

Article

The Elephant in the Machine: Proposing a New Metric of Data Reliability and its Application to a Medical Case to Assess Classification Reliability

Federico Cabitza ^{1,*†}, Andrea Campagner ^{1,†}, Domenico Albano ^{2,3}, Alberto Aliprandi ⁴, Alberto Bruno ³, Vito Chianca ², Angelo Corazza ², Francesco Di Pietro ⁵, Angelo Gambino ², Salvatore Gitto ⁶, Carmelo Messina ^{2,6}, Davide Orlandi ⁷, Luigi Pedone ², Marcello Zappia ^{8,9} and Luca Maria Sconfienza ^{2,6}

¹ Department of Informatics, Sistemics and Communication (DISCo), University of Milano-Bicocca, 20126 Milano, Italy; a.campagner@campus.unimib.it

² IRCCS Istituto Ortopedico Galeazzi, 20161 Milano, Italy; domenico.albano@grupposandonato.it (D.A.); vitochianca@gmail.com (V.C.); angelo.corazza@grupposandonato.it (A.C.); angelogambino@ymail.com (A.G.); carmelo.messina@unimi.it (C.M.); luigi.pedone@grupposandonato.it (L.P.); luca.sconfienza@unimi.it (L.M.S.)

³ Department of Biomedicine, Neuroscience and Advanced Diagnostics (BIND), University of Palermo, 90133 Palermo, Italy; bruno-alberto@hotmail.it

⁴ Unit of Radiology, Clinical Institutes Zucchi, 20900 Monza, Italy; a_aliprandi@yahoo.it

⁵ Diagnostic Imaging Department, Pineta Grande Hospital, 81030 Castel Volturno, Italy; fdipietro@gmail.com

⁶ Department of Biomedical Sciences for Health, Università degli Studi di Milano, 20122 Milano, Italy; Salvatore.gitto@unimi.it

⁷ Department of Radiology, Ospedale Evangelico Internazionale Genova, 16122 Genova, Italy; davide.orlandi@oeige.org

⁸ Department of Medicine and Health Sciences, University of Molise, 86100 Campobasso, Italy; marcello.zappia@unimol.it

⁹ Varelli Institute, 80126 Naples, Italy

* Correspondence: federico.cabitza@unimib.it

† These authors contributed equally to this work.

Received: 29 April 2020; Accepted: 4 June 2020; Published: 10 June 2020



Abstract: In this paper, we present and discuss a novel reliability metric to quantify the extent a ground truth, generated in multi-rater settings, as a reliable basis for the training and validation of machine learning predictive models. To define this metric, three dimensions are taken into account: agreement (that is, how much a group of raters mutually agree on a single case); confidence (that is, how much a rater is certain of each rating expressed); and competence (that is, how accurate a rater is). Therefore, this metric produces a reliability score weighted for the raters' confidence and competence, but it only requires the former information to be actually collected, as the latter can be obtained by the ratings themselves, if no further information is available. We found that our proposal was both more conservative and robust to known paradoxes than other existing agreement measures, by virtue of a more articulated notion of the agreement due to chance, which was based on an empirical estimation of the reliability of the single raters involved. We discuss the above metric within a realistic annotation task that involved 13 expert radiologists in labeling the MRNet dataset. We also provide a nomogram by which to assess the actual accuracy of a classification model, given the reliability of its ground truth. In this respect, we also make the point that theoretical estimates of model performance are consistently overestimated if ground truth reliability is not properly taken into account.

Keywords: inter-rater agreement; reliability; ground truth; machine learning; MRNet; knee; magnetic resonance imaging

1. Introduction

The research purpose of this paper is to shed light on the concept of the reliability of the decision supports that are developed by means of supervised techniques of Machine Learning (ML). In particular, our approach focuses on the reliability of the ground truth that is generated in multi-rater settings and used to train and validate such ML models. In this paper, starting from Section 2, we will also discuss the relationship between the reliability of the ground truth and the reliability of the resulting decision support.

In the ML literature, the ground truth is usually assumed to be 100% accurate, and as such, the ML predictive support's reliability is evaluated as its capability to reproduce these annotations with no error (i.e., through their accuracy). Nonetheless, it is well known that medical ground truths are far from perfect: medical experts involved in the annotation task are clearly not infallible (recent estimates of the diagnostic accuracy of model experts may range between 80% and 90% [1,2]); a large inter-rater disagreement among the experts involved is usually observed (e.g., an inter-rater disagreement of around 50% was reported in [3] on an X-ray-based diagnosis task); and in general, medical experts may have varying degrees of confidence in the annotations they produce (although this is seldom checked, if ever). Further, it has recently been highlighted that the data quality issues affecting the ground truth may severely impact the reliability of the ML predictive models that are trained and validated on them [4–6], likely making each estimate of their reliability optimistic.

While in the technical literature, the concept of reliability in multi-rater settings has usually been equated to inter-rater agreement [7], in this paper, we argue that agreement is but one component of reliability: indeed, as highlighted previously, reliability represents a multi-dimensional construct comprised of both confidence, which regards the extent the involved raters are certain of the expressed ratings, and competence, which regards the accuracy of the involved ratings. For these reasons, in this article, we focus on the study of how to assess the reliability of ground truth datasets in multi-rater settings by proposing a novel reliability metric that takes into account three dimensions: agreement (that is, how much a group of raters mutually agree on a single case); confidence (that is, how much a rater is certain of each rating expressed); and competence (that is, how accurate a rater is).

The rest of this article is structured as follows:

- In Section 2, we discuss the concept of reliability in decision support, and we will show, also in visual form, the relationship between this concept and the degree of agreement between raters, the quality of the ground truth in multi-rater settings, and the accuracy of the resulting models. To this aim, we will outline the main approaches proposed in the literature to measure the concept of reliability, and what are their main shortcomings that we aim to overcome with our proposal;
- In Section 3, we introduce and discuss the proposed weighted reliability scores; to this aim, we will present the underlying theoretical framework and its analytical derivations. Further, in Section 3.2, we discuss the main so-called paradoxes of reliability, which are intuitive properties that a measure of ground truth reliability should satisfy, but are violated by the most common and frequently adopted measures, and we will show that our metric is resistant to these paradoxes. Finally, in Section 3.3, we describe the design of a user study we performed to provide a proof of concept of our metric and to illustrate the main differences between the proposed solution and commonly adopted reliability metrics;
- In Section 4, we report on the results of the above user study, while in Section 5, we discuss these results and, more in general, how the proposed reliability metric can be used for, and put in relation to, the assessment of the reliability of an ML decision support;
- Finally, in Section 6, we summarize the main contributions of this paper and describe the motivations for further research.

2. Background and Motivations

Currently, the application of Machine Learning (ML) is recognized as a potential game changer in medicine [8]. Most of the systems that recently achieved—and even overcame—human performance in discriminative tasks, such as diagnosis and prognosis [9], are based on the methods and techniques known as supervised ML, in that the resulting models are optimized to discriminate among predefined conditions (i.e., classes). In supervised ML, by definition, human involvement is necessary to build the ground truth [10–12], that is the reference data on which the above discriminative models are trained. In ML, ground truth data are considered 100% accurate by definition. However, recent research has pointed out that a 100% accurate ground truth is possible only in fictional settings: indeed, the works in [5,13] highlighted a theoretical relationship existing between the number of involved experts and the accuracy of the resulting ground truth, whose analytical form is graphically depicted in Figure 1.

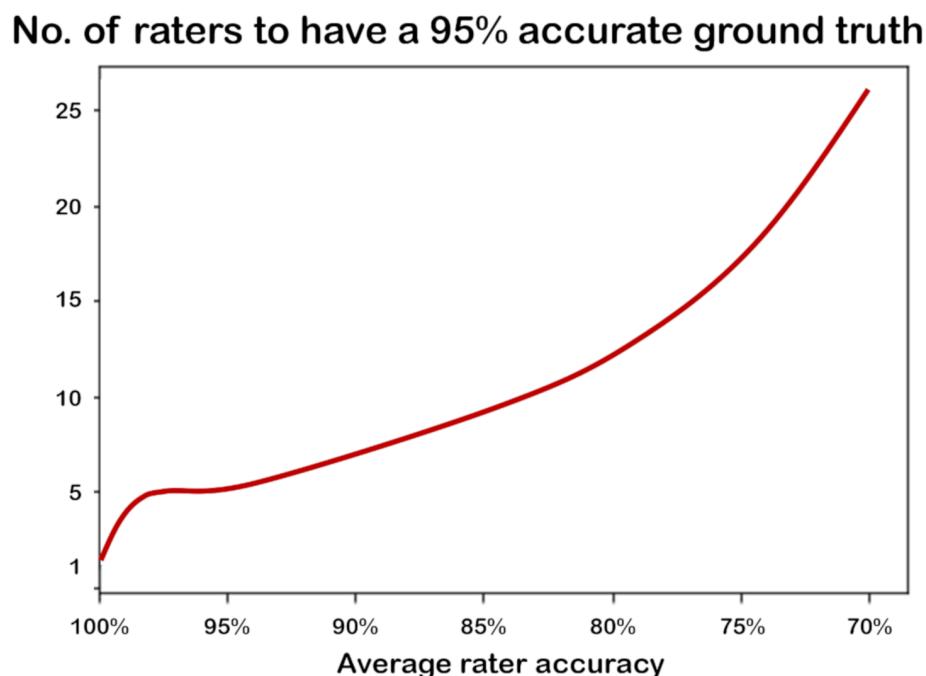


Figure 1. The figure depicts (on the *y*-axis) the number of raters that need to be involved to obtain a 95% accurate ground truth, as a function of the average accuracy of the raters involved (on the *x*-axis), if known. These estimates are obtained analytically and hence have general application.

As shown in the figure, a single infallible expert would be sufficient to obtain a 100% accurate ground truth. Unfortunately, human experts are all fallible, to some extent: in diagnostic tasks, the average accuracy of medical experts ranges from 80% to 90% [1,2], whereas the “average error rate among radiologists is around 30%” [14], although these figures can greatly vary, according to the specialty, exam modality, settings, and many other factors [15]. Therefore, if we take these estimates at face value, a simulation based on these estimates (see Figure 2) shows the intuitive notion that the more experts are involved in labeling a set of cases, the more accurate the resulting ground truth, when this latter is derived by applying the method of majority (or plurality) voting, that is simply assigning to each case the label that has been selected by the majority of the involved experts.

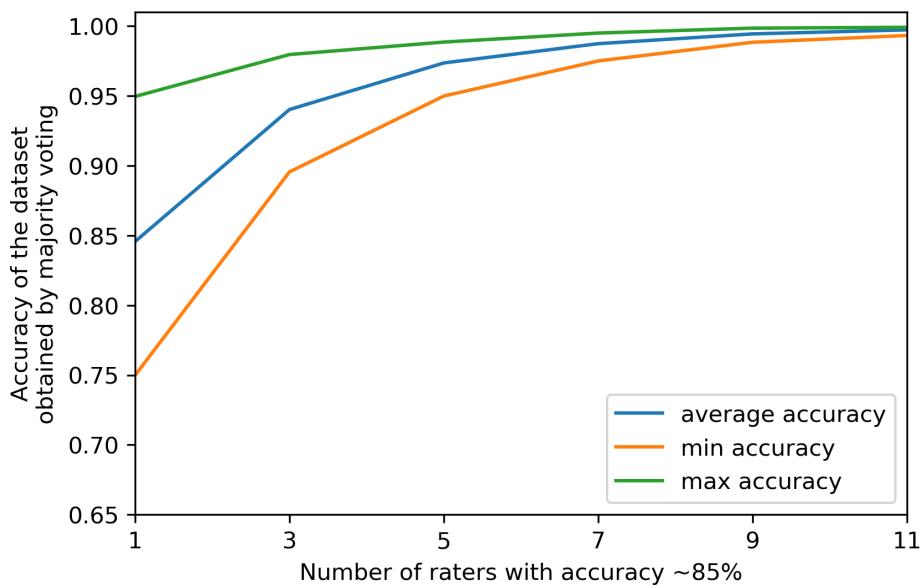


Figure 2. The figure depicts the average, minimum, and maximum accuracy of the datasets obtainable for a given number of raters, each with an average accuracy of $85\% \pm 1\%$, by simple majority voting among the raters for each case. The three estimates of accuracy were obtained by an analytical simulation, computed through random sampling from a simulated population of 100 individuals with the above-mentioned characteristics and by computing, respectively, the average, minimum, and maximum observed value among the extracted samples.

Looking at Figures 1 and 2 stimulates some reflections. First, the bad news is that normal (i.e., averagely accurate) raters will never yield a 100% accurate ground truth, although such a ground truth is assumed in all and every ML project. However, it is true that unity is strength (see also Figure 3): the smallest team of annotators to let a majority emerge (i.e., three raters) increases its collective accuracy by approximately eight percentage points with respect to the average accuracy of the single members of the team; likewise, a relatively small number of raters (like seven for binary tasks) is sufficient to achieve the high accuracy (although not perfect) of the ground truth. However, any estimation depicted in Figure 2 assumes to know the accuracy (that is, the competence) of each rater, or at least their average as a collective, and furthermore, that the errors that the raters make are non-correlated. This is seldom the case in general, and obviously impossible for the specific cases to label.

Furthermore, majority voting is not panacea: there will be times in which more raters are wrong than those who are right, thus leading to selecting the wrong label for a specific case. For instance, if the ML developers involve less than 10 raters (as is often the case), the number of times consensus is established on the basis of narrow or slim majorities (i.e., by only one rater) is higher than the opposite case (e.g., 55%, 63%, or 75% of the times, if the raters are 7, 5, or 3, respectively) (furthermore, for this reason, alternative methods to derive the ground truth from multi-rater labeling may be more powerful [5] than majority voting, although they are very seldom applied).

Thus, it is intuitive to understand that the higher the agreement, whatever it is, the “sounder” the majority and the more reliable the dataset. The easiest way to represent agreement is in terms of the percent of agreement (P_o), that is the ratio between the number of times the raters agree and the total number of joint ratings (see Figure 4). However, as is well known [7,16,17], this rate overestimates agreement (and hence reliability) in that (intuitively) raters can agree by chance, and not because they really agree on how to label a specific case. We call this latter case genuine agreement. To take chance into consideration, and discount it from agreement estimation, several metrics have been proposed [18] to assess the reliability of the data produced when at least one rater involved disagrees with the others in regard to some rating: here, we can mention the Pearson r , the Spearman Rho,

the intra-class correlation coefficient, the concordance correlation coefficient, and the most commonly adopted ones, that is Cohen's kappa (for two raters), the Fleiss kappa (for three or more raters) [19], and Krippendorff's α [20]. Although these metrics adopt a model of chance to account for the effect on agreement, they present some limitations, like restrictions on the data type they can be applied to, their ability to work with missing ratings, or their being undefined for single cases. Moreover, they are not free from known paradoxes, which have been discussed in some specialist works [7,21–25], among which we will discuss the main ones in Section 3.2 to show how our proposal is less prone to these shortcomings. Traditionally, inter-rater agreement scores are considered good proxy indicators of the reliability of data (e.g., [26]). However, as we argue in the following section, this may clearly be seen as a simplistic view that does not account for the multi-faceted nature of reliability.

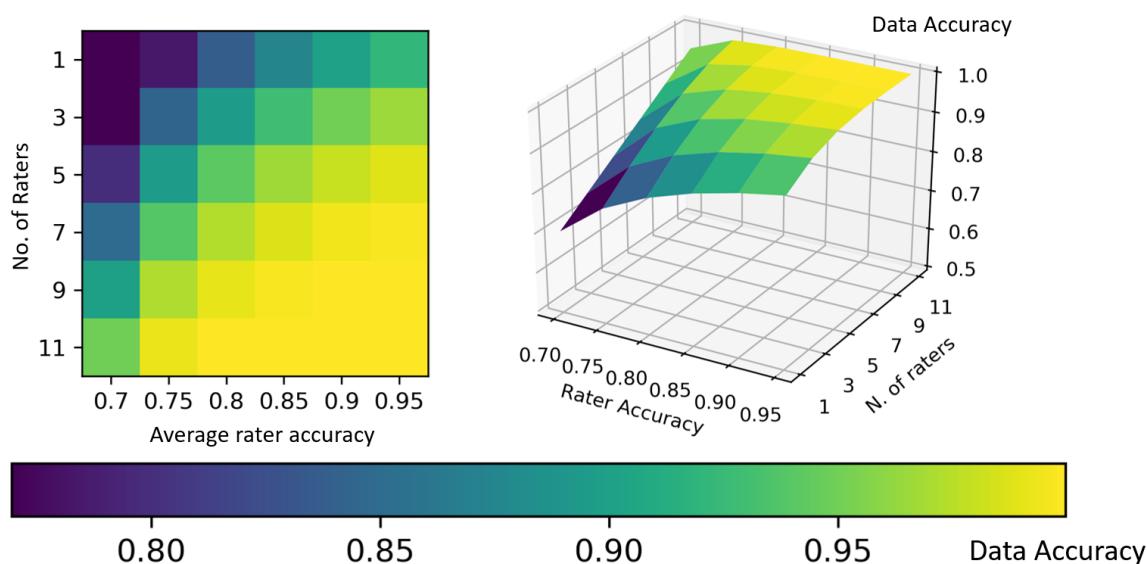


Figure 3. Graphical representation of the relationship between the number of raters involved in any process of ground truthing, their accuracy, and the accuracy of the resulting dataset (by majority voting). Furthermore, these estimates, as in the previous figures, are obtained analytically and hence have general application.

Reliability in Decision Support

In this paper, we focus on the concept of the reliability of decision supports that are developed by means of supervised techniques of Machine Learning (ML). As is widely known, this class of decision supports is not grounded on rules that are extracted from how human experts act and make decisions, or from directly asking them about their reasoning and decision criteria; rather, these systems ground their advice on data patterns and correlations that specific algorithms can detect among the features of large datasets and reproduce when fed with new data points. This characteristic motivates a first intuition, which we will investigate in our study: that the classification reliability of these systems (i.e., whether they are good classifiers or not) is related to the reliability of the data these systems have been trained on, i.e., their ground truth, to approximate and reproduce the patterns found therein (see Figure 4). A first consequence that follows from this intuition is that the actual accuracy of ML models is different from the “theoretical” accuracy that is evaluated by assuming the ground truth to be 100% accurate (right side of Figure 4), and that is generally considered the only accuracy we should care about and report in our scientific articles.

For instance, what would it mean that an ML model has been trained on a 85% accurate ground truth? If the model performance was reported to be 98% accurate, that is a very good performance (close to perfection in some respect), its actual accuracy would be 83% instead, which would be a much lower performance: in particular, this means that the ML model, instead of being almost perfect, actually makes mistakes with the same frequency with which a six can come out by rolling a dice.

For the above reasons, in this paper, we will pursue the research goal to assess how “reliable” the ground truth is in ML settings, because this reliability does affect what we are more interested in, that is the reliability of the decision support. This is a challenging task because this assessment must be obviously made in the absence of further reference data acting as a gold standard.

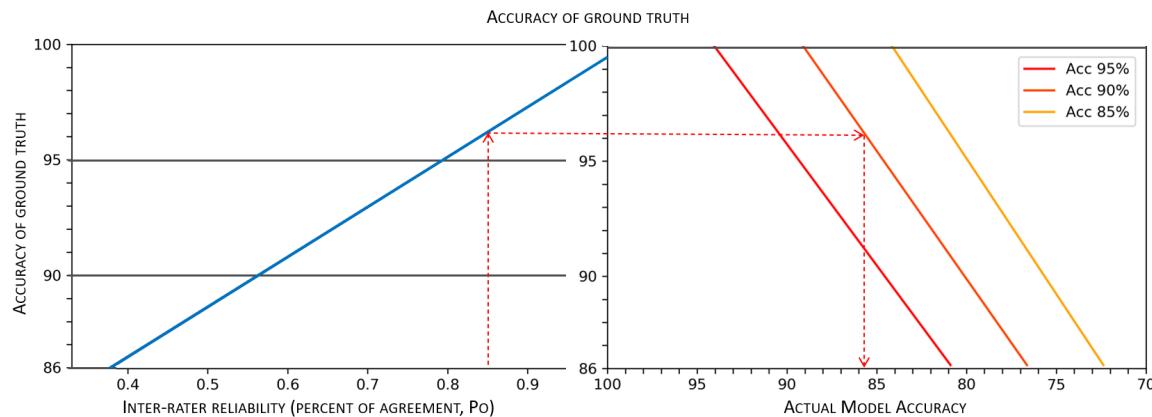


Figure 4. Representation of the general relationship between inter-rater agreement (measured as P_o), the accuracy of the ground truth, and the actual accuracy of an ML model trained on that ground truth. The estimates are obtained analytically and hence have general application. For this reason, the diagram can be used as a sort of a nomogram: given the level of agreement achieved to produce a ground truth and the accuracy of an ML model trained on this ground truth, the diagram can be used to obtain an estimate of the actual accuracy of the model.

Thus, this work is aimed at filling the gap in the literature to bridge the general notion of reliability to a more technical and specific notion in the decision support domain, a notion that directly combines the (theoretical) accuracy of ML models and the reliability of their training data (or ground truth), to get what we denote as actual accuracy (see Figures 4 and 5).

The concepts implied by the term reliability are wide and various enough to lay this latter open to multiple interpretations and, what is worse yet and is our concern, to a less than comprehensive and sound evaluation. As the common sense suggests, reliability regards the extent “something” is good enough to be relied upon. Hence, this term is often assimilated as trustworthiness and accuracy. Keeping aside the former concept, which also includes the users’ attitudes and perceptions, the latter concept can be applied to both machines (including computational decision aids) and measurements (that is, data). In the former case, a reliable predictive system is capable of giving the right answer (it is accurate); in the case of reliable data, these are such if they are an accurate representation of the reality of interest. However, both predictive systems and data present the same shortcoming: we cannot say whether the answer of a predictive system is the right one (otherwise, we would not need it), nor can we be sure our representation is true, unless compared with a true representation (which would be our data if it existed). Thus, while for predictive systems, we usually settle for a probabilistic estimation of their future accuracy, for data, we have to consider the second component of their accuracy (besides trueness), that is the precision of the process that generated them (not to cause confusion to the reader, we make it clear that in the social sciences, the term precision is related to the “resolution” of data [27]). This concept is usually equated to the replicability of measurement and to the mutual closeness of indications that come from multiple observations of the same object or phenomenon of interest [28]. Therefore, in any real-world multi-rater setting, the accuracy, and hence the reliability, of data can then be traced back to the agreement among the raters involved in producing them.

However, the main motivation for proposing a new reliability metric based on agreement evaluation lies in the recognition that agreement is but one component of reliability. In this paper, we make the point that reliability represents a multi-dimensional construct comprised of

both confidence, which regards the extent the involved raters are certain of the expressed ratings, and competence, which regards the accuracy of the involved ratings.

Our main assumption is that a sound reliability measure should not only value agreement that is purely due to chance less than genuine agreement, which is what reliability measures typically do, but should also value agreement on the wrong labels (also in those cases when the label trueness is not known *a priori*) less than agreement on the right labels: to make this aspect clear, it suffices to notice that a ground truth built by two raters who are always in mutual agreement, but systematically propose the wrong interpretation of the phenomenon under consideration should not be considered reliable at all. However, none of the proposed measures takes into account this aspect, only considering the agreement component of reliability [7].

These considerations bring us to consider three aspects of reliability which are seldom considered and will be addressed in Section 3:

- How to assess the extent the raters involved genuinely assert one specific rating instead of another, beyond self-assessment or naive chance models;
- How to take into consideration properly, in the definition of a measure of reliability, not only the mutual agreement of the involved raters, but also their competence in the specific task of annotation and the confidence they attach to their ratings;
- Lastly, how to combine these two components into a single coherent measure of the reliability of a set of annotated cases.

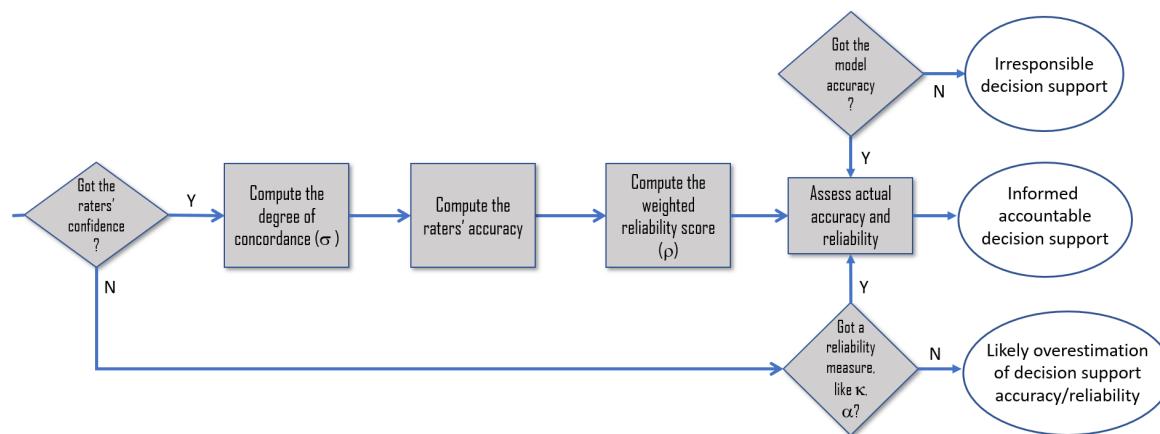


Figure 5. The general procedure proposed to get an estimate of the actual reliability of a classification model.

3. Method

As stated in Section 1, we consider the raters' reliability as a composite construct comprising the inter-rater agreement, the raters' confidence (i.e., the self-perceived ability to make a convincing decision for a specific case, or to avoid guessing), and their competence (i.e., the ability to make the right decision, in general).

In what follows, we will consider a rating an association between a case and a label (equivalently: the relationship between an instance and its target class): that is, a single labeling. Intuitively (see Figure 6, an agreement occurs when two raters assign the same label to the same case (or object, instance, phenomenon, etc.). We do not know whether a rating is right (i.e., including the true, but unknown label), but we assume that the more agreements (which, as mentioned above, are detected by the simple identity of responses given by a pair of raters) a rating gets, the stronger the evidence of it being right, and hence its reliability. In the “race” to become “truth” (or better yet, part of the ground truth), if we adopt majority voting (which is a simple, but effective method), the rating with the higher number of agreements wins.

However, we should not consider all agreements equal. In the literature, it is common to consider that some agreements are due to chance. Usually, the effect of chance is discounted on the basis of a model that is derived from the whole distribution of ratings in the dataset [7]. We consider this approach a limitation of these models, in that metrics that are adjusted in this way cannot be applied to single cases, but only to whole datasets, and the bigger these are, the better. In order to overcome this common limitation, we developed an alternative metric that first defines a single-agreement reliability, then a case reliability (averaging over all of the agreements), and finally, a dataset reliability (averaging over all of the cases included in the dataset).

To define an agreement reliability, our method leverages the confidence of the raters involved, when we ask them to estimate the confidence of their ratings. If this confidence is high, we assume the raters have not guessed. If both raters (who agree with each other) are highly confident, we assume their agreement is genuine, that is not due to chance.

However, we must also take into consideration the competence of the raters, as a way to discern between agreements made on a correct label and agreements that are wrong.

To this aim, we consider the rater's skill. This can be estimated in many ways, both qualitatively and quantitatively. In this latter case, a proxy dimension of the raters' competence is their average accuracy. The average accuracy of each single rater can then be estimated on the basis of a specific evaluation test or, more conveniently, on the basis of the majority voting of the other raters involved.

With reference to Figure 6, our method to quantify the reliability can be described as going backwards from an observed rating about which two raters agree, i.e., an agreement: first, we evaluate whether, based on the raters' confidence for the given rating, the observed agreement is genuine or only due to chance; second, we consider whether the agreed-upon rating is right or not, based on the raters' competence with respect to the discriminative task at hand.

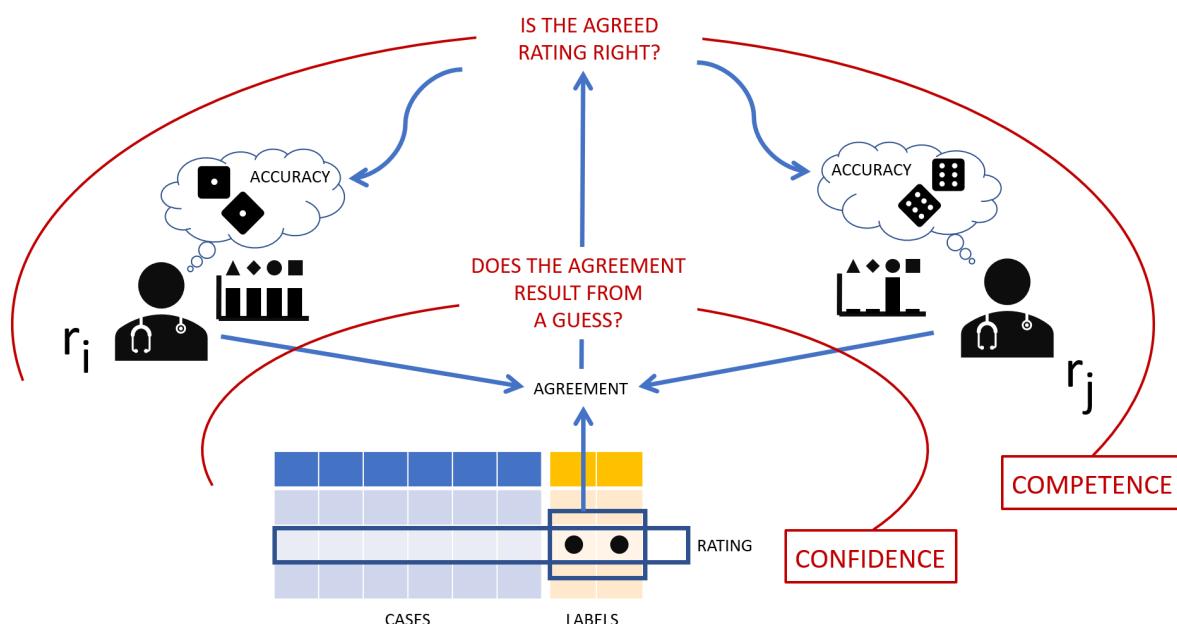


Figure 6. Graphical presentation of the general idea of the weighted reliability score, ρ . This reliability score is multi-dimensional as it encompasses both the confidence of the raters and their competence. In this graphical example, rater r_i is much less competent than rater r_j , and hence, the probability that her/his rating is correct is lower.

3.1. Derivation of the Weighted Reliability Score

We will assume the following setting: each rater in a group of k is asked to provide a label (from a given label set $C = \{1, \dots, n\}$, which we assume to be categorical) to each case x in a dataset S . We denote rater r_i 's labeling of case x as $r_i(x)$.

Furthermore, each rater r_i is also asked to report his/her confidence $c_i(x)$ in each label he/she gives. As previously stated, we consider reliability to be a composite construct, defined by the raters' confidence and competence.

As anticipated above, our method can be seen as going backwards from an observed agreement; thus, we first need to ask whether an observed agreement was genuine or due to chance. To this aim, we assume a two-step decision procedure, which can be seen as an instantiation of the standard decision-theoretic setting for decision under risk [29]: first, the decision maker flips a biased coin to choose between random choice (i.e., the random selection of a possible alternative according to an a priori probability distribution) with probability $1 - c_i(x)$ and peaked choice (i.e., the selection of an alternative according to an informed, case-dependent, distribution) with probability $c_i(x)$. Here, $c_i(x)$ is the confidence expressed by the rater in her/his decision $r_i(x)$, which we equate to the unconditional probability of the fact that the observed label $r_i(x)$ has been chosen genuinely and not simply by chance (i.e., $c_i(x) = P(r_i(x) \text{ genuinely})$).

Then:

- If rater r_i has “chosen” random choice, then we assume her/his label is selected according to a distribution $\mathbf{p}_{r_i}(x) = \langle p(1), \dots, p(n) \rangle$ where $p(i) = \frac{1}{n}$ (assuming a uniform distribution of the alternatives) or $p(i) = e(i)$ where $e(i)$ is an empirical prior estimate of the real prevalence of i (this can be derived from a ground truth labeling, if available, or considering all the labelings given by the multiple raters);
- If rater r_i has “chosen” peaked choice, then we assume her/his label is selected according to a distribution $\mathbf{d}_{r_i} = \langle d(1), \dots, d(n) \rangle$ where $\exists! i. d(i) = 1$.

The conditional estimate:

$$P(r_i(x) \text{ genuinely} | r_i(x))$$

can be obtained as:

$$\hat{c}_i(x) = \frac{c_i(x)}{c_i(x) + (1 - c_i(x))p(r_i(x))} \quad (1)$$

Thus, given two raters r_i, r_j who have been observed to agree on case x (i.e., $r_i(x) = r_j(x)$), we want to quantify the probability that the observed agreement is not due to chance; this means that both raters selected the peaked distribution, that is a Genuine Agreement (GA) occurred. Assuming that the two ratings were given independently, then the pairwise GA is defined as:

$$GA_x(r_i, r_j) = \begin{cases} 0 & r_i(x) \neq r_j(x) \\ \hat{c}_i(x)\hat{c}_j(x) & \text{otherwise} \end{cases} \quad (2)$$

We can now define the degree of concordance σ , as a measure of the proportion of genuine agreement (in either a single case or in a whole set of cases): this metric is simply defined as the average of GA both case-wise and sample-wise. First, we define the case-wise degree of concordance:

$$\sigma(x, R) = \binom{m}{2}^{-1} \sum_{r_i \neq r_j \in R} GA_x(r_i, r_j) \quad (3)$$

where the factor $\binom{m}{2}^{-1} = \frac{2(m-2)!}{m!}$ is the number of pairs of raters in R . We then define the degree of concordance as the average of the case-wise σ over the whole sample S , that is:

$$\sigma(S, R) = \frac{1}{|S|} \sum_{x \in S} \sigma(x, R) \quad (4)$$

Thus far, we only considered the agreement dimension of our definition of reliability; by following our backward reasoning (see Figure 6), we then discount the obtained degree of concordance with the raters' competence in the task at hand. To this aim, we need an estimate of the accuracy \hat{acc}_i of each rater i on each case x . If we were to know the correct label $cl(x)$ for each case x , then we would simply set $\hat{acc}_i(x) = \mathbb{1}_{r_i(x)=cl(x)}$. However, this is usually not possible, and we would need to resort to a probabilistic estimate of the rater's accuracy.

This latter can be set either on the basis of the empirical accuracy of the rater estimated on another dataset sample, which would however provide the same value \hat{acc}_i for each x , or on the basis of a model-based estimation, such as the Rasch model [30]. In this latter case, assuming that for each case x , an assessment of its complexity $diff(x)$ is available, the value of $\hat{acc}_i(x)$ can be estimated as follows:

$$\hat{acc}_i = \frac{e^{\exp_i - diff(x)}}{1 + e^{\exp_i - diff(x)}} \quad (5)$$

where \exp_i is an estimate of rater i 's ability or expertise.

Thus, having an estimate of each rater's accuracy on each case, we can quantify the probability that an observed agreement is right (i.e., equal to the true, unknown label) as:

$$P(r_i(x), r_j(x) \text{ correct}) = \frac{\hat{acc}_i * \hat{acc}_j}{\hat{acc}_i \hat{acc}_j + (1 - \hat{acc}_i)(1 - \hat{acc}_j)} \quad (6)$$

From the degree of concordance, we can then define a composite reliability measure, the so-called weighted reliability score, or ϱ , which also takes into account the accuracy of the raters and is defined at both the single case and whole dataset level:

$$\varrho(x, R) = \binom{m}{2}^{-1} \sum_{r_i \neq r_j \in R} GA_x(r_i, r_j) \cdot P(r_i(x), r_j(x) \text{ correct}) \quad (7)$$

$$\varrho(S, R) = \frac{1}{|S|} \sum_{x \in S} \varrho(x, R) \quad (8)$$

3.2. Paradoxes of Reliability Measures

Given the large number of proposed reliability measures, there has been an interest in proving one metric "better" than the others, in some respect. This has typically resulted in focusing on so-called paradoxes, i.e., statements asserting that a given class of reliability measures violates a set of intuitive properties. While we do not necessarily subscribe to viewing each and every violation as a paradox (e.g., see [25,31–33] for an account of this meta-perspective on so-called paradoxes of reliability), in what follows, we will show that the increased flexibility of the σ (resp. ρ) can be used to overcome (or simply shed light on) these properties. In particular, we will focus on two paradoxes that have been widely studied and discussed in the relevant literature [21,22,25,34] in regard to the most common and frequently adopted reliability measures, that is kappa and α .

Paradox 1. *High agreement, but low reliability: Let $C = \{0, 1\}$ be the set of possible classes. Suppose that all raters agree on assigning 0 to all cases but one, and on assigning 1 to the remaining case. Then, the observed agreement is perfect, but the value of kappa (or α) is ≤ 0 . If all the raters would have agreed on assigning the same class to all labels, then the value of kappa (or α) would be undefined.*

Paradox 2. In the case of two raters, when considering the distributions of labels for the raters, unbalanced distributions produce higher values of kappa (or α) than more balanced ones even when the observed agreement is the same.

We argue that both of these paradoxes arise from the interdependence in the definition of kappa and α between the model of the chance effects and the distribution of the labels among the raters. Thus, the paradoxical properties mentioned above are an intrinsic feature of any reliability score that does not separately models the self-agreement of the raters and their chance effects. Moreover, in regard to Paradox 2, we doubt this should be considered a negative feature: a sound reliability measure should exhibit this behavior whenever a model of the chance effects of the raters is not available: in the cases described by this paradox, if one of the two classes is much more represented in the dataset (and this mirrors the real population of interest), then the likelihood that the raters agreed just by chance is high, since they could have agreed just by selecting the most probable answer. In the extreme case where all the raters assigned all the cases to the same label, then the reliability of the dataset would be undefined as there is no way to understand if the observed agreement was genuine or due to chance. However, this observation does not apply to the weighted reliability score σ (nor to σ) due to the increasing flexibility of the chance model we adopted in its definition. In particular, the value of σ depends not only on the distribution of the classes in the dataset (which influences only the weighting factor \mathbf{p}_{r_i} in the definition of \hat{c}_i), but also on the confidence values c_i . Thus, formally:

Theorem 1. Let D be a dataset, and let $r_i(x)$ be the labeling given by rater i on case x , with C the set of class labels. Let $\forall i, x, j, y. r_i(x) = r_j(y) = c \in C$. Then, $\sigma > 0$, and further, σ depends only on $c_i(x)$ and $\mathbf{p}_{r_i}(x)$.

Proof. From the definition of σ in Equation (4), it is evident that $\sigma \geq 0$ and $\sigma = 0$ iff $\forall x \in D, i \in \{1, \dots, k\}. c_i(x) = 0$. Obviously, also, the definition of σ only depends on the stated factors. \square

Definition 1. We say that two probability distributions p_1, p_2 are marginal symmetric if there exists a permutation π s.t. $\pi(p_1), \pi(p_2)$ are both increasing, balanced if they are both the uniform distribution, and marginal asymmetric if there exists a permutation π s.t. $\pi(p_1), \pi(p_2)$ are inversely ordered.

Theorem 2. Let D_1, D_2, D_3 be three datasets, with the same observed agreement annotated by two raters r_1, r_2 . Let d_1^i (resp. d_2^i) be the distribution over the class labels given by rater r_1 (resp. r_2) for dataset D_i . Assume that d_1^1, d_2^1 are marginal symmetric, d_1^2, d_2^2 are balanced, and d_1^3, d_2^3 are marginally asymmetric. Furthermore, let \mathbf{e} be the same empirical distribution of the classes for each rater. Then, not necessarily, $\sigma(D_1) \leq \sigma(D_2) \leq \sigma(D_3)$.

Proof. By Definition 4 and Theorem 1, the value of σ only depends on $c_i(x)$ and $\mathbf{p}_{r_i}(x)$ (hence, in this case, on \mathbf{e}). Thus, it suffices to set the values accordingly for the three datasets. \square

3.3. User Study

In order to assess the feasibility and describe the usage of our metric, and more in general, describe the different dimensions that could affect the ground truth quality in a multi-rater setting, we designed a user study involving a team of domain radiology experts in a realistic task of data annotation for ground truthing purposes, namely the annotation of knee Magnetic Resonance Imaging (MRI). Specifically, we asked 13 fellowship-trained musculoskeletal radiologists in Italy to label 417 MRIs randomly extracted from the MRNet dataset [35] (<https://stanfordmlgroup.github.io/competitions/mrnet/>) so that the two classes were almost perfectly balanced, with normal MRIs accounting for 55% of the labeled samples and abnormal MRIs accounting for 45% of the labeled samples.

For each of these images, the doctors were asked to assess: the presence of abnormalities (therefore, it was a binary classification setting with $C = \{0, 1\}$) and their confidence in such annotations, on a 5-value ordinal scale. These annotations were also used to evaluate the discriminative competence

of the raters in terms of accuracy, by comparing their ratings with the MRNet ground truth taken as a reference.

4. Results

The average accuracy of the 13 radiologists involved was $\bar{acc} = 0.81 \pm 0.04$ (95% confidence interval). The radiologists' performance (in terms of the true positive rate, i.e., sensitivity, and the false positive rate) is shown in Figure 7a.

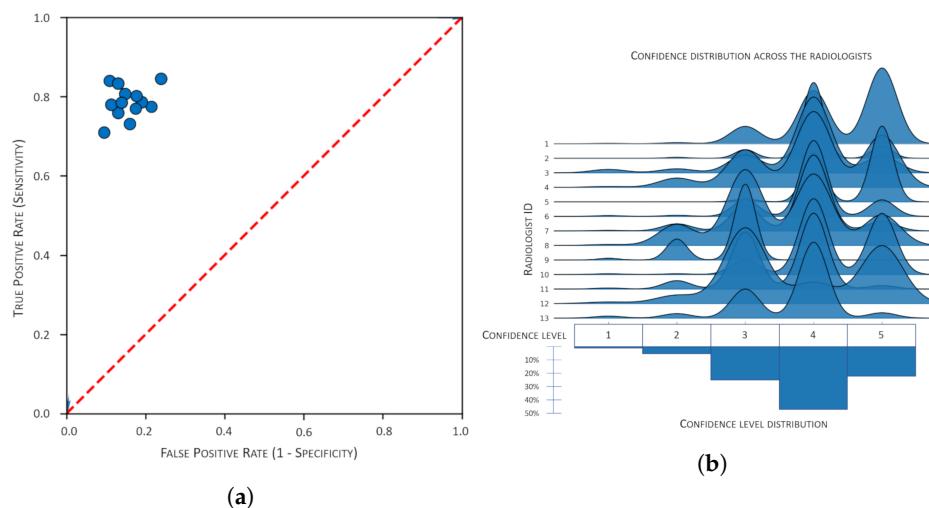


Figure 7. Results in graphical form. (a) The raters' performance in the ROC space: Circles represent single raters. The red line represents random guessing. (b) Joyplot and histogram illustrating the raters' confidence distributions.

The distribution of reported confidence for each of the radiologists is shown in Figure 7b, while Figure 8a,b depicts the relationship between the confidence expressed by the raters and, respectively, their accuracy and the difficulty of the cases they were supposed to rate.

As regards the confidence expressed by the involved medical experts, we note that the radiologists involved ranged from specialists with a few years of practice to senior practitioners with more than 20 years of experience and image reading; all of them were nevertheless involved in one of the most important orthopedic institutes in Italy. Therefore, it is little wonder that their confidence in the ratings was generally high, with four out of five the level most frequently mentioned (see Figure 7b). However, as a first observation, we noticed that being confident in a rating did not mean this was necessarily correct: Figure 8a shows this discrepancy, where 57% of wrong labels were nevertheless associated with the highest levels of confidence. This first reflection highlighted that confidence (intended as a subjective measure of the soundness of the proposed annotation) and accuracy were orthogonal dimensions that both impacted on the reliability of a multi-rater annotation.

As regards agreement, we observed that for 322 cases (75%), the radiologists expressed a statistically significant (or overwhelming) majority (which occurred whenever at least 10 of them agreed on a rating, out of 13). However, in regard to 61 cases (14%), the majority of the raters chose the wrong label (with respect to the MRNet ground truth). In those cases, majority voting would lead to an error in the derived ground truth. Even more surprisingly, in 41% of these latter cases (25, or 6% of the total), the wrong majority was overwhelming from the statistical point of view mentioned above. These observations showed that also agreement and accuracy were orthogonal dimensions providing different information about the reliability of the multi-rater ground truth.

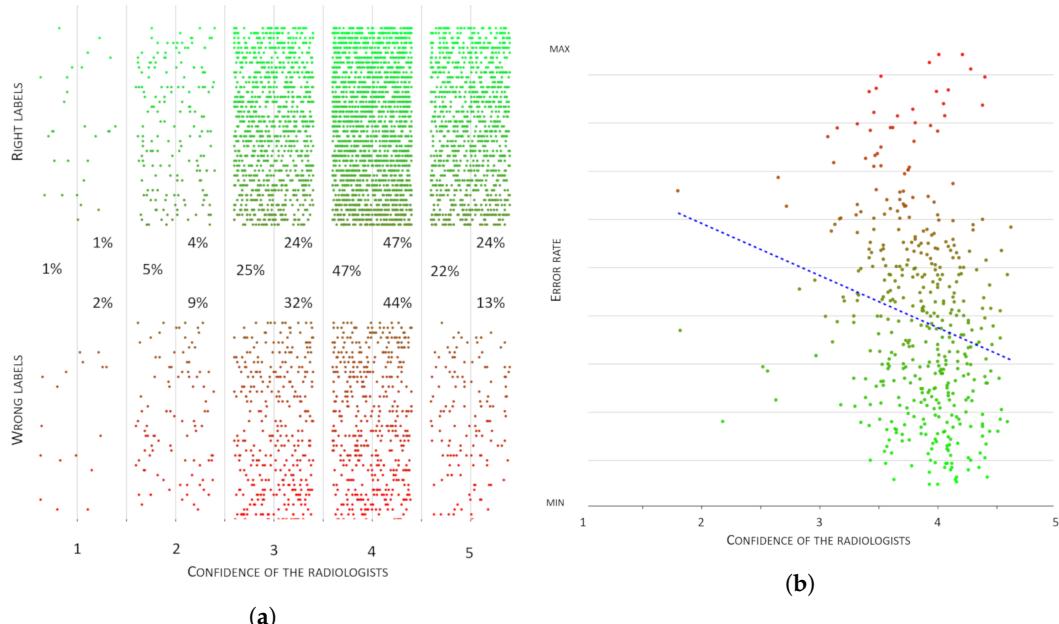


Figure 8. Results in graphical format: (a) The distribution of the labels in terms of right/wrong annotation, with respect to the confidence levels. For each confidence level, the percentages in the left column indicate the confidence level proportion with respect to all the cases; the top and bottom percentages in the right column indicate the proportions of right (resp. wrong) labels in the corresponding confidence level. (b) Relationship between the raters' confidence and the probability of error; the regression line suggests a mild inverse proportion.

The degree of concordance (σ) observed between the 13 radiologists was 0.62. Besides this degree, we computed two different weighted reliability scores, according to the method to assess the competence of the raters: when we considered the accuracy (with respect to the MRNet ground truth) of each rater, the resulting ϱ_a was 0.58; when we employed the Rasch model estimate (see Equation (5)), the resulting ϱ_R was 0.45. The values of Krippendorff's α and Fleiss' k were both equal to 0.63, with a percent of agreement P_o of 0.82.

5. Discussion

The user study reported in the previous sections confirmed the simulations presented in Section 1: teams of domain experts could mislabel cases, even when they were large enough to reach statistically significant consensus, and this was achieved by an overwhelming majority of raters, as it happened in approximately one case out of 20 in our study. This suggested that any metric, used for assessing the reliability of a given multi-rater ground truth, that did not take into account the accuracy of the involved raters was likely to provide an over-estimated and over-optimistic assessment.

On the other hand, as regards confidence, we could make two observations. First, as can be seen from Figure 8b, the raters' confidence had only a very weak correlation with the raters' error: indeed, as can be seen from Figure 8a, the two highest confidence levels accounted for over 50% of all the annotation errors. This suggested that confidence and accuracy indeed represented two orthogonal concepts. Second, we could observe that the values of α , k , and σ were relatively close ($\alpha = k = 0.65$, $\sigma = 0.62$): the small difference between the α and kappa measures could be explained by the fact that the class proportions in the dataset were almost perfectly evenly balanced, as both metrics considered class balance in order to model chance effects; in contrast, the small difference between the degree of concordance σ and the above-mentioned measures could be seen as a confirmation that leveraging the self-reported confidence in the definition of ϱ (through σ) was not only meaningful, as the confidence levels were (intuitively) inversely correlated with random guesses by the raters, but it also provided a practical and flexible model of chance effects that did not necessarily depend on the

whole dataset distribution (as is the case for other common reliability measures), but yielded equivalent results. Interestingly, making a naive estimate on the value of σ , based on the average confidence, we obtained (by multiplication with $P_o = 0.82$) an expected σ equal to 0.61, which was almost equivalent to the observed one. This suggested that the proposed σ could also be interpreted (provided that the distribution of confidence had small variance around its average) as a chance-corrected modification of the agreement P_o , in which the chance-correction was given by the average confidence among the raters.

On the other hand, we detected a large difference between the weighted reliability score ϱ and the reliability measures based only on an estimate of the chance-adjusted agreement (i.e., α and κ). This difference, as previously mentioned, could be explained by the fact that our score also took the competence of the raters into account and thus provided a more robust and comprehensive account of ground truth reliability. As we argued previously, this behavior is desirable as it allows assigning more weight to agreements that are likely to be true compared to agreements on possibly wrong labels. Importantly, the estimate we produced on the basis of the Rasch model showed that ϱ could provide a sound and conservative estimation of the ground truth quality even when a separate ground truth was not available, which is the most common case when the multi-rater setting is adopted (otherwise, it would be pointless to collect the multi-rater annotations).

Another indication of the robustness and flexibility of the proposed measures derived from the analytical results discussed in Section 3.2. These results shed light on the fact that the two above paradoxes were, as anticipated therein, intrinsic in any measure of reliability where the model of the chance effect was dependent on the distribution of the labels in the annotated dataset. On the other hand, the increased flexibility and more elaborate model of chance in the definition of the ϱ allowed “circumventing” the above paradoxes and made the proposed metric a sound measure of reliability.

In regard to the application of ϱ in real-world settings, in order to assess if a given ground truth is sufficiently reliable to support further analysis such as the training of predictive models, it would be natural to ask what threshold level should be used for such an assessment. This problem is called setting the so-called smallest acceptable reliability threshold [20]. Unfortunately, any proposal of such a threshold would be laden with some extent of arbitrariness. Nevertheless, Krippendorff suggested to not accept data with reliability estimates whose confidence interval (computed via bootstrapping or a permutation procedure) reaches below the smallest acceptable reliability set at 0.667 [20] (p. 242). Any reliability score below 0.667 would mean that only two thirds of the data are labeled to a degree better than chance. This recommendation, while still arbitrary to some degree, challenges a much more popular way to interpret agreement scores in vogue since the 1970s [36], which is much more indulgent (a score of 0.21 is considered an indicator of fair agreement, 0.41 moderate, and 0.61 substantial), and it offers a more robust and conservative procedure by which to evaluate reliability.

In Figure 9, we can see a nomogram, similar to that depicted in Figure 4, which is a two-dimensional diagram designed to show the mathematical association between measures of the weighted reliability of a multi-rater dataset, the accuracy of the related ground truth (obtained by majority voting), and the corresponding actual predictive accuracy of models, which in Figure 9 are associated with a “theoretical” accuracy of 95%, 90%, and 85%, respectively. This nomogram can be used for any value observed along these three dimensions (namely, data reliability, ground truth accuracy, and model accuracy), but as an example, we show the losses in accuracy associated with the minimum reliability threshold mentioned above (i.e., 0.67) for those models for which developers boast an accuracy of 95%, 90%, and 85%: the deviation for all these cases is approximately 6%, which is a margin that is much greater than what is usually tolerated to choose the best model after a cross-validation session.

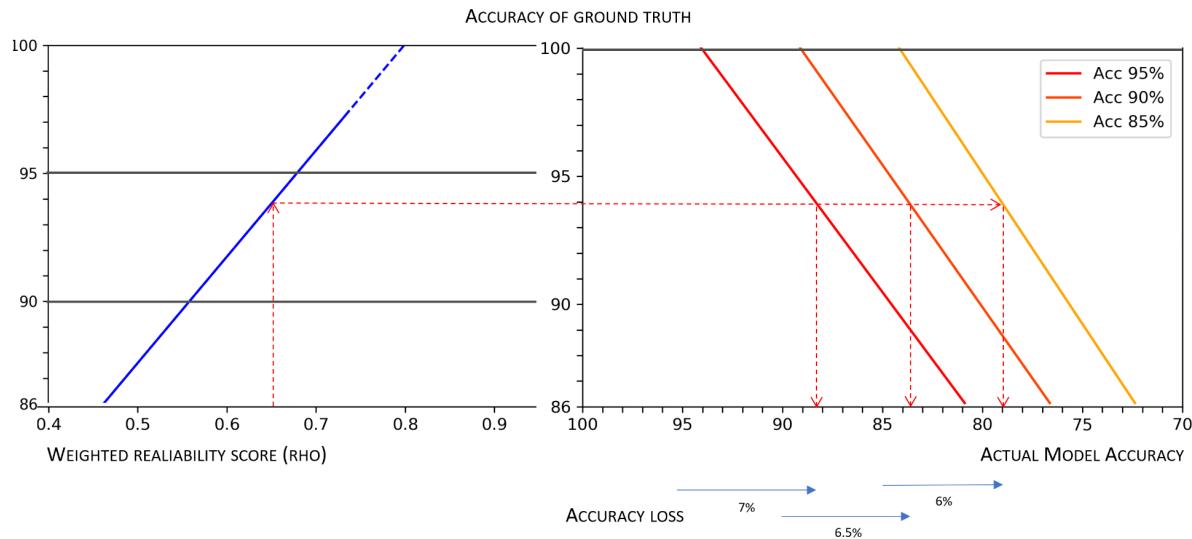


Figure 9. Representation of the relationship between the multi-rater reliability (measured as ρ), the accuracy of the ground truth, and the actual accuracy of an ML model trained on that ground truth. The figure can be used as a sort of nomogram: given the level of reliability for a given ground truth and the accuracy of an ML model on such a ground truth; the diagram can be used to obtain an estimate of the actual accuracy of the model.

The nomogram depicted in Figure 9 also allows a further, important, consideration. Estimates of ρ allow establishing a threshold for adequate reliability, which is based on a minimum acceptable accuracy value: having fixed a minimum acceptable value of actual accuracy acc (for a model whose measured level of accuracy is x), we should require ρ to be high enough to result in a ground truth quality g such that $x * g \geq acc$. As an example, let us assume that we need a predictive model exhibiting actual accuracy $\geq 90\%$, and further assume that, by training the model on a perfect ground truth, we would be able to obtain a model whose measured accuracy is equal to 95% . By observing the nomogram depicted in Figure 9, we can easily see that to achieve our requirement, we would need a ground truth at least 95% accurate. Thus, we set the minimum acceptable reliability at $\rho \sim 0.7$.

6. Conclusions

In this article, we introduced a new model to quantify the reliability of ground truth data in multi-rater settings. This model provided a more comprehensive and multi-dimensional definition of reliability, which took into account not only agreement, but also the confidence and competence of the involved raters: in very short terms, in our proposal, we related the reliability of the data to the reliability of those who produced the data. This model allowed us to define a novel reliability metric, which we called the weighted reliability score ρ ; therefore, is this yet another reliability metric?

To motivate this research endeavor, we made some consequent points: in order to have good data, it was necessary to involve more raters, not just one or few (see Figure 1). The quality of data affected the quality of the predictions of predictive models (see Figure 4) and of the decision making process itself, that is their reliability. Since this was what we really cared about in computer-based decision support, it was necessary to assess the reliability of the models' ground truth in a sound manner (see Figure 5): to this aim, we needed a metric that took into account the reliability of the raters involved and their ratings' (see Figure 6), as the weighted reliability ρ score did, and no other metric did, to our knowledge. Lastly, neglecting the assessment of data reliability in multi-rater settings meant to equate theoretical accuracy and actual accuracy and hence to be content with a harmful overestimation of the models' capability to truly support decision makers (see Figure 9).

In this paper, we also provided a formal proof that our metric was a sound measure and was not subject to common paradoxes that affect other common reliability measures. We also showed the

robustness and conservative nature of the proposed measure in the context of a realistic annotation task and illustrated how ϱ could be used to evaluate the reliability of a real-world ground truth in light of the required performance of a predictive model to be trained on it.

As we believe that reliability is a primary concern in any data analysis task, in what follows, we enumerate some still open issues that we believe should be further addressed and that motivate our future research agenda:

- We intend to extend the proposed model of reliability so that it can be applied also in the case of annotation tasks in which the target annotations are either numeric or ordinal values, as well as in the case of missing annotations and incomplete ratings, that is when, for a given case, one or more raters do not provide an annotation;
- We intend to further investigate the relationship between the ground truth reliability (as measured by the ϱ score) and the actual model accuracy to obtain more robust and precise estimates based on computational learning theory;
- As shown in [37], the accuracy of raters is also heavily affected by the complexity and difficulty of the cases considered (or similarly, by the proportion of really hard cases to interpret in the ground truth): difficulty can be another contextual (i.e., case-specific) factor in reducing the probability that a specific label is correct, both confidence and competence being equal. Thus, also this parameter should be collected from the team of raters involved, even if in a necessarily subjective and qualitative way, and be factored in the derivation of the weighted reliability score ϱ from the degree of concordance σ .
- We also intend to further enrich our ϱ metric, by considering not just the number of agreements and their reliability, taken individually, for each case, but also whether the number of agreements is in a scarce majority configuration (i.e., in one of those majorities where one rating could change the assigned label), with the assumption that these cases are intrinsically less “reliable” than the cases where disagreements do not affect the majority decision.

To conclude: In recent times and in an increasing number of application domains, classification models have been aimed at bringing (or could bring) a “significant effect” (cf. GDPR) to the life of human beings. Such domains are not limited to healthcare and medical applications, but encompass also other delicate domains like, e.g., employee selection, credit scoring, the assessment of the risk of recidivism, and more in general, any kind of profiling and automated decision making that can impact the health, reputation, or economic situation of the people involved. In all of these domains, reliability should be the most important aspect of those models to guarantee, as this is also the basis for the user trust in their role in decision making [38]. In this respect, assessing such a quality dimension—along with other common dimensions like accuracy, precision, and the like—and performing this assessment by means of a sound and conservative metric are crucial to improve the quality of the overall decision making process. We believe that such an assessment is also worthy of the collection of additional information from the experts involved in ground truthing (like the confidence in their ratings), in light of the impact that data reliability may have on decision making and on the accuracy (that is reliability) of the predictive models that may automate or support such decision making.

We discussed this impact and showed how it could have detrimental effects on the quality of the decision making if ground truth reliability is not taken into account and factored in, at least qualitatively: for its pervasiveness, and yet the relatively low debate about it, this problem looks like the notorious “elephant in the room” (hence, the title of this contribution) of decision support reliability assessment. For this reason, we assert that any responsible decision maker should always evaluate the reliability of the ground truth, to check the extent that these latter data can be rightly considered “the truth” and, hence, ultimately, improve the quality of the decision making process. The weighted reliability score is a tool to support such a responsible and accountable use of any classification models in settings where the life of human beings is at stake.

Author Contributions: Idea conception, F.C.; theoretical conceptualization, A.C. (Andrea Campagner); methodology, F.C. and A.C. (Andrea Campagner); software, A.C. (Andrea Campagner); formal analysis, A.C. (Andrea Campagner); data collection and curation, D.A., A.A., A.B., V.C., A.C. (Angelo Corazza), F.D.P., A.G., S.G., C.M., D.O., L.P., M.Z., and L.M.S.; writing and editing, F.C. and A.C. (Andrea Campagner). All authors read, reviewed and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GDPR	General Data Protection Regulation
IRCCS	Istituto di Ricovero e Cura a Carattere Scientifico
ML	Machine Learning

References

1. Qukel, L.G.; Kessels, A.G.; Goei, R.; van Engelshoven, J.M. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* **1999**, *115*, 720–724. [[CrossRef](#)] [[PubMed](#)]
2. Graber, M.L. The incidence of diagnostic error in medicine. *BMJ Qual. Saf.* **2013**, *22*, ii21–ii27. [[CrossRef](#)] [[PubMed](#)]
3. Jewett, M.A.; Bombardier, C.; Caron, D.; Ryan, M.R.; Gray, R.R.; Louis, E.L.S.; Witchell, S.J.; Kumra, S.; Psihramis, K.E. Potential for inter-observer and intra-observer variability in x-ray review to establish stone-free rates after lithotripsy. *J. Urol.* **1992**, *147*, 559–562. [[CrossRef](#)]
4. Cabitza, F.; Ciucci, D.; Rasolini, R. A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In *Organizing for the Digital World*; Springer: Cham, Switzerland, 2019; pp. 121–136.
5. Cabitza, F.; Campagner, A.; Ciucci, D. New Frontiers in Explainable AI: Understanding the GI to Interpret the GO. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 27–47.
6. Svensson, C.M.; Hübner, R.; Figge, M.T. Automated classification of circulating tumor cells and the impact of interobserver variability on classifier training and performance. *J. Immunol. Res.* **2015**, *2015*, 573165. [[CrossRef](#)] [[PubMed](#)]
7. Gwet, K.L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*; Advanced Analytics, LLC: Piedmont, CA, USA, 2014..
8. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [[CrossRef](#)] [[PubMed](#)]
9. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297. [[CrossRef](#)]
10. Beigman, E.; Klebanov, B.B. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1–Volume 1*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 280–287.
11. Beigman Klebanov, B.; Beigman, E. From annotator agreement to noise models. *Comput. Linguist.* **2009**, *35*, 495–503. [[CrossRef](#)]
12. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *New Engl. J. Med.* **2019**, *380*, 1347–1358. [[CrossRef](#)] [[PubMed](#)]
13. Heinecke, S.; Reyzin, L. Crowdsourced PAC learning under classification noise. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*; AAAI Press: Palo Alto, CA, USA, 2019; Volume 7, pp. 41–49.
14. Pinto, A.; Brunese, L. Spectrum of diagnostic errors in radiology. *World J. Radiol.* **2010**, *2*, 377. [[CrossRef](#)] [[PubMed](#)]
15. Brady, A.P. Error and discrepancy in radiology: Inevitable or avoidable? *Insights Imaging* **2017**, *8*, 171–182. [[CrossRef](#)] [[PubMed](#)]

16. Hripcsak, G.; Heitjan, D.F. Measuring agreement in medical informatics reliability studies. *J. Biomed. Infor.* **2002**, *35*, 99–110. [[CrossRef](#)]
17. Hunt, R.J. Percent agreement, Pearson’s correlation, and kappa as measures of inter-examiner reliability. *J. Dent. Res.* **1986**, *65*, 128–130. [[CrossRef](#)] [[PubMed](#)]
18. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica Biochem. Medica* **2012**, *22*, 276–282. [[CrossRef](#)]
19. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378. [[CrossRef](#)]
20. Krippendorff, K. *Content Analysis: An Introduction to its Methodology*; Sage Publications: Thousand Oaks, CA, USA, 2018.
21. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [[CrossRef](#)]
22. Cicchetti, D.V.; Feinstein, A.R. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 551–558. [[CrossRef](#)]
23. Hayes, A.F.; Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [[CrossRef](#)]
24. Powers, D.M. The problem with kappa. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Avignon, France, 23–27 April 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 345–355.
25. Zhao, X.; Feng, G.C.; Liu, J.S.; Deng, K. We agreed to measure agreement—Redefining reliability de-justifies Krippendorff’s alpha. *China Media Res.* **2018**, *14*, 1.
26. Duffy, L.; Gajree, S.; Langhorne, P.; Stott, D.J.; Quinn, T.J. Reliability (inter-rater agreement) of the Barthel Index for assessment of stroke survivors: Systematic review and meta-analysis. *Stroke* **2013**, *44*, 462–468. [[CrossRef](#)] [[PubMed](#)]
27. Brancati, D. *Social Scientific Research*; Sage: Thousand Oaks, CA, USA, 2018.
28. Costa Monteiro, E.; Mari, L. Preliminary notes on metrological reliability. In Proceedings of the 21st IMEKO World Congress on Measurement in Research and Industry, Prague Congress CentrePrague, Prague, Czech Republic, 30 August–4 September 2015..
29. Resnik, M.D. *Choices: An Introduction to Decision Theory*; Ned - New Edition; University of Minnesota Press: Minneapolis, MN, USA, 1987.
30. Rasch, G. *Probabilistic Models for some Intelligence and Attainment Tests 1960*; Danish Institute for Educational Research: Copenhagen, Denmark, 1980.
31. Charles Feng, G.; Zhao, X. Do not force agreement: A response to Krippendorff (2016). *Methodology* **2016**, *12*, 145–148. [[CrossRef](#)]
32. Krippendorff, K. Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Ann. Int. Commun. Assoc.* **2013**, *36*, 481–499. [[CrossRef](#)]
33. Krippendorff, K. Misunderstanding reliability. *Methodology* **2016**, *12*, 139–144. [[CrossRef](#)]
34. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [[CrossRef](#)] [[PubMed](#)]
35. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K.; et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med.* **2018**, *15*, e1002699. [[CrossRef](#)] [[PubMed](#)]
36. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometric* **1977**, *33*, 159–174. [[CrossRef](#)]
37. Campagner, A.; Sconfienza, L.; Cabitza, F. H-accuracy, an alternative metric to assess classification models in medicine. In *Digital Personalized Health and Medicine*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2020; Volume 270.
38. Cabitza, F.; Campagner, A.; Balsano, C. Bridging the “last mile” gap between AI implementation and operation: “data awareness” that matters. *Ann. Transl. Med.* **2020**, *8*, 501. [[CrossRef](#)]

