A p_B \left(1 - p_A - p_B\right) \geq 0.$$

Secondly, we will assume that $p_A + p_B > 1$. In this case we get

$$f\left(p_{AB}, p_A, p_B\right) > f\left(p_A + p_B - 1, p_A, p_B\right) = p_A^r \left(1 - p_A\right)^r \left(p_A p_B - p_A - p_B + 1\right) \geq 0.$$

Since GM is symmetric and class symmetric the proof is complete. □

In contrast to the binary case, CC is not strongly monotone. Consider a confusion matrix $\mathcal{C} = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, then increasing c_{11} we get that the value of the measure decreases.

To prove that Acc is strongly monotone we note that $(a + 1)/(b + 1) > a/b$ for $0 < a < b$. So, Acc increases if we simultaneously increment c_{ii} and n. If we increment n and c_{ij} for $i \neq j$ then Acc decreases, which proves strong monotonicity.

The same reasoning works for BA and SBA as well. So, these measures are also strongly monotone.

C.5 Constant baseline

Exact constant baseline To check the constant baseline property, we will use the following lemma:

Lemma 3. *Suppose that the fixed true labeling A has fixed class-sizes $a_1, \ldots, a_m$ while the predicted labeling $B \sim U(b_1, \ldots, b_m)$ is random. Then $\mathbb{E}_{B \sim U(b_1, \ldots, b_m)} c_{ij} = a_i b_j / n$.*

Proof. To prove this equality we simply note that

$$\mathbb{E}_{B \sim U(b_1, \ldots, b_m)} c_{ij} = \sum_{x \in A_i} \mathbb{E}\ \mathbb{1}\{x \in B_j\} = a_i \mathbb{P}\left(\tilde{x} \in B_j\right),$$

where $\tilde{x}$ is an arbitrary element of A_i. So, it remains to check that $\mathbb{P}\left(\tilde{x} \in B_j\right) = b_j / n$. It follows from the following calculations

$$\mathbb{P}\left(\tilde{x} \in B_j\right) = \mathbb{E} \sum_{y \in B_j} \mathbb{1}\{\tilde{x} = y\} = \frac{b_j}{n}.$$

□

Suppose measure $M\left(\mathcal{C}\right)$ to be a linear on c_{ii} for fixed a_i and b_j (see Subsection C.4). Then, using linearity of expectation, we reduce the verification of constant baseline property to the following question: "does $M\left(\mathcal{C}\right)$ depends on a_i, b_j if all $c_{ij} = a_i b_j / n$" , which is equivalent to checking approximate constant baseline.

The following statement allows us to prove that all measures having no approximate constant baseline also do not have exact one.

Lemma 4. *Let us assume that a measure $M(\mathcal{C})$ is scale invariant (see Definition 11), continuous ($\implies$ bounded) and has an exact constant baseline. Then it also has an approximate constant baseline.*

Proof. Fix non-negative numbers $\{a_i\}_{i=0}^{m-1}, \{b_i\}_{i=0}^{m-1}$ such that $\sum_{i=0}^{m-1} a_i = \sum_{i=0}^{m-1} b_i = n$. Then consider a fixed classification NA taken from $U(Na_1, \dots, Na_m)$ and a random classification NB taken from $U(Nb_1, \dots, Nb_m)$. These classifications have Nn elements, but for any $i \in \{1, \dots, m\}$ they keep fractions of elements classified as i.

Let us prove that for any $i, j \in \{1, \dots m\}$ the random variable c_{ij}/N converges to $a_i b_j / n$ in L_2 as $N \to \infty$. From Lemma 3 we know that $\mathbb{E}(c_{ij}/N) = a_i b_j / n$. Now we compute $\mathbb{V}ar(c_{ij})$. To do this, we recall that $c_{ij} = \sum_{x \in NA_i} \mathbb{1}\{x \in NB_j\}$, then

$$\mathbb{V}ar(c_{ij}) = \sum_{x,y \in NA_i} \mathbb{C}ov(\mathbb{1}\{x \in NB_j\}, \mathbb{1}\{y \in NB_j\}).$$

It remains to compute $\mathbb{C}ov(\mathbb{1}\{x \in NB_j\}, \mathbb{1}\{y \in NB_j\})$ in two different cases: $x = y$ and $x \neq y$. To do this, we note that

$$\mathbb{P}(x, y \in NB_j) = \frac{Nb_j(Nb_j - 1)}{Nn(Nn-1)}, \text{ for } x \neq y.$$

and $\mathbb{P}(x \in NB_j) = b_j/n$. Then

$$\mathbb{C}ov(\mathbb{1}\{x \in NB_j\}, \mathbb{1}\{y \in NB_j\}) = \mathbb{P}(x, y \in NB_j) - \mathbb{P}^2(x \in NB_j) = O(1/N).$$

Thus, we get that $\mathbb{V}ar(c_{ij}/N) = O(N)/N^2 = O(1/N)$ and prove L_2 convergence.

Now we are ready to prove the lemma. Denote by M a scale invariant, continuous measure that has constant baseline. Then

$$c_{\text{base}} = \mathbb{E}_{NB \sim U(Nb_1, \dots, Nb_m)} M(\mathcal{C}_N) = \mathbb{E} M\left(\frac{\mathcal{C}_N}{N}\right) \xrightarrow[N \to \infty]{} M\left(\frac{a_i b_j}{n}\right),$$

where $\mathcal{C}_N$ is a confusion matrix between NA and NB. It follows from the fact that convergence in distribution is implied by L_2 convergence. □

Using this argument, we reduce the verification of constant baseline to the next paragraph.

Approximate constant baseline It is clear that substituting $c_{ij} = a_i b_j / n$ into GM, CC, BA, κ and SBA we get that the results do not depend on a_i, b_j. Thus, these measures have an approximate constant baseline.

Substituting $c_{ij} = a_i b_j / n$ in $-$CE we get that the result depends on a_i and b_j. So, taking $(a_0, a_1) = (2, 1), (b_0, b_1) = (1, 2)$ and $(a_0, a_1) = (0, 3), (b_0, b_1) = (1, 2)$ we get different values of $-$CE and disprove approximate constant baseline.

C.6 Micro averaging

Let us prove that micro averaging preserves maximal agreement. If a confusion matrix $\mathcal{C}$ is diagonal, then $n - \sum_{i=0}^{m-1} c_{ii} = 0$ and $FP = FN = 0$. Substituting these values in a binary measure M, we get $c_{\max}$, which does not depend on c_{ii}. If $\mathcal{C}$ is not diagonal, then $FP = FN = n - \sum_{i=0}^{m-1} c_{ii} > 0$ and the result of the averaging will be strictly lower than $c_{\max}$.

Also, micro averaging preserves monotonicity: Increasing c_{ii} for fixed n we increase TP and TN leaving $TP + FP, TP + FN, TN + FP, TN + FN$ unchanged.

Additionally, (class) symmetry are preserved by the averaging as well.

However, strong monotonicity can be violated by averaging: incrementing c_{ij} for $i \neq j$ we increase n, so $TN = (m-2)n + \sum_{i=0}^{m-1} c_{ii}$ also increases and the averaged measure may increase. For example, consider a strongly monotone binary measure $TP + TN - FP - FN$. Then, after micro-averaging, it reduces to nm, which violates strong monotonicity.

Similarly, minimal agreement is not preserved by averaging since $TN = (m-2)n > 0$ in the case of zero-diagonal confusion matrix. Consider a measure $\mathbb{1}\{TP + TN > 0\}$ satisfying minimal agreement property, then after averaging it is constant, thus it violates minimal agreement.

To prove that micro averaging preserves distance, we consider it as a result of the following procedure. We use one-hot encoding to map each class to a binary vector. Then, we map a classificaton vector A of size n to the binary vector A^2 of size nm composed of one-hot encoding binary vectors. After that, we compute confusion matrices between the resulting binary vectors A^2, B^2, C^2 corresponding to classifications A, B, C. So, this procedure proves that for any multiclass labeling A, B, C, there exists binary labelings A^2, B^2, C^2 with confusion matrices corresponding to the result of micro averaging. The triangle inequality for micro averaged matrices then follows from the binary counterpart.

Finally, approximate constant baseline can be violated by averaging: let us take $c_{ii} = a_i b_i / n$, then after the averaging we get $TP = \sum_{i=0}^{m-1} a_i b_i / n$, which is not necessary equal to $(TP + FN)(TP + FP)/(mn) = n/m$. As an example we can consider a measure $TP - (TP + FP)(TP + FN)/(TP + FP + FN + TN)$ having constant baseline. Thus, the averaged measure is $\sum_{i=0}^{m-1} c_{ii} - n/m$, which does not have an approximate constant baseline.

C.7 Macro averaging

We start with checking maximal agreement. Consider a binary measure M satisfying maximal agreement. If $\mathcal{C}$ is diagonal then the result of the averaging is $\frac{1}{m}\sum_i M(c_{ii}, 0, 0, n - c_{ii}) = c_{\max}$ independently on n and c_{ii}. If $\mathcal{C}$ is not diagonal then one of $a_i - c_{ii} > 0$ and the averaged measure is strictly lower than $c_{\max}$.

However, minimal agreement property can be violated, since for a zero-diagonal confusion matrix the result of the averaging looks like $\frac{1}{m}\sum_i M(0, a_i, b_i, n - a_i - b_i)$. So, we can see that the fourth argument may not be equal to zero. Illustrating this, we take $\mathbb{1}\{TP + TN > 0\}$ satisfying minimal agreement property in the binary case. Then taking $\mathcal{C}_1 = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ and $\mathcal{C}_2 = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ we get that the averaging has different values on these matrices, thus minimal agreement property does not hold.

Symmetry, class symmetry and monotonicity are clearly satisfied.

The same is true for the distance property: let A, B and C be multiclass classifications with n elements and m classes. Then for each $i \in \{1, \ldots, m\}$ we can build binary labelings A^2, B^2, C^2 with "ones" on those position where the base labeling has i, on other positions we place "zeroes". For such binary labelings we get that $TP = c_{ii}$, $FN = a_i - c_{ii}$, $FP = b_i - c_{ii}$ and $TN = n - a_i - b_i + c_{ii}$. Then, applying triangle inequality to them and summing these inequalities over all $i \in \{1, \ldots, m\}$ we prove that macro averaging preserves distance.

However, strong monotonicity does not necessary hold: let us increase c_{ij}, then for $k \neq i, j$ we get that c_{kk}, a_k, b_k do not change while n increases. To illustrate this, we take the same counterexample as in the case of micro averaging: $TP + TN - FP - FN$. Then, after averaging, it takes the form of $\left(n(m-4) + 4\sum_{i=0}^{m-1} c_{ii}\right)/m$. This measure is not strongly monotone.

However, approximate and exact constant baseline are clearly preserved by this averaging due to linearity of mathematical expectation.

C.8 Weighted averaging

Almost all of the previous analysis applies to weighted averaging, except of class symmetry property: replacing a_i by b_i we may change value of the averaging. Thus we get that distance property does not preserve.

To construct counterexample for minimal agreement property we take $\mathbb{1}\{TP + TN > 0\}$ again. Then taking $\mathcal{C}_1 = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ and $\mathcal{C}_2 = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ we get different values of the averaging.

As an counterexample to strong monotonicity we can take $M = TP + TN - FP - FN$ and $\mathcal{C} = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{smallmatrix}\right)$. Then the averaged measure increase during c_{12} increment.

D Theoretical analysis

In this section, we perform a theoretical analysis of binary classification measures: first, we generalize the definition of constant baseline and theoretically compare the two non-linear distance-transformations of Matthews correlation coefficient. Then, we derive the class of measures that satisfy all properties except distance.

D.1 Higher-order approximate constant baseline

Before we generalize our definition of the constant baseline, we first introduce some additional properties. These properties differ from the properties introduced in the main text in the sense that they are not desirable in themselves, but are rather *instrumental* for the analysis of other desirable properties.

Definition 11. *A measure M is* scale-invariant *if, for any scalar $\alpha > 0$ and confusion matrix $\mathcal{C}$, $M(\alpha\mathcal{C}) = M(\mathcal{C})$.*

We remark that all measures of Table 1 are scale invariant. Furthermore, note that any binary classification measure can be written as a function of the four variables c_{11}, a_1, b_1 and n, as $c_{10} = a_1 - c_{11}$, $c_{01} = b_1 - c_{11}$ and $c_{00} = n - a_1 - b_1 + c_{11}$. Therefore, any scale-invariant binary classification measure can be written as a function of the three fractions $p_{AB} = c_{11}/n$, $p_A = a_1/n$ and $p_B = b_1/n$. Therefore, we will use the shorthand notation $M(\mathcal{C}) = M(p_{AB}, p_A, p_B)$ for the remainder of this analysis. We will write P_{AB} instead of p_{AB} whenever B is random. Note that for $B \sim U(p_B n, (1-p_B)n)$, it holds that $\mathbb{E}_{B\sim U(p_B n,(1-p_B)n)}[P_{AB}] = p_A p_B$. Thus, it can readily be seen that the approximate constant baseline is satisfied whenever $M(p_A p_B, p_A, p_B) = c_{\text{base}}$. We introduce one additional property that ensures that the measure is a well-behaved function in terms of these variables.

Definition 12. *A scale-invariant measure M is* smooth *if, for any $p_A, p_B \in (0,1)$, the Taylor series of $M(p_{AB}, p_A, p_B)$ around the point $p_{AB} = p_A p_B$ converges absolutely on the interval $p_{AB} \in [0, \min\{p_A, p_B\}]$. That is, for all $p_A, p_B \in (0,1)$ and $p_{AB} \in [0, \min\{p_A, p_B\}]$, we have*

$$\sum_{k=0}^{\infty} \left| \frac{(p_{AB} - p_A p_B)^k}{k!} \frac{\partial^k}{\partial p_{AB}^k} M(p_A p_B, p_A, p_B) \right| < \infty.$$

Note that such absolute converges implies that the Taylor series converges to $M(p_{AB}, p_A, p_B)$. We remark that all constant-baseline measures of Table 1 (main text) are actually linear functions in p_{AB} for fixed p_A, p_B. Thus, each of these is smooth. Furthermore, because Matthews Correlation is linear in p_{AB}, we have that for any transformation $f(\text{CC})$, the Taylor expansion of $f(CC)$ is given by substituting CC in the Taylor expansion of f. Thus, since the Taylor expansion of $f_1(x) = \frac{1}{\pi}\arccos(x)$ and $f_2(x) = \sqrt{2(1-x)}$ around $x = 0$ converges for $x \in [-1, 1]$, we have that $\text{CD}= f_1(\text{CC})$ and $\text{CD}' = f_2(CC)$ are also smooth measures.

This allows us to express the expected value of a measure in terms of the central moments of P_{AB}:

$$\begin{aligned}\mathbb{E}[M(P_{AB}, p_A, p_B)] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(P_{AB} - p_A p_B)^k}{k!} \frac{\partial^k}{\partial p_{AB}^k} M(p_A p_B, p_A, p_B)\right] \\ &= \sum_{k=0}^{\infty} \frac{\mathbb{E}[(P_{AB} - p_A p_B)^k]}{k!} \frac{\partial^k}{\partial p_{AB}^k} M(p_A p_B, p_A, p_B).\end{aligned}$$

Here, the absolute convergence helps bound the term inside the expectation so that the Dominated Convergence Theorem allows us to interchange summation and expectation. In this expression, the first-order term vanishes as $\mathbb{E}[P_{AB}] = p_A p_B$. Thus, we have that

$$\mathbb{E}[M(P_{AB}, p_A, p_B)] = M(p_A p_B, p_A, p_B) + \sum_{k=2}^{\infty} \frac{\mathbb{E}[(P_{AB} - p_A p_B)^k]}{k!} \frac{\partial^k}{\partial p_{AB}^k} M(p_A p_B, p_A, p_B).$$

Note that for large numbers of items, P_{AB} is highly concentrated around $p_A p_B$. Thus, the contribution of the higher-order central moments is relatively small. This leads to the following generalization of the constant baseline:

Definition 13. *A smooth measure M has a k-th* order approximate constant baseline*, if there exists a constant c_{base} such that $M(p_A p_B, p_A, p_B) = c_{base}$, while for all $\ell \in \{2, \dots, k\}$, it holds that*

$$\frac{\partial^\ell}{\partial p_{AB}^\ell} M(p_A p_B, p_A, p_B) = 0.$$

Thus, first-order constant baseline is equivalent to the approximate constant baseline. Furthermore, note that ∞-th order approximate constant baseline implies exact constant baseline since then

$$\mathbb{E}[M(P_{AB}, p_A, p_B)] = M(p_A p_B, p_A, p_B) = c_{\text{base}}.$$

While it seems likely that the exact constant baseline also implies ∞-th order constant baseline, we were not able to formally prove this. However, all constant-baseline measures of Table 1 (main text) also satisfy ∞-th order constant baseline. For this reason, we will use ∞-th order constant baseline as a substitute for the exact constant baseline when deriving measures from properties.

D.2 Constant baseline order of distance transformations

We now show that the constant baseline of CD$= \frac{1}{\pi}\arccos(\text{CC})$ is one order higher than CD$' = \sqrt{2(1-\text{CC})}$.

Statement 6. *CD$= \frac{1}{\pi}\arccos(CC)$ has a second-order approximate constant baseline while CD$' = \sqrt{2(1-CC)}$ only has a first-order approximate constant baseline.*

Proof. Note that Matthews correlation is given by

$$\text{CC}(p_{AB}, p_A, p_B) = \frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}},$$

so that it is indeed a linear function in p_{AB} for fixed p_A, p_B. Therefore, the Taylor expansions of CD and CD$'$ are obtained by simply substituting CC into the Taylor expansions of $\frac{1}{\pi}\arccos(x)$ and $\sqrt{2(1-x)}$ respectively. We have

$$\tfrac{1}{\pi}\arccos(x) = \frac{\pi}{2} - \sum_{k=0}^{\infty} \frac{(2k)! x^{2k+1}}{4^k (k!)^2 (2k+1)}, \quad \text{and} \quad \sqrt{2(1-x)} = \sqrt{2} - \sqrt{2}\sum_{k=0}^{\infty} \frac{2}{k+1}\binom{2k}{k}\left(\frac{x}{4}\right)^{k+1}$$

Thus, we see that $\sqrt{2(1-x)}$ we have a quadratic term, which we do not have for $\frac{1}{\pi}\arccos(x)$. This shows that CD has a second-order constant baseline while CD$'$ only has a first-order constant baseline. □

D.3 Deriving a class of measures satisfying all properties except distance

Let us derive the class of measures satisfying all properties from Table 2 except distance. We will also use ∞-th order constant baseline instead of the exact constant baseline as this property is easier to analyze while it coincides with exact constant baseline for all measures of Table 2 in the main text.

Theorem 3. *Let M be a smooth binary classification measure that satisfies the following properties:*

1. *∞-th order constant baseline with constant 0,*
2. *symmetry,*
3. *class symmetry,*
4. *maximal agreement with constant 1,*
5. *minimal agreement with constant -1,*
6. *strong monotonicity.*

Then, it is of the following form:

$$M(p_{AB}, p_A, p_B) = s(p_A, p_B)(p_{AB} - p_A p_B),$$

where s satisfies the following properties:
$s(p_B, p_A) = s(p_A, p_B) = s(1-p_A, 1-p_B)$,
$s(p_A, p_A) = s(p_A, 1-p_A) = \dfrac{1}{p_A(1-p_A)}$,
$s(p_A, p_B) < \min\left\{\dfrac{1}{p_A p_B}, \dfrac{1}{(1-p_A)(1-p_B)}, \dfrac{1}{\min\{p_A(1-p_B), (1-p_A)p_B\}}\right\}$ *for* $p_B \notin \{p_A, 1-p_A\}$,
$\dfrac{1}{s}\left(p_A\dfrac{\partial}{\partial p_A} + p_B\dfrac{\partial}{\partial p_B}\right)s \in \left[\min\left\{-2, -1 - \dfrac{p_A p_B}{(1-p_A)(1-p_B)}\right\}, \max\left\{\dfrac{2p_B-1}{1-p_B}, \dfrac{2p_A-1}{1-p_A}\right\}\right]$,
$\dfrac{1}{s}\left((1-p_A)\dfrac{\partial}{\partial p_A} - p_B\dfrac{\partial}{\partial p_B}\right)s \in \left[\min\left\{2 - \dfrac{1}{p_A}, 2 - \dfrac{1}{1-p_B}\right\}, \max\left\{1 + \dfrac{p_B(1-p_A)}{p_A(1-p_B)}, 2\right\}\right]$

Proof. From the definition of ∞-th order constant baseline, we have that $M(p_{AB}, p_A, p_B)$ must be a linear function in p_{AB} for fixed p_A, p_B. Thus, it must be of the form

$$M(p_{AB}, p_A, p_B) = c_{\text{base}} + (p_{AB} - p_A p_B)s(p_A, p_B) = (p_{AB} - p_A p_B)s(p_A, p_B).$$

Now symmetry requires $M(p_{AB}, p_B, p_A) = M(p_{AB}, p_A, p_B)$ which leads to $s(p_B, p_A) = s(p_A, p_B)$ while class-symmetry requires $M(1 - p_A - p_B + p_{AB}, 1-p_A, 1-p_B)$, leading to $s(1-p_A, 1-p_B) = s(p_A, p_B)$. For maximal agreement, we have $M(p_{AB}, p_A, p_B) \leq 1$ with equality only if $p_{AB} = p_A = p_B$, i.e., $M(p_A, p_A, p_A) = 1$, leading to $s(p_A, p_A) = \frac{1}{p_A(1-p_A)}$. Furthermore, $M(p_{AB}, p_A, p_B) \leq M(\min\{p_A, p_B\}, p_A, p_B) < 1$ for $p_A \neq p_B$ is satisfied by

$$s(p_A, p_B) < \frac{1}{\min\{p_A, p_B\} - p_A p_B} = \frac{1}{\min\{p_A(1-p_B), (1-p_A)p_B\}}.$$

Minimal agreement requires $M(p_{AB}, p_A, p_B) \geq -1$ with equality only if $p_{AB} = 0, p_B = 1 - p_A$. For equality, we need

$$s(p_A, 1-p_A) = \frac{1}{p_A(1-p_A)}.$$

While for $p_A \neq p_B$, we need $M(p_{AB}, p_A, p_B) \geq M(\max\{0, p_A + p_B - 1\}, p_A, p_B) > -1$, leading to

$$s(p_A, p_B) < \frac{1}{\max\{p_A p_B, (1-p_A)(1-p_B)\}}.$$

Combined, this gives

$$s(p_A, p_B) < \frac{1}{\max\{p_A p_B, (1-p_A)(1-p_B), \min\{p_A(1-p_B), (1-p_A)p_B\}\}}.$$

For the remainder of the proof, we will derive that strong monotonicity is satisfied when the last two conditions of Theorem 3 hold. The first one will be derived from the increasingness of M in N_{00} while the second one will be derived from decreasingness in N_{10}. Increasingness in N_{11} and decreasingness in N_{01} will then follow from class symmetry and symmetry respectively.

We rewrite the condition $\frac{d}{dN_{00}}M$ to

$$\frac{d}{dN_{00}}M\left(\frac{N_{11}}{N_{11}+N_{10}+N_{01}+N_{00}}, \frac{N_{11}+N_{10}}{N_{11}+N_{10}+N_{01}+N_{00}}, \frac{N_{11}+N_{01}}{N_{11}+N_{10}+N_{01}+N_{00}}\right)$$
$$= -\frac{1}{N}\left[p_{AB}\frac{\partial}{\partial p_{AB}} + p_A\frac{\partial}{\partial p_A} + p_B\frac{\partial}{\partial p_B}\right]M(p_{AB}, p_A, p_B).$$

Since we want $\frac{d}{dN_{00}}M > 0$, we need

$$\left[p_{AB}\frac{\partial}{\partial p_{AB}} + p_A\frac{\partial}{\partial p_A} + p_B\frac{\partial}{\partial p_B}\right]M(p_{AB}, p_A, p_B) < 0.$$

We compute the partial derivatives of M:

$$\begin{aligned}
\frac{\partial}{\partial p_{AB}}M &= s, \\
\frac{\partial}{\partial p_A}M &= -p_B s + (p_{AB} - p_A p_B)\frac{\partial}{\partial p_A}s, \\
\frac{\partial}{\partial p_B}M &= -p_A s + (p_{AB} - p_A p_B)\frac{\partial}{\partial p_B}s.
\end{aligned} \tag{3}$$

Thus, we need

$$(p_{AB} - 2p_Ap_B) \cdot s + (p_{AB} - p_Ap_B) \left[p_A \frac{\partial}{\partial p_A} + p_B \frac{\partial}{\partial p_B} \right] s < 0,$$

for all $p_{AB} \in [\max\{p_A + p_B - 1, 0\}, \min\{p_A, p_B\}]$. Since the left hand side is linear in p_{AB}, we only need to check the upper and lower limit. Substituting $p_{AB} = \min\{p_A, p_B\}$ leads to

$$\begin{aligned}
\left[p_A \frac{\partial}{\partial p_A} + p_B \frac{\partial}{\partial p_B} \right] s &< \frac{2p_Ap_B - \min\{p_A, p_B\}}{\min\{p_A, p_B\} - p_Ap_B} s \\
&= \left(\frac{p_Ap_B}{\min\{p_A(1-p_B), p_B(1-p_A)\}} - 1 \right) s \\
&= \max\left\{ \frac{p_B}{1-p_B} - 1, \frac{p_A}{1-p_A} - 1 \right\} s \\
&= \max\left\{ \frac{2p_B - 1}{1-p_B}, \frac{2p_A - 1}{1-p_A} \right\} s,
\end{aligned}$$

while substituting $p_{AB} = \max\{0, p_A + p_B - 1\}$ gives a lower bound

$$\begin{aligned}
\left[p_A \frac{\partial}{\partial p_A} + p_B \frac{\partial}{\partial p_B} \right] s &> -\frac{2p_Ap_B - \max\{0, p_A + p_B - 1\}}{p_Ap_B - \max\{0, p_A + p_B - 1\}} s \\
&= -\left(1 + \frac{p_Ap_B}{\min\{p_Ap_B, (1-p_A)(1-p_B)\}} \right) s \\
&= -\max\left\{ 2, 1 + \frac{p_Ap_B}{(1-p_A)(1-p_B)} \right\}.
\end{aligned}$$

Combining this, we conclude that increasingness in N_{00} is satisfied whenever it holds that

$$\frac{1}{s} \left(p_A \frac{\partial}{\partial p_A} + p_B \frac{\partial}{\partial p_B} \right) s \in \left[\min\left\{ -2, -1 - \frac{p_Ap_B}{(1-p_A)(1-p_B)} \right\}, \max\left\{ \frac{2p_B - 1}{1-p_B}, \frac{2p_A - 1}{1-p_A} \right\} \right],$$

as required. The condition for decreasingness in N_{10} is obtained similarly. The condition $\frac{d}{dN_{10}} M < 0$ can be rewritten to

$$\left[-p_{AB} \frac{\partial}{\partial p_{AB}} + (1-p_A) \frac{\partial}{\partial p_A} - p_B \frac{\partial}{\partial p_B} \right] M(p_{AB}, p_A, p_B) < 0.$$

Substituting the partial derivatives from (3) gives

$$s \cdot (-p_{AB} - (1-p_A)p_B + p_Ap_B) + (p_{AB} - p_Ap_B) \left((1-p_A) \frac{\partial}{\partial p_A} - p_B \frac{\partial}{\partial p_B} \right) s < 0.$$

Again, this linear inequality should hold for all p_{AB} so that we only need to test the extremes. For $p_{AB} = \min\{p_A, p_B\}$, we find the upper bound

$$\begin{aligned}
\frac{1}{s} \left((1-p_A) \frac{\partial}{\partial p_A} - p_B \frac{\partial}{\partial p_B} \right) s &< \frac{\min\{p_A, p_B\} + (1-p_A)p_B - p_Ap_B)}{\min\{p_A, p_B\} - p_Ap_B} \\
&= \frac{\min\{p_A(1-p_B) + p_B(1-p_A), 2p_B(1-p_A)\}}{\min\{p_A(1-p_B), p_B(1-p_A)\}} \\
&= \max\left\{ 1 + \frac{p_B(1-p_A)}{p_A(1-p_B)}, 2 \right\}.
\end{aligned}$$

Substituting $p_{AB} = \max\{0, p_A + p_B - 1\}$, we get the following upper bound

$$\begin{aligned}
\frac{1}{s} \left((1-p_A) \frac{\partial}{\partial p_A} - p_B \frac{\partial}{\partial p_B} \right) s &> \frac{\max\{0, p_A + p_B - 1\} + (1-p_A)p_B - p_Ap_B)}{\max\{0, p_A + p_B - 1\} - p_Ap_B} \\
&= -\frac{\max\{p_B(1-2p_A), p_A + 2p_B - 1 - 2p_Ap_B\}}{\min\{p_Ap_B, (1-p_A)(1-p_B)\}} \\
&= \min\left\{ 2 - \frac{1}{p_A}, 2 - \frac{1}{1-p_B} \right\}.
\end{aligned}$$

Combined, we obtain the desired condition

$$\frac{1}{s}\left((1-p_A)\frac{\partial}{\partial p_A} - p_B\frac{\partial}{\partial p_B}\right)s \in \left[\min\left\{2-\frac{1}{p_A}, 2-\frac{1}{1-p_B}\right\}, \max\left\{1+\frac{p_B(1-p_A)}{p_A(1-p_B)}, 2\right\}\right]$$

□

The Generalized Means measure GM_r corresponds to $s(p_A, p_B) = M_r(p_A(1-p_A), p_B(1-p_B))^{-1}$, where M_r is the generalized mean with exponent r. It holds that this $s(p_A, p_B)$ satisfies all the conditions of Theorem 3, since it is proven in Section C that this measure indeed satisfies all these properties. The first three conditions can also be easily verified by substituting this $s(p_A, p_B)$. Verifying the last two conditions require a bit more work. The partial derivatives of $s(p_A, p_B)$ are given by

$$\begin{aligned}
&\frac{\partial}{\partial p_A}\left[\frac{1}{2}\left(p_A(1-p_A)\right)^r + \frac{1}{2}\left(p_B(1-p_B)\right)^r\right]^{-\frac{1}{r}} \\
&= -\frac{1}{r}\frac{\frac{r}{2}\left(p_A(1-p_A)\right)^{r-1}(1-2p_A)}{\left[\frac{1}{2}\left(p_A(1-p_A)\right)^r + \frac{1}{2}\left(p_B(1-p_B)\right)^r\right]^{\frac{r+1}{r}}} \\
&= \frac{2p_A-1}{p_A(1-p_A)}\frac{\left(p_A(1-p_A)\right)^r}{\left(p_A(1-p_A)\right)^r + \left(p_B(1-p_B)\right)^r}s,
\end{aligned}$$

and similarly

$$\frac{\partial}{\partial p_B}s = \frac{2p_B-1}{p_B(1-p_B)}\frac{\left(p_B(1-p_B)\right)^r}{\left(p_A(1-p_A)\right)^r + \left(p_B(1-p_B)\right)^r}s.$$

Substituting this into the condition for N_{00}-monotonicity, we get

$$\begin{aligned}
&\frac{1}{s}\left(p_A\frac{\partial}{\partial p_A} + p_B\frac{\partial}{\partial p_B}\right)s = \\
&\frac{2p_A-1}{1-p_A}\frac{\left(p_A(1-p_A)\right)^r}{\left(p_A(1-p_A)\right)^r + \left(p_B(1-p_B)\right)^r} + \frac{2p_B-1}{1-p_B}\frac{\left(p_B(1-p_B)\right)^r}{\left(p_A(1-p_A)\right)^r + \left(p_B(1-p_B)\right)^r}.
\end{aligned}$$

Note that the two large fractions sum to 1, so that we recognize this as the weighted average of $(2p_A-1)/(1-p_A)$ and $(2p_B-1)/(1-p_B)$, which are exactly the two terms in the maximum of the upper bound of the N_{00}-monotonicity condition. Furthermore, note that both these terms are larger than -1, so that the lower bound is also satisfied.

Similarly, for the condition corresponding to N_{10}-monotonicity, we get

$$\begin{aligned}
&\frac{1}{s}\left((1-p_A)\frac{\partial}{\partial p_A} - p_B\frac{\partial}{\partial p_B}\right)s = \\
&\left(2-\frac{1}{p_A}\right)\frac{\left(p_A(1-p_A)\right)^r}{\left(p_A(1-p_A)\right)^r + \frac{1}{2}\left(p_B(1-p_B)\right)^r} + \left(2-\frac{1}{1-p_B}\right)\frac{\left(p_B(1-p_B)\right)^r}{\left(p_A(1-p_A)\right)^r + \left(p_B(1-p_B)\right)^r}.
\end{aligned}$$

Again, we recognize this as the weighted average of $2-p_A^{-1}$ and $2-(1-p_B)^{-1}$, which are the terms in the minimum of the required lower bound, so that this is always satisfied. Finally, the corresponding upper bound is always satisfied since $2-p_A^{-1}$ and $2-(1-p_B)^{-1}$ can both be upper-bounded by 1. We thus conclude that GM_r indeed lies inside this class of measures for all r.

GM generalizes CC and SBA Let us show that Generalized Means indeed generalizes both CC and SBA. Recall that

$$\text{GM} = c_{\text{base}} + \frac{nc_{11} - a_1b_1}{\left(\frac{1}{2}\left(a_1^r a_0^r + b_1^r b_0^r\right)\right)^{\frac{1}{r}}}.$$

Taking $r=-1$ and $c_{\text{base}}=1$ we get a measure proportional to SBA:

$$\begin{aligned}
1 + \frac{1}{2}\left(\frac{nc_{11}}{a_0a_1} + \frac{nc_{11}}{b_0b_1} - \frac{b_1}{a_0} - \frac{b_0}{a_1}\right) &= \frac{1}{2}\left(\frac{c_{11}(a_0+a_1)}{a_0a_1} + \frac{c_{11}(b_0+b_1)}{b_0b_1} - \frac{b_1}{a_0} - \frac{b_0}{a_1} + 2\right) \\
&= \frac{1}{2}\left(\frac{c_{11}}{a_1} + \frac{c_{11}}{b_1} + \frac{n-a_1-b_1+c_{11}}{a_0} + \frac{n-a_1-b_1+c_{11}}{b_0}\right).
\end{aligned}$$

Table 9: Examples of triplets discriminating all pairs of different measures: the upper table lists the triplets, the lower table specifies which triplet discriminates a particular pair

	A	B_1	B_2
Triplet 1	(1, 1, 1, 0, 1, 1, 0, 1, 1, 0)	(1, 1, 1, 0, 1, 0, 1, 1, 1, 1)	(1, 0, 0, 1, 0, 1, 0, 1, 1, 0)
Triplet 2	(0, 1, 1, 1, 1, 0, 1, 1, 0, 1)	(1, 0, 0, 1, 0, 1, 0, 1, 1, 0)	(0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Triplet 3	(0, 0, 0, 0, 1, 1, 1, 0, 1, 0)	(1, 1, 1, 1, 1, 1, 1, 1, 0, 1)	(0, 1, 1, 1, 1, 0, 1, 1, 0, 1)
Triplet 4	(0, 1, 1, 1, 1, 0, 1, 1, 0, 1)	(1, 1, 1, 1, 1, 1, 1, 1, 0, 1)	(0, 1, 0, 1, 1, 1, 1, 1, 0, 1)
Triplet 5	(0, 0, 0, 0, 1, 1, 1, 0, 1, 0)	(0, 1, 1, 0, 0, 1, 0, 0, 0, 1)	(0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
Triplet 6	(1, 1, 1, 1, 1, 1, 1, 1, 0, 1)	(1, 1, 1, 0, 1, 1, 0, 1, 1, 0)	(0, 1, 1, 0, 0, 1, 0, 0, 0, 1)

	Acc	BA	F_1	κ	CE	GM_1	CC	SBA
Acc	—	1	2	6	6	1	5	5
BA	1	—	1	1	1	3	3	1
F_1	2	1	—	2	2	1	2	2
κ	6	1	2	—	4	1	3	3
CE	6	1	2	4	—	1	3	3
GM_1	1	3	1	1	1	—	5	1
CC	5	3	2	3	3	5	—	4
SBA	5	1	2	3	3	1	4	—

Now, let us confirm that taking $r \to 0$ and $c_{\text{base}} = 0$ leads to CC. Let $x = b_0 b_1/(a_0 a_1)$, then $\left(\frac{1}{2}(a_1^r a_0^r + b_1^r b_0^r)\right)^{\frac{1}{r}}$ can be rewritten to

$$a_0 a_1 \left(\frac{1}{2}\left(1 + x^r\right)\right)^{\frac{1}{r}} = a_0 a_1 \exp\left(\frac{1}{r} \ln\left(\frac{1}{2}\left(1 + x^r\right)\right)\right).$$

We take the limit of the exponent and use l'Hôpital to find that

$$\lim_{r \to 0} \frac{\ln\left(\frac{1}{2}\left(1 + x^r\right)\right)}{r} = \lim_{r \to 0} \frac{\ln(x) x^r}{1 + x^r} = \frac{1}{2} \ln x,$$

so that the measure indeed converges to CC.

E Additional experimental results

E.1 Binary measures

Distinguishing binary measures Let us show triplets of labelings (A, B_1, B_2) discriminating all pairs of measures in the binary classification case. Each triplet consists of the true labeling A and two predicted labelings B_1 and B_2. We say that two measures are strictly inconsistent if, according to the first one, B_1 is closer to A, while, according to the second one, B_2 is closer to A (comparing to the main text, here we consider only strict inequalities). Table 9 lists 6 triplets, where all labelings are of size $n = 10$. It also specifies which triplet discriminates each pair of measures.

Experiment within a weather forecasting service In this section, we provide a detailed analysis of the precipitation prediction task discussed in Section 5.1.

In Figure 1, we show the dependence of measures on the threshold that is used to convert soft predictions to binary labels. This is done separately for two prediction horizons: ten minutes and two hours. We make the following observations. For the ten-minute horizon, most of the measures agree that the optimal threshold is 0.9. However, confusion entropy favors the largest threshold, while balanced accuracy favors the smallest one. Interestingly, the behavior of measures significantly differs for the two-hour prediction interval. In this case, many of the measures favor either 0.6, 0.7, or 0.77. However, accuracy and CE prefer the largest threshold, while BA and SBA prefer the smallest one. Interestingly, this is the only experiment where we observe that SBA has such a noticeable disagreement with GM_1 and CC.

Table 10: Inconsistency of binary measures for rain prediction, horizon 10 minutes, %

	Acc	BA	F_1	κ	CE	GM_1	CC	SBA
Acc	—	93.3	14.4	14.4	3.3	14.4	15.0	15.0
BA	93.3	—	78.9	78.9	96.7	78.9	78.3	78.3
F_1	14.4	78.9	—	0.0	17.8	0.0	0.6	0.6
κ	14.4	78.9	0.0	—	17.8	0.0	0.6	0.6
CE	3.3	96.7	17.8	17.8	—	17.8	18.3	18.3
GM_1	14.4	78.9	0.0	0.0	17.8	—	0.6	0.6
CC	15.0	78.3	0.6	0.6	18.3	0.6	—	0.0
SBA	15.0	78.3	0.6	0.6	18.3	0.6	0.0	—

Table 11: Inconsistency of binary measures for rain prediction, horizon 2 hours, %

	Acc	BA	F_1	κ	CE	GM_1	CC	SBA
Acc	—	98.3	63.3	58.3	1.7	61.1	72.2	91.7
BA	98.3	—	35.0	39.4	100	37.2	25.6	6.1
F_1	63.3	35.0	—	4.4	65.0	2.2	8.9	28.3
κ	58.3	39.4	4.4	—	60.0	2.2	13.3	32.8
CE	1.7	100	65.0	60.0	—	62.8	73.9	93.3
GM_1	61.1	37.2	2.2	2.2	62.8	—	11.1	30.6
CC	72.2	25.6	8.9	13.3	73.9	11.1	—	18.9
SBA	91.7	6.1	28.3	32.8	93.3	30.6	18.9	—

To better understand the differences between measures, let us list average confusion matrices for the ten-minute and two-hour prediction horizons depending on a threshold (in increasing order). Here we show the relative values in percentages.

For 10 minutes:

$$\begin{pmatrix} 93.55 & 1.12 \\ 0.22 & 5.11 \end{pmatrix} \quad \begin{pmatrix} 93.76 & 0.91 \\ 0.29 & 5.04 \end{pmatrix} \quad \begin{pmatrix} 93.84 & 0.83 \\ 0.33 & 5.01 \end{pmatrix} \quad \begin{pmatrix} 93.91 & 0.76 \\ 0.36 & 4.97 \end{pmatrix} \quad \begin{pmatrix} 94.10 & 0.57 \\ 0.49 & 4.85 \end{pmatrix} \quad \begin{pmatrix} 94.33 & 0.34 \\ 0.75 & 4.59 \end{pmatrix}$$

For 2 hours:

$$\begin{pmatrix} 90.41 & 4.25 \\ 1.47 & 3.87 \end{pmatrix} \quad \begin{pmatrix} 91.32 & 3.34 \\ 1.74 & 3.60 \end{pmatrix} \quad \begin{pmatrix} 91.67 & 2.99 \\ 1.87 & 3.47 \end{pmatrix} \quad \begin{pmatrix} 91.96 & 2.70 \\ 1.99 & 3.35 \end{pmatrix} \quad \begin{pmatrix} 92.94 & 1.72 \\ 2.51 & 2.83 \end{pmatrix} \quad \begin{pmatrix} 93.98 & 0.68 \\ 3.43 & 1.91 \end{pmatrix}$$

Consider, for instance, the two smallest thresholds for the ten-minute horizon. It is easy to see that accuracy grows from 98.66% to 98.80%. In contrast, for balanced accuracy, the difference between the values can be written as:

$$\Delta\text{BA} = \frac{\Delta c_{00}}{a_0} + \frac{\Delta c_{11}}{a_1} \approx \frac{0.21}{94.67} + \frac{-0.07}{5.33} < 0.$$

So, balanced accuracy favors the smallest threshold. This can be explained by the fact that it normalizes true positives (c_{11}) by a much smaller value, so that the impact of c_{11} is much higher.

More interesting is the fact that for the ten-minute horizon, SBA agrees with most of the measures and strongly disagrees with BA. This can be explained by the fact that SBA also takes into account the distribution of predicted labels. For instance, for the two smallest thresholds, the difference becomes:

$$\Delta\text{SBA} \approx \frac{0.21}{94.67} + \frac{-0.07}{5.33} + \left(\frac{93.76}{94.05} - \frac{93.55}{93.77}\right) + \left(\frac{5.04}{5.95} - \frac{5.11}{6.23}\right) > 0.$$

Here the difference between the last two terms is positive and dominates all other differences. This happens because the false positive rate becomes significantly smaller.

Tables 10 and 11 summarize inconsistency between different measures for the ten-minute and two-hour horizons. In particular, we can see that SBA and CC always agree for the ten-minute horizon, while they have almost 20% disagreement for two hours.

E.2 Multiclass measures

Image classification The extended results are shown in Table 12. The models are the following:[7]

[7] https://github.com/rwightman/pytorch-image-models/blob/master/results/results-imagenet.csv

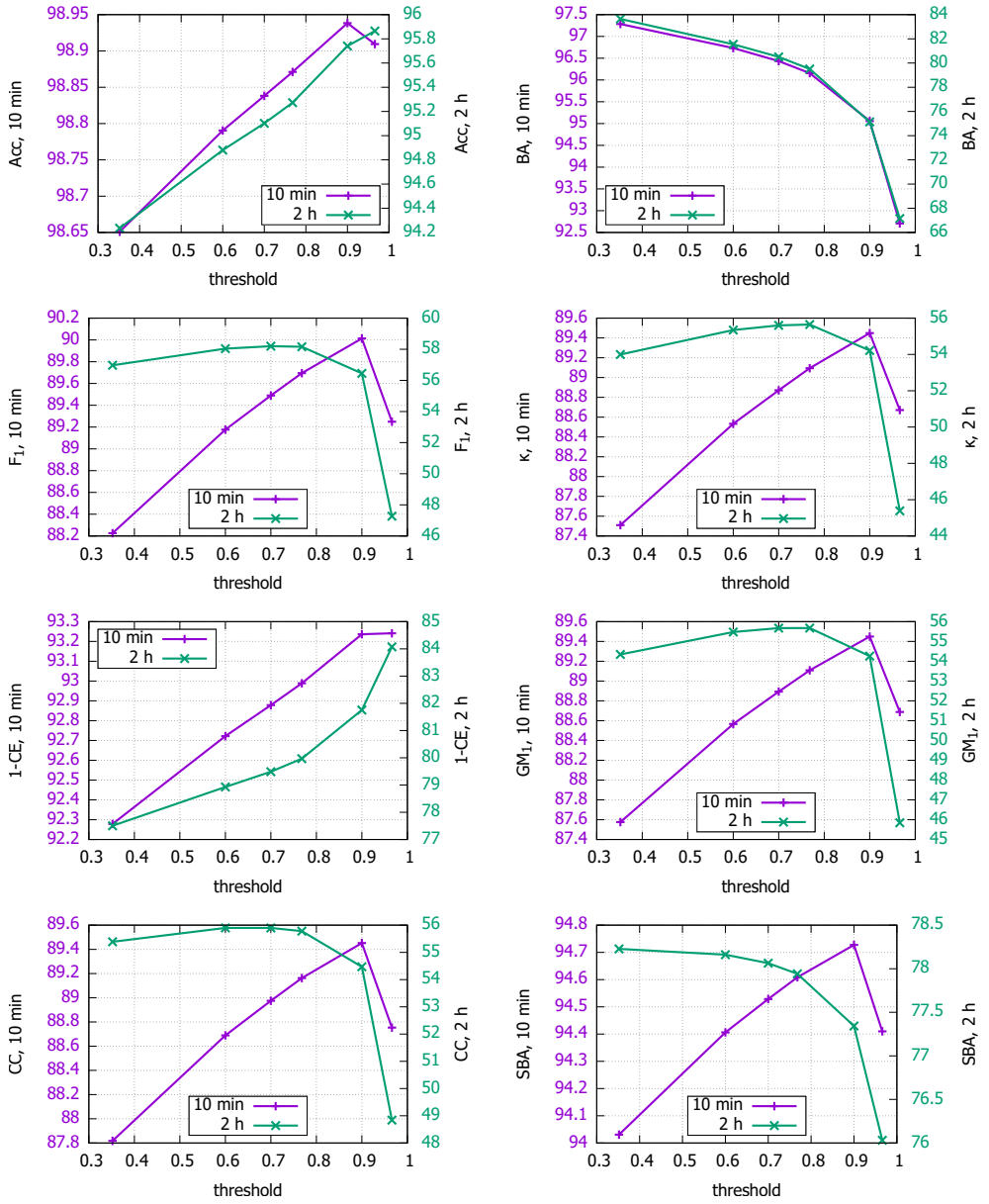

Figure 1: Dependence of measures on a threshold, for ten-minute and two-hour prediction horizons, the values are multiplied by 100

Table 12: Extended results for ImageNet, the values are multiplied by 100, inconsistencies are highlighted

	Acc/BA	F_1	J	κ	1−CE	GM_1	CC	CC^{mac}	SBA
1	88.33	88.21	80.43	88.32	94.42	88.19	88.32	88.31	88.44
2	88.23	88.08	80.25	88.21	94.38	88.07	88.22	88.20	88.35
3	87.15	87.01	78.63	87.13	93.86	87.00	87.13	87.14	87.30
4	86.83	86.64	78.08	86.82	93.64	86.63	86.82	86.78	86.95
5	86.46	86.30	77.525	86.44	93.41	86.28	86.44	86.419	86.57
6	86.43	86.27	**77.531**	86.42	**93.51**	86.26	86.42	**86.423**	**86.60**
7	86.32	86.17	77.311	86.30	93.37	86.16	86.30	86.31	86.48
8	86.31	86.12	**77.314**	86.29	**93.41**	86.10	86.30	86.28	86.47
9	86.08	85.89	76.97	86.06	93.21	85.87	86.07	86.02	86.19
10	85.72	85.55	76.51	85.70	93.05	85.53	85.70	85.70	85.89

1. tf_efficientnet_l2_ns
2. tf_efficientnet_l2_ns_475
3. swin_large_patch4_window12_384
4. tf_efficientnet_b7_ns
5. tf_efficientnet_b6_ns
6. swin_base_patch4_window12_384
7. swin_large_patch4_window7_224
8. dm_nfnet_f6
9. tf_efficientnet_b5_ns
10. dm_nfnet_f5

Note that the dataset is balanced, so accuracy coincides with BA, and weighted average coincides with macro average.

Inconsistency for Yeast dataset In this experiment, we consider the Yeast dataset[8] from the UCI repository [8]. The task is to predict protein localization sites among 10 possible variants. The class sizes are {463, 429, 244, 163, 51, 44, 35, 30, 20, 5}, so the dataset is highly unbalanced.

To this dataset, we apply the following algorithms from the scikit-learn library [20]: DecisionTree, ExtraTree, ExtraTreesEnsemble, NearestNeighbors, RadiusNeighbors, RandomForest, BernoulliNB, GaussianNB, LabelSpreading, QuadraticDiscriminantAnalysis, LinearDiscriminantAnalysis, NearestCentroid, MLPClassifier, LogisticRegression, LogisticRegressionCV, RidgeClassifier, RidgeClassifierCV, LinearSVC. Thus, there are 18 algorithms giving 153 possible pairs. For each pair of measures and each pair of algorithms we check whether the measures are consistent. Aggregating the results over all pairs of algorithms, we obtain Table 13.

We can see that for some measures the disagreement can be significant. Inconsistency is particularly high for confusion entropy, which does not satisfy most of the properties. Interestingly, the best agreement is achieved by CC and κ.

Finally, Table 14 shows inconsistency of different averagings.

References

[1] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

[2] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

[8] https://archive.ics.uci.edu/ml/datasets/Yeast

Table 13: Inconsistency of multiclass measures on the Yeast dataset, %

	Acc	BA	F_1	J	κ	CE	GM_1	CC	SBA
Acc	—	11.8	13.7	11.1	4.6	47.7	11.1	3.3	17.0
BA	11.8	—	9.8	8.5	7.2	52.9	7.2	8.5	11.8
F_1	13.7	9.8	—	2.6	10.5	48.4	5.2	10.5	4.6
J	11.1	8.5	2.6	—	9.2	49.7	6.5	9.2	7.2
κ	4.6	7.2	10.5	9.2	—	49.7	7.8	1.3	13.7
CE	47.7	52.9	48.4	49.7	49.7	—	51.0	48.4	45.1
GM_1	11.1	7.2	5.2	6.5	7.8	51.0	—	7.8	8.5
CC	3.3	8.5	10.5	9.2	1.3	48.4	7.8	—	13.7
SBA	17.0	11.8	4.6	7.2	13.7	45.1	8.5	13.7	—

Table 14: Inconsistency of averagings on the Yeast dataset

	F_1^{mic}	F_1^{mac}	F_1^{wgt}
F_1^{mic}	—	13.73	3.27
F_1^{mac}	13.73	—	10.46
F_1^{wgt}	3.27	10.46	—

	J^{mic}	J^{mac}	J^{wgt}
J^{mic}	—	11.11	2.61
J^{mac}	11.11	—	8.50
J^{wgt}	2.61	8.50	—

	CC	CC^{mic}	CC^{mac}	CC^{wgt}
CC	—	3.27	0.00	0.65
CC^{mic}	3.27	—	0.00	0.65
CC^{mac}	0.00	0.00	—	0.65
CC^{wgt}	0.65	0.65	0.65	—

	CD	CD^{mic}	CD^{mac}	CD^{wgt}
CD	—	3.27	0.00	0.65
CD^{mic}	3.27	—	0.00	0.65
CD^{mac}	0.00	0.00	—	0.65
CD^{wgt}	0.65	0.65	0.65	—

	GM_1^{mic}	GM_1^{mac}	GM_1^{wgt}
GM_1^{mic}	—	11.76	7.19
GM_1^{mac}	11.76	—	7.19
GM_1^{wgt}	7.19	7.19	—

[3] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[5] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16(16):313–320, 2004.

[6] R. Delgado and J. D. Núñez-González. Enhancing confusion entropy (cen) for binary and multiclass classification. *PloS one*, 14(1):e0210264, 2019.

[7] R. Delgado and X.-A. Tibau. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916, 2019.

[8] D. Dua and C. Graff. UCI machine learning repository, 2017.

[9] L. Flight and S. A. Julious. The disagreeable behaviour of the kappa statistic. *Pharmaceutical statistics*, 14(1):74–78, 2015.

[10] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374, 2004.

[11] M. Gösgens, L. Prokhorenkova, and A. Tikhonov. Systematic analysis of cluster similarity indices: How to validate validation measures. *International Conference on Machine Learning (ICML)*, 2021.

[12] M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[13] O. Koyejo, N. Natarajan, P. Ravikumar, and I. S. Dhillon. Consistent multilabel classification. In *NIPS*, volume 29, pages 3321–3329, 2015.

[14] V. Lebedev, V. Ivashkin, I. Rudenko, A. Ganshin, A. Molchanov, S. Ovcharenko, R. Grokhovetskiy, I. Bushmarinov, and D. Solomentsev. Precipitation nowcasting with satellite imagery. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2680–2688, 2019.

[15] C. X. Ling, J. Huang, H. Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, volume 3, pages 519–524, 2003.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[17] D. M. W. Powers. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355, 2012.

[18] P. Rao. Fine grained sentiment classification. `https://github.com/prrao87/fine-grained-sentiment`, 2021.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[20] Scikit-learn. Clustering algorithms. `https://scikit-learn.org/stable/supervised_learning.html`, 2021.

[21] F. Sebastiani. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 11–20, 2015.

[22] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[23] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[24] V. Van Asch. Macro-and micro-averaged evaluation measures. *Technical report*, 2013.

[25] J.-M. Wei, X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37(5):3799–3809, 2010.

[26] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.