# Decisions are not all equal. Introducing a utility metric based on the case-wise raters' perceptions

Andrea Campagner
DISCo, University of Milano-Bicocca
Italy

Enrico Conte
DISCo, University of Milano-Bicocca
Italy

Federico Cabitza
DISCo, University of Milano-Bicocca
Italy

## ABSTRACT

In this article we discuss a novel utility metrics for the evaluation of AI-based decision support systems, which is based on the users' perceptions of the relevance of the training cases. Relevance is a multifaceted concept, which can encompass other domain-specific dimensions, like complexity, rarity and severity in medicine. We discuss the relationship between the proposed metric and other previous proposals in the specialist literature; in particular, we show that our metric generalizes the well-known *Net Benefit*, and other similar utility-based measures. More in general, we make the point for having utility as the prime dimension to optimize machine learning models in critical domains, like the medical one, and to evaluate their potential impact on real-world practices.

**ACM Reference Format:**
Andrea Campagner, Enrico Conte, and Federico Cabitza. 2021. Decisions are not all equal. Introducing a utility metric based on the case-wise raters' perceptions. In *Proceedings of CHI '21: Workshop on Realizing AI in Healthcare: Challenges Appearing in the Wild (CHI '21).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The interest in medical AI has markedly risen in the last few years, with an increasing number of studies that show how Machine Learning (ML) models can achieve performances on par with human clinicians [20] in several diagnostic tasks. However, most of these studies regard controlled settings and are performed on retrospective data (e.g. [12, 15]), while still few studies show significant effects in real-life practice on prospectively collected data and cases (e.g. [1, 26]).

In this light, the use of reliable metrics, through which both vendors and certification bodies can attest the *validity* of the performance of ML-based decision support applications[8], is of critical importance. Such metrics should allow for sound health technology assessment, product pre-certification, and version-based approval renewals that deal with the continuous evolution and update of these systems; these metrics should also give prospective users insights on the practical usefulness of these otherwise "opaque" and difficult-to-scrutinize systems [27].


*CHI '21, May 08–09, 2021,*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Over time, many metrics to assess and report the performance of discriminative models [21, 25] have been developed; in the clinical domain, the most common metrics are based on error rates and are aimed at informing the decision maker on the capability of the model to provide the right recommendation for a new case on the basis of its performance on other past (similar) cases: accuracy, specificity, sensitivity, the $F_1$ score (that is the harmonic mean of the latter and positive predictive value) , and the area under the ROC curve (AUROC), that is an estimate of the probability that the model will rightly discriminate between a positive and negative case. Less frequently, also utility-based metrics are proposed: these are aimed at informing the decision maker on the capability of the model to limit risks and costs, in terms of pre-defined utilities: here we mention the net benefit [24] and its standardized version [18], and the relative utility [2].

Even though the former class of metrics (and, in particular, the accuracy and the AUROC) have become increasingly more popular for the evaluation of ML-based support systems, there are problems with using them as quality criteria [3]. Indeed, despite their apparent simplicity, error rate-based metrics can be misleading: A potential source of bias lies in highly imbalanced test datasets for which a so-called *accuracy paradox* has been observed [23] in that even high accuracy is not an indicator of high classifier performance. The same problem also affects the F1 score [9] and the AUROC [4].

While utility-based metrics or balanced error-based metrics (e.g. the *balanced accuracy*, the *Matthew correlation coefficient* [9] or the *Youden Index*) address these biases, by taking into account the costs (in the former case) or prevalence (in the latter one) associated with different classes, also these metrics neglect other meaningful aspects of the training and validation datasets. These include latent variations (so called *hidden stratifications* [19]) within individual cases which, in turn, could lead to variations in terms of treatment risks and benefits; or the individual perceptions of the clinicians involved in the decision making, which could influence the relative importance of correctly identifying different cases. As an example of this concept, one can easily see that the impact associated with the different decisions (e.g. treat vs not treat) is not, in general, constant and can differ from case to case: two patients, despite being affected by the same condition and in such a way that the same given treatment is suggested to both of them, may have different attitudes towards medical intervention, and the same could hold for the involved clinicians.

In this article we attempt to address these shortcomings by proposing a novel utility-based metrics, which we call *weighted utility* (*wU*). This metric generalizes existing attempts by taking into account variations in terms of impact and relevance of the individual instances on which the ML-based system is trained and

evaluated on. To present this metric we will proceed as follows: in Section 2, we will present the mathematical formulation of our proposal; then, we will show that our method encompasses, generalizing them, existing utility-based metrics, which then can be seen as a special case of our metric. Finally, we will illustrate the application of our metric in a realistic user study in the domain of radiological diagnosis (MRI of knee lesions).

## 2 METHODS

### 2.1 Weighted Utility Metrics

In this Section, we describe the proposed utility metric and we derive its relationship with other existing utility metrics.

Let $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$ be a dataset where $x_i \in X$ is an instance and $y_i \in \{0, 1\}$ is the associated target label (thus, we consider only binary classification problems): generally, we associate *normality* with class 0, and *abnormality* (that is *presence of disease* or *treatment required*) with class 1.

Our goal is to evaluate the performance of a ML model $h$: we will generally assume that such a ML model provides, for each $x_i$ a probabilistic score; in particular, with $h(x_i)$ we denote the probability score that $h$ assigns, for instance $x_i$, to the positive class (that is $h(x_i) = P(y_i = 1|x_i, h)$).

Let $\mathbf{r} : X \mapsto [0, 1]$ be an *relevance* function: this function defines, for each instance $x_i \in X$, the extent the decision maker considers it *relevant* or otherwise "how important it is that the model $h$ correctly classifies $x_i$". We note that *relevance* could represent multiple properties of instance $x_i$, for example its complexity or its rarity: we will discuss this aspect further in Section 4: here we simply assume that $\mathbf{r}$ is an abstract weighting function.

Let $\tau : X \mapsto [0, 1]$ be a probability threshold, which, for instance $x$, defines the threshold $\tau(x)$ at which one should be maximally undecided between assigning any of the 2 target labels to $x$: under this semantics, a classifier $h$ such that $h(x) \geq \tau(x)$ would be interpreted as pointing towards the positive class, while the opposite case (i.e. $h(x) < \tau(x)$) would be interpreted as evidence towards the negative class. As it is usual for utility-based metrics (where, however, the $\tau$ function is assumed to be a constant, depending only on the classification task), the $\tau$ function can be associated with a utility-based interpretation, noting that $\tau(x) = \frac{|C(x)|}{|B(x)|+|C(x)|}$, where $B(x)$ (resp. $C(x)$) is the benefit (resp. cost) of undergoing treatment (associated with the positive class) *for* case $x$.[1]

Then the *weighted utility* metrics for dataset $S$ and model $h$ is defined as:

$$wU(\tau, \mathbf{r}, S, h) = \frac{1}{\mathbf{r}(Pos)} \sum_{x_i : y_i = 1} r(x_i) \cdot \mathbf{1}_{h(x_i) \geq \tau(x_i)} \quad (1)$$

$$- \frac{1}{\mathbf{r}(Pos)} \sum_{x_i : y_i = 0} r(x_i) \cdot \frac{\tau(x_i)}{1 - \tau(x_i)} \mathbf{1}_{h(x_i) \geq \tau(x_i)}. \quad (2)$$

Where $\sigma_\gamma(\mathbf{h}(x)|\tau(x_i))$ is a monotone incresing weighting function of the parameter $\gamma \in [0, 1]$, s.t. if $\mathbf{h}(x) \geq \tau(x_i)$ then $\sigma_\gamma(\mathbf{h}(x)|\tau(x_i)) = 1$ and more in general:

$$\sigma_\gamma(\mathbf{h}(x_i)|\tau(x_i)) = \begin{cases} 0 & \mathbf{h}(x_i) < \gamma\tau(x_i) \\ \frac{\mathbf{h}(x_i) - \gamma\tau(x_i)}{(1-\gamma)\tau(x_i)} & \gamma\tau(x_i) \leq \mathbf{h}(x_i) \leq \tau(x_i) \\ 1 & \mathbf{h}(x_i) > \tau(x_i) \end{cases} \quad (3)$$

In what follows, we give an informal explanation of the above expression: we propose to see utility as the difference between the normalized and weighted true positive rate and the normalized and weighted false positive rate; intuitively a decision support is useful if the number of times it is right in detecting a problem is higher than the number of times it is wrong so. The value of our proposal lies in the concept of *weight*: true positive cases are weighted for their (case-wise) relevance ($r(x_i)$), e.g., complexity and difficulty to detect, as this aspect is perceived by the ground-truth raters. Cases are considered so according to the confidence the classifier attaches to their classification: if this is above a certain threshold, the classification is counted as positive, otherwise as negative. The same logic applies also to the 'false positive' part of the equation; however, to that respect we also consider the risk (i.e., impact, negative importance) associated with giving a wrong advice for positivity (that is in regard to actually negative cases), like e.g., over-diagnosis and over-treatment ($\tau$). The $wU$ metric allows to make all these considerations at the level of single instances: this obviously encompass the more general case, when the same weights (relevance and positivity risk) are constantly associated with all of the instances[2].

Next, we show that our the $wU$ metric represents a natural generalization of the (standardized) Net Benefit. This is defined as [25]:

$$NB(\tau) = TPR_\tau * \pi - (1-\pi) * \frac{\tau}{1-\tau} FPR_\tau \qquad sNB(\tau) = \frac{NB(\tau)}{\pi} \quad (4)$$

where $NB(\tau)$ is the Net Benefit and $\pi$ is the proportion of positive cases in S. In the following derivations, we assume that in the definition of $wU$ we have $\gamma = 1$.

Theorem 2.1. *Let, for each $x$, $\tau(x) = \tilde{\tau}$ (where $\tilde{\tau}$ is a constant) and $\mathbf{r}_1(x) = 1$. Then $wU(\tilde{\tau}, \mathbf{r}_1) = sNB(\tilde{\tau}) = \frac{NB(\tau)}{\pi}$*

Interestingly, the *weighted utility* can also be related to a certain class of accuracy metrics that account for label imbalance.

Lemma 2.2. *Let $\forall x_i \in S . \tau(x_i) = \tilde{\tau}$, $\mathbf{r}_+$ be s.t. $\mathbf{r}_+ = 1 \; \forall x_i . y_i = 1$ and $\mathbf{r}_+ = 0 \; \forall x_i . y_i = 0$, $\mathbf{r}_- = 1 - \mathbf{r}_+$ and $S^c$, $\mathbf{r}_1$ as in Theorem ??. Then, the*

[1] In this footnote, we give informal proof of this equivalence in the scenario of establishing whether a medical intervention (e.g., a treatment) should be performed or not. Let us then denote with $a, b, c, d$, respectively, the utilities related to the possible events *disease + treatment, no disease + treatment, disease + no treatment, no disease + no treatment*. Then, the $NB$ can be derived [24], by assuming that $\tau$ is a probability threshold at which a patient would be uncertain between accepting the treatment or not (that is between being considered at high-risk or not [18]), by setting $a - c = 1$ (i.e. the utility of a true positive) and solving for $b - d$ (i.e. the utility of a false positive) in $\frac{a-c}{d-b} = \frac{B}{C} \frac{1-\tau}{\tau}$, where $B$ (resp. $C$) is the benefit (resp. cost) of undergoing treatment, thus $\tau = \frac{C}{B+C}$.

[2] Likewise, considering the true positive and false positive cases is dual with respect to, respectively, false negative and true negative cases.

*following hold:*

$$TPR_{\tilde{\tau}} = wU(\tilde{\tau}, i_+, S, h) \tag{5}$$

$$TNR_{\tilde{\tau}} = wU(\tilde{\tau}, i_-, S^c, 1-h) \tag{6}$$

$$PPV_{\tilde{\tau}} = \frac{\mathbf{r}(P)wU(\tilde{\tau}, \mathbf{r}_+, S, h)}{D_{PPV}} \tag{7}$$

$$NPV_{\tilde{\tau}} = \frac{\mathbf{r}(N)wU(\tilde{\tau}, \mathbf{r}_-, S^c, 1-h)}{D_{NPV}} \tag{8}$$

$$\tag{9}$$

*where:*

$$\begin{aligned} D_{PPV} &= \mathbf{r}(P)wU(\tilde{\tau}, \mathbf{r}_+, S, h) \\ &+ \frac{1-\tilde{\tau}}{\tilde{\tau}}(\mathbf{r}(P)wU(\tilde{\tau}, \mathbf{r}_+, S, h) - \mathbf{r}(P)wU(\tilde{\tau}, \mathbf{r}_1, S, h)) \end{aligned} \tag{10}$$

$$\begin{aligned} D_{NPV} &= \mathbf{r}(N)wU(\tilde{\tau}, \mathbf{r}_-, S^c, 1-h) \\ &+ \frac{1-\tilde{\tau}}{\tilde{\tau}}\mathbf{r}(N)wU(\tilde{\tau}, \mathbf{r}_-, S^c, 1-h) \\ &- \frac{1-\tilde{\tau}}{\tilde{\tau}}\mathbf{r}(N)wU(\tilde{\tau}, \mathbf{r}_1, S^c, 1-h)) \end{aligned} \tag{11}$$

Theorem 2.3. *Let $TPR_{\tilde{\tau}}$, $TNR_{\tilde{\tau}}$, $PPV_{\tilde{\tau}}$ and $NPV_{\tilde{\tau}}$ be the true positive rate, false positive rate, positive predictive value and negative predictive value at threshold $\tilde{\tau}$. Then:*

$$BalAcc = \frac{TPR_{0.5} + TNR_{0.5}}{2} \tag{12}$$

$$J = TPR_{0.5} + TNR_{0.5} - 1 \tag{13}$$

$$Markedness = PPV_{0.5} + NPV_{0.5} - 1 \tag{14}$$

*where BalAcc is the Balanced Accuracy and J is the Youden index.*

## 2.2 Experimental Evaluation

In this Section, we report on a user-based study that we conducted in order to evaluate the viability of the proposed metrics. To this purpose, we involved 13 board-certified radiologists from several Italian hospitals, asking them to annotate a sample of 417 cases randomly extracted from the MRNet dataset[3]. This dataset encompasses 1,370 knee MRI exams performed at the Stanford University Medical Center (with 81% abnormal exams, and in particular 319 Anterior Cruciate Ligament (ACL) tears and 508 meniscal tears).

In the study we used an online questionnaire platform (Limesurvey, version 3.18[4]) and invited the participants by personal email. As anticipated above, we involved 13 radiologists (of different MRI reading skills, which we stratified in two subgroups, higher- and lower-proficiency according to self-assessment), in a diagnostic task where they were called to discriminate the MRNet cases that were positive, and indicate whether these regarded either ACL or meniscal tears: in particular they had to say whether the presented imaging presented a case of ACL tear (yes/no), or a meniscal tear (yes/no): hence two decisions in total. The radiologists were also requested to assess each case in terms of complexity (or difficulty in giving the correct answer) on an 5-level ordinal scale, and the confidence with which they classified the case, on an 6-level ordinal scale.

[3]https://stanfordmlgroup.github.io/competitions/mrnet/.
[4]https://www.limesurvey.org/

The subjective complexity and confidence ratings of the involved clinicians were then used to define the case-wise *relevance function* $\mathbf{r}$ and the case-wise probability threshold $\tau$.

In order to illustrate the application of the $wU$ metric, we developed a Deep Learning classification model, trained to perform the same task of the clinicians: more precisely, we trained an InceptionV3 Convolutional Neural Network model to discriminate between abnormal cases (that is, cases affected by either a meniscal or ACL tear) and normal cases. The training set was a collection of MRI exams taken from the MRNet dataset that were not given to the radiologists: consequently, the ML model was trained on a collection of 953 individual exams, each of which was composed of a variable number of images. The ML model was then evaluated on the 417 re-annotated images using different metrics, namely the accuracy, balanced accuracy, AUROC, (standardized) net benefit (at different threshold values) and the our *weighted utility*.

As regards the *relevance* function, we simply used the average reported complexity rating, for each case, maximum normalized so to obtain numbers in $[0, 1]$. As regards the case-wise $\tau$ values, we considered three different definitions (we will discuss the semantics behind these three definitions in Section 4):

$$\begin{aligned} \tau_{confidence}(x_i) = \frac{1}{\text{n° raters}}( &\sum_{r \text{ rater}: r(x_i)=1} \frac{c_r(x_i)+1}{2} \\ &+ \sum_{r \text{ rater}: r(x_i)=0} \frac{1-c_r(x_i)}{2}) \end{aligned} \tag{15}$$

$$\begin{aligned} \tau_{persuasion}(x_i) = \frac{1}{\text{n° raters}}( &\sum_{r \text{ rater}: r(x_i)=0} \frac{c_r(x_i)+1}{2} \\ &+ \sum_{r \text{ rater}: r(x_i)=1} \frac{1-c_r(x_i)}{2}) \end{aligned} \tag{16}$$

$$\tau_{auto-bias}(x_i) = \begin{cases} \frac{d(x_i)}{2} & |\{r : r(x_i) = 1\}| \geq |\{r : r(x_i) = 0\}| \\ \frac{2-d(x_i)}{2} & otherwise \end{cases} \tag{17}$$

where $r(x_i) \in \{0, 1\}$ is the label annotation reported by rater $r$ for case $x_i$, $c_r(x_i) \in [0, 1]$ is the (normalized) confidence reported by rater $r$ for their annotation of case $x_i$, and $d(x_i) = \frac{4}{(\text{n° raters}+1)(\text{n° raters}-1)} \sum_{r \neq r' \text{ raters}} \mathbf{1}_{r(x_i) \neq r'(x_i)}$ is the *disagreement rate*[5]. In short, $\tau_{confidence}$, for an instance to be classified as positive, requires the model's probability score to be at least as high as the average of the probabilities expressed by the raters; $\tau_{persuasion}$ requires the model's probability score to be higher than the probability that the raters assigned to the negative class; while for $\tau_{auto-bias}$ the required probability score is defined based on the disagreement among the raters.

## 3 RESULTS

The performance of the raters and of the AI model, in the ROC space, is reported in Figure 1. The average perceived case complexity was 0.70 (95% C.I [0.69, 0.71], IQR [0.63, 0.77]), the average $\tau_{confidence}$ was 0.72 (95% C.I. [0.71, 0.74], IQR [0.61, 0.87]), the average $\tau_{persuasion}$ was 0.55 (95% C.I. [0.52, 0.58], IQR [0.17, 0.85]), and the average $\tau_{auto-bias}$ was 0.54 (95% C.I [0.51, 0.58], IQR [0.14, 0.86]). The performance of the AI model, in terms of $wU$ (with three

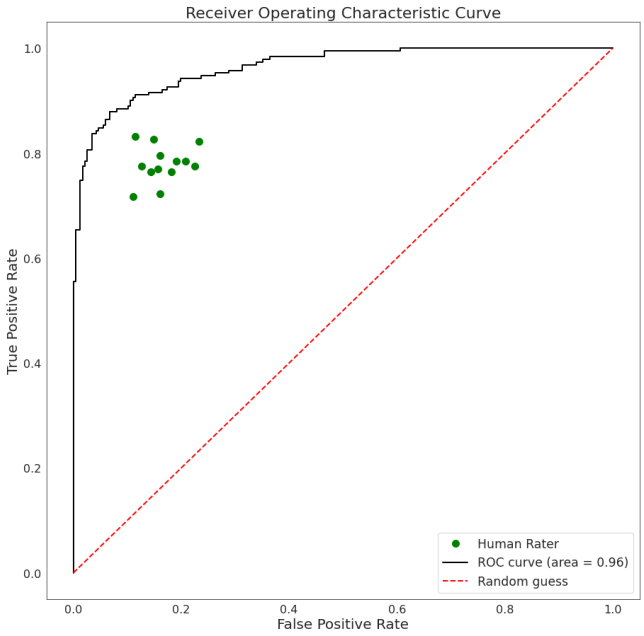


**Figure 1: Performance of the raters and of the AI model, in the ROC space**

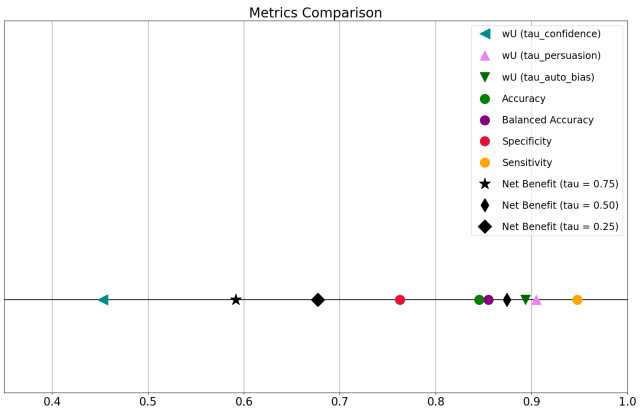


**Figure 2: The performance of the AI model, in terms of the three different versions of $wU$, and a collection of other pertinent metrics.**

different settings of the $\tau$ function), and other metrics, is reported in Figure 2.

## 4 DISCUSSION

Commenting the results, the first observation regards the large differences observed among the different values of the proposed $wU$ metrics, computed according to the three definitions of $\tau$ reported in Section 2: indeed, we can see that $wU(\tau_{confidence})$ was largely smaller than both $wu(\tau_{auto-bias})$ and $wU(\tau_{persuasion})$, while these latter two were more similar. These numerical differences reflect different semantics underlying the three definitions of $\tau$:

- As regards $\tau_{confidence}$, this parameter reflects the fact that the confidence acts as a threshold for the probability score of the model: for the classification provided by the ML model to be considered useful and reliable by a rater, the former one's confidence (i.e. the probability score) should be at least as high the rater's one. Indeed, we can see that the value of $\tau$ monotonically increases with $c_r$ when the label annotation reported by $r$ was 1, with range in $[0.5, 1]$: if the probability score $h(x_i)$ of the model is at least as high as $c_r$ (appropriately rescaled), then the model is considered to be supporting class 1. Dually, the value of $\tau$ monotonically decreases with $c_r$ when the label annotation reported by $r$ was 0, with range in $[0, 0.5]$: since in the formulation of $wU$ the value $h(x_i)$ refers to the probability score *of class 1*, we note that to be considered as supportive of class 0, the model's probability scores should be such that $h(x_i)$ is lower than $c_r$;
- On the other hand, the definition of $\tau_{persuasion}$ reflects the fact that to *persuade* a human rater in changing its opinion, the ML model should be very confident in the advice it provides.
  From the perspective of a single rater $r$, the value of $\tau$ is monotonically increasing with $c_r$ when the rater's annotation was 0: in this case, the model probability score $h(x_i)$ should be higher then $c_r$ to persuade the rater that the most appropriate label should be 1. Similarly, when the rater's annotation was 1, only a very low value of $h(x_i)$ (thus, a high probability score for class 0) would be able to persuade them of the contrary;
- Finally, $\tau_{auto-bias}$ shares some similarities with the other two definitions, but its semantics is more directly related to *risk* (albeit, at a qualitative level) and to the notion of *automation bias* [14], that is when doctors make a mistake because they over-rely on a wrong advice. Consider the case in which we have an odd number of raters (so to ensure that a qualified majority always exists), and suppose that the label reported by the majority of the raters was 0. If all raters favored option 0, then the model would be able to nudge the final decision toward option 1 only if its probability score $h(x_i) \sim 1$: in this case, however, if the model suggestion was wrong, it would have negatively affected the raters, thus incurring in greater costs (e.g. suggest treatment for an healthy patient). By contrast, if only half (plus one) of the raters selected 0, a probability score $h(x_i) \geq 0.5$ would cancel the majority and increase uncertainty: compared to the former case, such a decision would incur in lower risks, as the raters may decide to perform additional investigations or to otherwise engage in further discussion about the case at hand. Similar reasoning can justify the behavior of this definition of $\tau$ when the majority selection was class 1.

As briefly discussed in Section 2, and as widely known in the specialized literature [2], in general the definition of a probability threshold should be based on a cost-benefit analysis: nonetheless, it is not always easy to properly quantify such constructs, even more so under the usual requirements that are imposed according to standard decision theory (e.g. linearity). Indeed, this has generally been acknowledged as one of the main difficulties in applying

utility-based metrics (compared with error rate-based ones) [16]. With this respect, the three criteria that we proposed to define $\tau$ could be useful to simplify the definition of such a threshold, as these definitions only rely on simple, self-quantified constructs (i.e. confidence) or on objectively measurable characteristics (i.e. disagreement rate), while still being intuitively related to the notion of risk (as discussed for the case of $\tau_{auto-bias}$).

*Comparing wU and the Net Benefit.* As a second observation, we compare the (standardized) Net Benefit and the $wU$. In Section 2, we proved that $wU$ provides a generalization of the Net Benefit (and related metrics) by allowing the probability threshold $\tau$ to vary with the individual cases, and by attaching a degree of *relevance* to each individual case. The first factor allows to evaluate the costs and benefits of treatment vs non-treatment on an individual, case-wise basis, whenever we are to validate a ML model: this provides the $wU$ with an increased level of flexibility, as it allows to differentiate between two cases that, although identical in terms of condition (e.g. same disease and/or stadiation), still differ with respect to the risks of undergoing treatment, for example due to the different inclinations of the individual patients towards a possibly invasive treatment option, or due to characteristics that are not datafied and are thus not available to the ML model [6]. The second term, on the other hand, allows to capture case-wise differences in the perceived importance of correctly identifying a case with respect to others. In this paper we focused on *complexity* as a dimension to define relevance, because we related relevance to the importance of receiving a right advice in case of difficult to detect (i.e., complex) cases; however, other dimensions could be of interest as well: some examples include *rarity*, *severity*, *impact* (if the condition gets undetected), or any combination thereof [22]. Interestingly, it is not hard to conjecture that in certain settings one could be interested in the *dual* of these constructs: for example, a ML model which is more apt at identifying *simple* and *routine* cases could be useful to partially automate the workload of expert clinicians and allow them to better focus on more complex and multi-faceted cases. Conversely, in primary healthcare settings, a ML model that is better at difficult cases could be more useful in that it better complements the general skills of the doctors therein employed. In any case, in regard to the mathematical formulation of $wU$, the relevance factor is an agnostic factor that we introduced with the simple aim of capturing the central notion of *relative importance*.

As a consequence of this increased flexibility, we can easily notice in our exemplificatory study that in no case (i.e for no risk threshold in the definition of the standardized Net Benefit, and for no *definition* of the $\tau$ in the $wU$) the $wU$ and the Net Benefit were exactly the same. However, the Net Benefit at $\tau = 0.5$, and the $wU$ based on either $\tau_{persuasion}$ or $\tau_{auto-bias}$ were quite similar (i.e. 0.85 vs 0.90 and 0.89, respectively): this can be explained by noting that both $\tau_{persuasion}$ and $\tau_{auto-bias}$ had an average value close to 0.5 (albeit with a relatively large IQR), and the case complexity was relatively stable across the dataset. On the other hand, even though the average value of $\tau_{confidence}$ was close to 0.75, its value was noticeably smaller than the Net Benefit at $\tau = 0.75$ (0.45 vs 0.59). This difference can be explained by noting that there was a significant difference in the cases associated with an high probability score (i.e. $h(x_i) \geq 0.75$) and a correct (resp. wrong) diagnosis provided by the ML model, both in terms of case complexity ($0.75 \pm 0.02$ vs $0.69 \pm 0.00$, respectively) and value of $\tau_{confidence}$ ($0.85 \pm 0.02$ vs $0.93 \pm 0.01$, respectively): we recall that, in particular, the value of $\tau_{confidence}$ directly impacts the penalty assigned to False Positive cases.

*Comparing wU and Error Rate-based Metrics.* Similar comments can be made with respect to the comparison between $wU$ and the considered error rate-based metrics: indeed, also the Sensitivity (resp. Specificity) can be easily understood as a special case of $wU$ (resp., $wU$ on the dual dataset, in which positive and negative labels are interchanged), when we set the relevance of the negative (resp. positive) cases to 0: since the distribution of relevance was significantly skewed towards the positive class ($0.75 \pm 0.01$ vs $0.66 \pm 0.01$), both $\tau_{persuasion}$ and $\tau_{auto-bias}$ were slightly skewed toward Sensitivity rather than Specificity. Interestingly, we also proved that other error-rate based metrics (including the Balanced Accuracy, the Markedness, the Youden Index and, more in general, all metrics that can be described in terms of the confusion matrix, such as the F-score or the Matthews Correlation Coefficient). In addition to being capable to capture the same information (by combining the $wU$ and its dual), the $wU$ can also represent the appropriateness of the decision model to the risk profile used to determine the costs and benefits of the clinical intervention, thus providing a clear indication of the accuracy of the ML model in terms of *percentage of correct predictions completed with a confidence appropriate to the risk threshold required by the clinical intervention protocol and the specific characteristics of the case at hand*. The flexibility of $wU$ provides decision makers with a straightforward and comprehensive indicator of the whole clinical process that can be further analyzed, identifying the impact of the decision model and of the clinical decision strategy.

*Future works.* In light of the relationship between the $wU$ and other error rate-based metrics, we think that a first interesting future problem regards investigating the theoretical properties of $wU$ as an evaluation metrics, e.g. with respect to the mathematical framework proposed in [21]. The investigation of such empirical and theoretical properties, in turn, can be important for the development of regulatory standards with the goal of establishing the appropriateness of different metrics as tools to evaluate and validate ML model for use in the real world: indeed, despite the abundance of metrics, all aimed at quantifiying "how accurate the predictions are of a predictive model", reaching consensus on what measure should be used, even in a specific application domain like medicine, has been so far an oft neglected objective. Indeed, even recent recommendations developed for the reporting of prediction models, like TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis [11]) or MINIMAR (MINimum Information for Medical AI Reporting [17]), do not generally recommend any specific measure, despite their comprehensiveness. With this respect, and focusing in particular on the usage of evaluation metrics as tools for the *certification* of ML models, further research should also be devoted at establishing appropriate threshold values for claiming *validity* (so-called *minimum acceptable accuracy* [7])

Finally, let us consider other possible uses of our proposed metric: we notice that, even though in this article we mainly focused on metrics as tools to *evaluate* ML models (so as to validate them and

understand their suitability in real-world settings), also other uses are possible. More precisely, performance metrics can be seen as tools to drive either model training or model selection (equivalently, hyper-parameter optimization). In regard to the latter use case, it is largely known that ML models typically depend on "tunable" parameters (usually called hyper-parameters) that are not directly learned during model training, but optimized through computational procedures (e.g. grid or stochastic search) that automatically select the configuration that optimizes the value of a given metrics [13]. As widely known, employing error rate-based metrics to this aim is common (and largely undisputed) practice. However, one could envision the application of utility-based metrics instead, so as to represent more clearly the costs and benefits involved in the application of the ML model to be trained [10]: with this respect, the use of $wU$ would allow to more naturally capture the characteristics of the considered cases and the perceptions of the involved actors. By contrast, in the case of model training, it is noteworthy that neither error rate-based nor utility-based metrics are typically used as an optimization target. In their place, so-called *surrogate* metrics are typically used: these refer to approximations of a given target metrics that exhibit properties (e.g. differentiability, convexity, smoothness) that make them more amenable to optimization through standard black-box algorithms. Therefore, we believe that further research should be aimed at the development of appropriate surrogates for $wU$.

## 5 CONCLUSIONS

In this paper, we introduced a novel utility metrics, called *weighted Utility* ($wU$) and discussed its relationships with other existing metrics. The potentiality of the metric was demonstrated proving that it generalizes state-of-the-art metrics like the *Net Benefit* and *Standardized Net Benefit*. The $wU$ metrics allows the description of the same information provided by the above metrics, *but* it is also informed by additional information of the whole clinical process, including information about the individual cases and the perceptions of raters involved in the annotation and decision making process, when compared with other existing metrics. We believe this makes $wU$ measures more indicative of the real usefulness of a classification model when it comes to considering the skills and expectations of the intended users and the kind of decisions these are called to make. Further research is needed to validate this claim and research direction.

## REFERENCES

[1] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. 2018. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine* 1, 1 (2018), 39.

[2] S. G. Baker, N. R. Cook, A. Vickers, and B. S. Kramer. 2009. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society* 172, 4 (2009), 729–748. https://doi.org/10.1111/j.1467-985X.2009.00592.x

[3] E.S. Berner. 2003. Diagnostic Decision Support Systems: How to Determine the Gold Standard? *Journal of the American Medical Informatics Association* 10, 6 (2003), 608–610.

[4] William M Briggs and Russell Zaretzki. 2008. The skill plot: a graphical technique for evaluating continuous diagnostic tests. *Biometrics* 64, 1 (2008), 250–256.

[5] Federico Cabitza, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, Francesco Di Pietto, Angelo Gambino, Salvatore Gitto, et al. 2020. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences* 10, 11 (2020), 4014.

[6] Federico Cabitza, Andrea Campagner, and Clara Balsano. 2020. Bridging the "last mile" gap between AI implementation and operation:"data awareness" that matters. *Annals of translational medicine* 8, 7 (2020).

[7] Federico Cabitza, Andrea Campagner, Francesco Del Zotti, Alice Ravizza, and Federico Sternini. 2020. All You Need is Higher Accuracy? On the Quest for Minimum Acceptable Accuracy for Medical Artificial Intelligence. In *E-Health 2020: Proceedings of the 14th Multi Conference on Computer Science and Information Systems*. IADIS.

[8] Federico Cabitza and Jean-David Zeitoun. 2019. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of translational medicine* 7, 8 (2019).

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

[10] Enrico Coiera. 2019. Assessing Technology Success and Failure Using Information Value Chain Theory. *Stud Health Technol Inform* 263 (2019), 35–48.

[11] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC medicine* 13, 1 (2015), 1.

[12] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[13] Matthias Feurer and Frank Hutter. 2019. Hyperparameter optimization. In *Automated Machine Learning*. Springer, Cham, 3–33.

[14] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[15] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[16] Steve Halligan, Douglas G Altman, and Susan Mallett. 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology* 25, 4 (2015), 932–939.

[17] Tina Hernandez-Boussard, Selen Bozkurt, John PA Ioannidis, and Nigam H Shah. 2020. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association* 27, 12 (2020), 2011–2015.

[18] Kathleen F Kerr, Marshall D Brown, Kehao Zhu, and Holly Janes. 2016. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology* 34, 21 (2016), 2534.

[19] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2019. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *arXiv preprint arXiv:1909.12475* (2019).

[20] Jiayi Shen, Casper JP Zhang, Bangsheng Jiang, Jiebin Chen, Jian Song, Zherui Liu, Zonglin He, Sum Yi Wong, Po-Han Fang, and Wai-Kit Ming. 2019. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR medical informatics* 7, 3 (2019), e10010.

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

[22] Federico Sternini, Alice Ravizza, and Federico Cabitza. 2020. How Accurate Do You Want It? Defining Minimum Required Accuracy for Medical Artificial Intelligence. In *E-Health 2020: Proceedings of the 14th Multi Conference on Computer Science and Information Systems*. IADIS.

[23] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. 2014. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PLOS ONE* 9, 1 (01 2014), 1–10. https://doi.org/10.1371/journal.pone.0084217

[24] Andrew J. Vickers and Elena B. Elkin. 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making* 26, 6 (2006), 565–574. https://doi.org/10.1177/0272989X06295361

[25] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. 2016. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj* 352 (2016).

[26] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, et al. 2019. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 68, 10 (2019), 1813–1819.

[27] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *Bmj* 364 (2019).