Dissimilar Similarities: Comparing Human and Statistical Similarity Evaluation in Medical AI

Federico Cabitza[1,2], Lorenzo Famiglini[1], Andrea Campagner[2], Luca Maria Sconfienza[2,3], Stefano Fusco[3], Valerio Caccavella[4], and Enrico Gallazzi[4]

[1] University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy
federico.cabitza@unimib.it
[2] IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milano
[3] Università degli Studi di Milano, Milano, Italy
[4] ASST Gaetano Pini - CTO, Milan, Italy

Abstract. This study explores the concept of similarity in machine learning (ML) and its congruence with human judgment in medical contexts, focusing primarily on radiology. We conducted a user study involving two radiologists and two orthopedic and spine surgeons. These experts evaluated the similarity of 72 cases, selected from a larger dataset by an ML model based on Cosine and Euclidean distances, in comparison to 18 representative base cases of vertebral fractures. Our analysis focused on correlating these ML-derived distances with the experts' assessments. The findings reveal that: (1) both Cosine and Euclidean distances had limited correlation with human judgments; (2) Cosine distances showed a marginally higher correlation than Euclidean distances; despite the limitations due to the small samples of evaluations and evaluators, our findings emphasize the necessity for ongoing research to enhance AI similarity metrics, aiming for greater human-centricity and relevance, particularly considering their critical role in ML training and inference. Our study's implications are far-reaching, advocating for a comprehensive reevaluation of similarity assessments in AI to achieve a closer alignment with human cognitive processes, extending well beyond the realm of medical imaging.

Keywords: Similarity, Machine Learning, Radiology, Orthopedics, User Study

1 Introduction

The aim of this study is to explore the concept of similarity in machine learning (ML) and its congruence with human judgment in decision-making scenarios. Indeed, in the application of ML to human decision-making, the concept of similarity is key: to give but two examples, several models rely on the definition of an underlying similarity measure [1], which would then have to conform with the human assessment of similarity to produce reliable and comprehensible support, while retrieval of similar cases has long been one of the most relevant forms of explainable AI [23]. Yet, the notion of similarity, and that its computational

evaluation could be useful to support human decision-making, grounds on common assumptions that are seldom verified or empirically tested: that in a certain domain, or for a specific task, all users (or models) have the same perception of the difficulty of cases that are to decide upon, or that they consider the same cases to be similar to each other in the same way. In other words, that similarity is a stable, objective, specific extension concept characterizing the relationship between two entities or phenomena.

In another study [2] we investigated the effectiveness (and perceived usefulness) of ML-based support systems that extracted similar cases with their respective ground-truth labels as decision support (for the task of spinal fracture diagnosis): one of our findings was that different users consider similar cases differently. Thus, in this article, we designed and executed a user study, in which we investigated the reliability of traditional computational methods to compute the similarity and the usefulness of measuring this information to provide more personalized support based on the decision maker's profile.

Thus, in our study, we investigate whether AI systems developed using machine learning techniques can extract similar cases from training cases that are indeed perceived as semantically similar by subject-matter experts, and, consequently, whether one similarity evaluation technique is to produce better results than another. In the following, we will therefore first frame the concept of similarity, both in human judgment and in algorithmic classification; then we will introduce the two similarity calculation techniques that we have compared; and finally, we will present the findings of the empirical user study, in which we collected the similarity ratings of four subject-matter experts, two board-certified radiologists, and two board-certified spine surgeons, to understand how the two above metrics correlate with the human ratings.

2 The importance of similarity in ML classification

Classification is an important part of human cognition in that by grouping similar objects into more general classes (categories), we organize and make sense of the world around us, identify patterns and relationships, and make predictions and decisions [26]. On the other hand, by recognizing and assessing the similarity between objects and concepts we perform important cognitive processes for our survival and success, such as analogical and metaphorical reasoning [12].

Both categorization (that is creating a set of categories) and classification (that is associating an instance to any of the predefined categories, taken as a class) ground on the concept of similarity, as we assume that instances said to belong to the same class, denoted with the same category, all share a set of common characteristics, which all make up a sort of "class-ness": such a relationship has indeed also been recognized in computational frameworks and theories aimed at formalizing the notion of categorization [18, 24].

Recently, much research has been devoted to computational methods to automate these two fundamental cognitive processes, and machine learning, as a general computational approach, encompasses many of these methods. The con-

cept of similarity is fundamental in machine learning research, where it is used in different ways and for different purposes depending on the type of task being performed. Indeed, similarity is one of the main criteria [24] by which ML systems determine the relationship between different data points or distributions and by which machine learning models are trained and their performance is evaluated. A trivial example is spam detection [22]: if a system is to classify emails as spam or non-spam, it needs to determine the degree of similarity between a given email and the group of emails that have been labeled as spam. Then, as for the less common tasks here we mention the assessment of the extent to which a collection of data points is similar to a given training set: In this case, the goal is determining the extent any new instance is representative of the same underlying distribution from which the training data points have been extracted, so that we can assume that the performance exhibited by the classifier on will also be guaranteed when applied to new data points [4]. Finally, we also mention the essential role similarity plays in the assessment of the extent to which two data distributions are similar [17].

There are various ways to measure similarity: the main ones include measures derived from distance metrics, such as (the dual of) Euclidean distance and Cosine distance; kernels [24], such as the radial basis function kernel; and measures drawn from fuzzy relations [16]. If one has to choose one metric over the others, this is usually chosen on the basis of traditional performance metrics (such as accuracy, area under the curve (AUC), sensitivity, specificity, their harmonic average (F1 score) and the like) applied to some downstream task. Notwithstanding this variety of performance metrics, it is usually assumed that the choice of the most suitable similarity measure will depend on the nature of the data and the goals of the ML task; however, clear guidelines or heuristics for this choice are still missing.

Moreover, researchers face several challenges when they are to evaluate whether a given similarity metric captures how human raters consider two cases similar; besides the fact that it is difficult to provide an appropriate definition of similarity for a given task, the main challenge lies in the subjectivity of the task itself: as illustrated in [5, 13], and widely known in some scientific field (like medicine), human judgment is affected by some amount of intrinsic and unavoidable subjectivity what Kahneman with a term calls noise [13]. For this reason, different people may have different opinions either on what constitutes similarity or on the extent two objects are similar to each other. Most relevantly, this issue is further exacerbated when one compares human and computational assessments of similarity: indeed, while an appropriate matching of these two evaluations is of critical importance if one aims at using similarity to support human decision making, these assessments could ground on entirely different principles and features and hence provide discording results.

Indeed, some previous work studied the match between human and computational definitions of similarity: Diaz et al. [10] studied similarity in the context of artwork, showing how traditional similarity metrics on images (based either on color histograms, or qualitative features, such as elicited emotions, knowledge

about the artworks and their context, or tag-based content similarity) poorly explained the human perception of similarity, which was furthermore shown to significantly vary across subjects; Towne et al. [25] studied similarity in the context of document-topic classification, by comparing similarity assessments reported by humans to those obtained through topic models, and found only a moderate agreement between humans' and LDA reported similarity assessments; similarly, Colla et al. [9] studied the relationship between computational and human assessments of similarity in the context of concept similarity, showing only a moderate correlation which increased only when humans were provided with a textual explanations for the computational similarity score, making them aware of linguistical features that they did not consider but were instead central in the algorithmic evaluation of similarity. Thus, these results highlight how it is difficult to determine whether a similarity metric is capturing the way that human raters consider two cases similar, for different reasons: because we would always refer to the average human rater; because there is not a clear right or wrong answer when it comes to determining the similarity between two cases; because similarity is strongly context-dependent, and hence results obtained in a given setting cannot be straightforwardly generalized to other ones. Furthermore, we note that previous work, such as the ones cited above, did not focus on similarity in a decision making setting, which is the more typical one when considering the application of AI and XAI systems in human practice.

The rest of the work presents the study we performed to address two main research questions: 1) is there any difference between common similarity metrics with respect to their capability to reflect how humans see different objects as similar? 2) if any such a difference exists, which similarity metric is better for a given task (in our case, selecting the most similar diagnostic images) and is this adequate with respect to this task? And lastly, 3) are the images retrieved from a case repository with common machine learning techniques perceived as useful by experts in their case interpretation and decision making?

3 The similarity assessment user study

In this section we present the user study designed and performed to address the research questions proposed in Section 2: we will outline the methods applied, and report the collected results. A discussion and concluding sections will follow after having presented the results.

3.1 Methods

Choice of the similarity metrics A wide range of metrics for assessing similarity between datasets, encompassing both tabular and image datasets, has been explored, debated, and implemented within the domain of specialized machine learning literature. Among the most recognized are the Euclidean Distance, Cosine Similarity, Pearson Correlation Coefficient, and specifically for images, the Structural Similarity Index (SSIM). Focusing on a subset, we hypothesized

that Euclidean Distance and Cosine Similarity are the most prevalent metrics. To validate this claim we performed two queries on Google Scholar: the query ("radiology" AND ("machine learning" OR "deep learning") AND "similarity") reported 18,200 results[5]; the more specific query ("radiology" AND ("machine learning" OR "deep learning") AND "similarity" AND ("Euclidean" OR "Cosine")), which is equivalent but focused on works explicitly mentioning either the term Euclidean or Cosine yielded 12,500 results. Thus, we concluded that at least one of this similarity measures was used in more than two thirds of the retrieved articles. This motivated us to focus on Cosine and Euclidean similarity to apply in our user study. In what follows, we briefly recall the definition of these metrics.

Embedding Extraction for Similarity Computation Consider $\mathbb{R}^n$ as the n-dimensional vector space over the real numbers. The vectors used for computing similarity and dissimilarity metrics are derived from a fine-tuned deep-learning model. Specifically, we employed a transfer-learning approach with a pre-trained ResNeXt-50 model, which we adapted for binary classification by modifying the last dense layer to have two neurons. This model was trained using the softmax function and cross-entropy loss. After training, we extracted embeddings from the last dense layer before the output layer, representing high-level features of the X-ray images in $\mathbb{R}^n$. We computed the similarity between these embeddings using Cosine and Euclidean distances to retrieve the most similar cases. This methodology allowed us to evaluate and compare the model-derived similarity metrics with human judgments.

The chosen metrics Let us denote with $\mathbb{R}^n$ the n-dimensional vector space over the real numbers. The Euclidean distance is the natural distance metric between two vectors $x, y \in \mathbb{R}^n$. It is calculated as:

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

The Cosine similarity, by contrast, is a measure of similarity between two vectors that measures the Cosine of the angle between them. It is calculated as:

$$s_{\cos}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{2}$$

where $||x||$ denotes the Euclidean (or l_2) norm of x.

Notice that, theoretically speaking, Cosine similarity and Euclidean distance have semantics that are dual to each other: indeed, whereas the first quantifies

[5] This query and the next one were performed on the 23rd of February 2024

the similarity between two objects, the second one quantifies their dissimilarity or difference. Nonetheless, and importantly for what follows, both can be expressed as metric distances: indeed, from the Cosine similarity one can easily define a Cosine distance as:

$$d_c(\mathbf{x}, \mathbf{y}) = 1 - s_{\cos}(\mathbf{x}, \mathbf{y}) \tag{3}$$

Thus, in the following, we will refer to the two distance metrics defined above.

Comparison methods To compare how humans perceive similarity to the scores generated by the two metrics described in Section 3.1, we calculated the Spearman's rank correlation coefficients and the Krippendorff's alpha. The Spearman's coefficients were used to assess the degree of correlation of the absolute values of similarity ratings and scores, while Krippendorff's alpha was applied to evaluate the level of agreement across the corresponding distribution quartiles. The fundamental concept here is that the method exhibiting the highest correlation and agreement, in absolute terms, would most closely align with how humans interpret clinical similarity of cases.

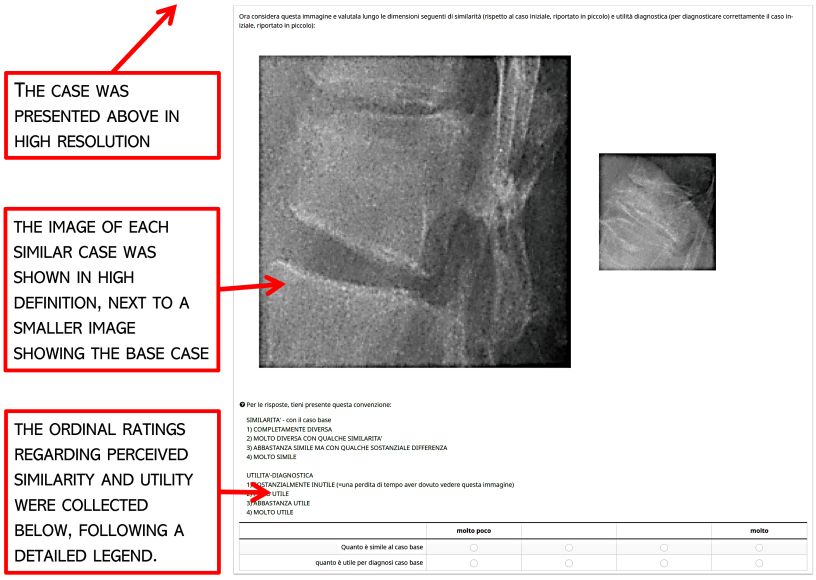

Fig. 1. Screenshot of the online questionnaire used to collect the users' ratings and perceptions.

The user study design To perform our similarity assessment study, we involved four experts in x-ray reading: two board-certified orthopedists and two

orthopedic radiologists were asked to independently consider 18 x-ray cases suspected of presenting some vertebral lesion. These cases had been previously selected for another study by one of the orthopedists involved to be representative of medium-to-high complexity and diagnostic difficulty (see [3]).

For each case (base case), the clinicians had to assess to which extent four *other* cases, retrieved from the training set and presented to them, were to be considered *clinically similar* to the base case, on a four-value ordinal scale ranging from 'completely different' (1) to 'very similar' (4)[6]. In order to detect *clinical similarity* between the diagnostic images shown to the clinicians, we agreed with them that they had to consider the anatomical site, the main anatomical elements involved, the features of the fracture (if any), and just *any* other analogy or similarity that could make a case "similar" and *educationally instructive* to inform the most accurate and most efficient interpretation of the base case.

For both the base case as well as the similar cases that were retrieved, the human experts evaluated the original images extracted from the hospital PACS system. By contrast, the similarity scores used to retrieve the similar cases were computed on images that had been previously normalized in terms of brightness and contrast to ensure consistency across the training set. The model was trained on an augmented dataset, which included typical procedures such as rotation, flipping, and scaling of the normalized images.

Moreover, we also asked the raters to assess the utility of considering the cases retrieved by similarity for the correct interpretation of the base case: also in this case, the raters employed a four-value ordinal scale ranging from 'substantially useless (1) to 'very useful' (4). We agreed that the meaning of the lowest score was associated with 'a waste of time having had to see also this image with respect to the base case'[7]. The clinicians' ratings were collected by means of an online multi-page questionnaire, developed on the Limesurvey platform (see Figure 1).

3.2 The user study results

In Table 1 we present the degree of correspondence between human perceptions of similarity (72 cases, 288 ratings) and between those perceptions and the computed similarities. We represent this degree both in terms of correlation (Spearman rho) and agreement (Krippendorff's alpha).

Regarding Figure 2, the Spearman correlation between the median perceived similarity (that is the extent the base case and the cases retrieved by the AI were found similar) and the Cosine similarity scores were found to be significant but only of moderate strength ($r(72) = .41$, $p = <.001$); while the correlation between median perceived similarity and the Euclidean distance scores was found

[6] The other anchors of the scale were 'very different with some similarity' (2) and 'quite similar but with some substantial difference' (3). The four-value scale was adopted to mitigate label noise (with few values) and central tendency bias (with an odd number of values).

[7] The other values of the scale were 'poorly useful' (2) and 'fairly useful' (3).

Table 1. Comparative Analysis of Inter-rater Agreement and Correlation between human raters and the analyzed metrics. Correlations are reported in terms of Spearman's rank correlation coefficients, while agreement is reported in terms of Krippendorff's alpha. Correlation scores are all highly significant, with p-values lower than .01. Correlation legend (based on [20]): very strong correlation; strong correlation; moderate correlation; weak correlation.

Comparison of similarity perceptions	Correlation ρ	Agreement α
Btw Cosine & Euclidean	-.83	-.77
Among humans	.64	.54
Among radiologists	.64	.42
Among orthopedists	.76	.75
Btw Radiologists & Orthopedists	.61	.50.
Btw Cosine & Humans	.41	.43
Btw Euclidean & Human	-.34	-.38

to be significant but weak (r(72) = -.34, p = .004). The difference between these correlations is not significant: as the confidence interval ([-.31, .15]) of the difference between the two correlations includes 0, we fail to reject the null hypothesis of no difference [11].

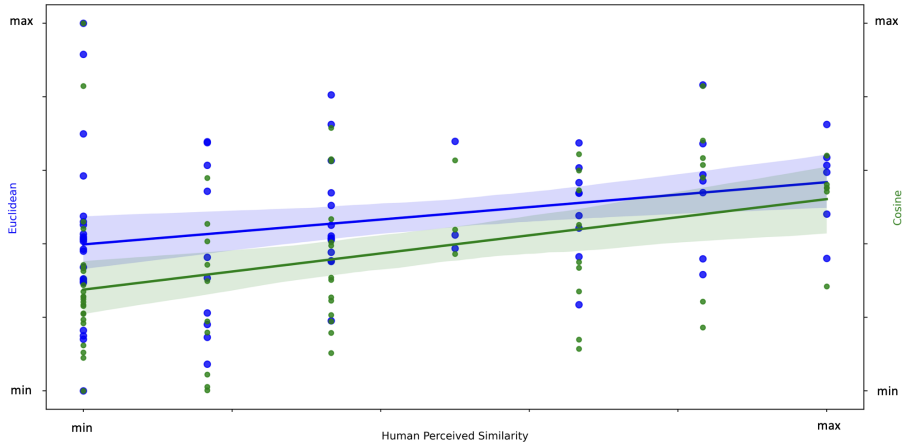

Fig. 2. This scatterplot illustrates the correlation between Euclidean (left, blue) and Cosine similarity metrics (right, green), both normalized to the [0,1] interval (min - max), thereby emphasizing similarity over distance between pairs of instances (denoted by blue and green circles, respectively), in comparison with the human-perceived similarity for the same instance pairs. A more robust correlation, as reflected by a steeper slope, signifies a higher concordance between the computational measures and human judgment.

In regard to perceived utility, the medical raters did not find considering the retrieved cases particularly useful to make a better, i.e., more accurate, interpretation of the base case (see Figure 3). The average utility score (i.e., 2) was significantly lower than the middle level (2.5). Moreover, we could not find any significant difference with respect to the class (positive vs negative) of the base case (p-value= .56, Z= 0.58, U=1438.5, standardized effect size= .057).

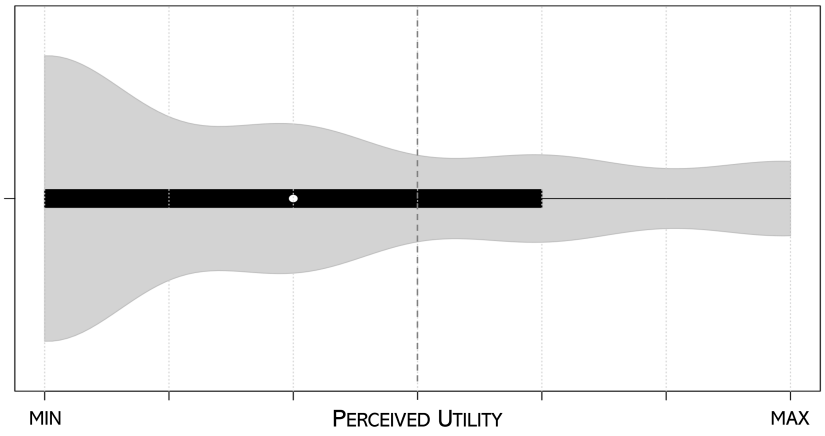

Fig. 3. Violin plot (embedding a box plot) of the perceived usefulness of considering the similar cases identified in the user study (N=288 ratings). The dotted line indicates the midline of the scale, suggesting perceived utility was generally low.

4 Discussion

The main findings of our study, as reflected in Table 1 can be summarized as follows: Cosine-based and Euclidean similarities do not differ too much, as also reflected by the high level of agreement between their scores [14]. Conversely, human perceptions of similarity are less correlated, and their agreement is likewise lower, on the threshold of sufficient reliability for ground truthing requirements [14], although the agreement between orthopedists was much higher (and hence adequate also for ground truth annotation) than that observed between radiologists. This suggests that the concept of similarity between cases is inherently vague, and that for similarity annotation ML practitioners should only rely on the perceptions of renowned experts with a strong common background rather than involving people who not only are not professionals but also do not share the same specialty. Furthermore, no computational metric achieved a level of alignment with human perception comparable to the level of intrinsic alignment between humans. In fact, cases extracted on the basis of either metrics were not found to be informative or useful for the interpretation of a given case.

That notwithstanding, Cosine-based similarity was found to be more similar to the human similarity than the Euclidean one. This means that automating the similarity assessment could lead to unreliable evaluations, and that, if this is deemed necessary in some application tasks, it is better to adopt Cosine-based similarity.

The primary limitation of this study is its relatively small sample size, a common challenge encountered in exploratory research such as ours. Despite this, our research is underpinned by 18 authentic cases, meticulously selected by a subject matter expert to encapsulate the breadth of complexity observed in real-world decision-making scenarios. Moreover, this study engaged four experts to assess 72 similarity relationships, yielding nearly 300 independent evaluations. Based on these figures, and notwithstanding these limitations, our research provides meaningful insights into the field of radiological image interpretation, highlighting several critical aspects. First, similarity metrics are not universally effective in retrieving pertinent and relevant cases, with significant variations in their alignment with human experts' perceptions of similarity. This seems to be due to the fact that commonly used similarity metrics in medical domains fail to capture what physicians consider similar, indicating a discrepancy between technical metrics and clinical relevance. Our analysis thus reveals a need for further research to identify or develop the most suitable similarity metrics for specific tasks, enhancing case retrieval and diagnostic support [19]. Conversely, reliance on less effective techniques could lead to suboptimal outcomes, providing clinically insignificant information. Despite its poor correlation with human perceptions, Cosine distance scores were nevertheless found to be more aligned with physicians' perceptions than Euclidean similarity scores; therefore Cosine-based similarity should be the metric of choice should one wish to treat similarity, at least in contexts of orthopedic image interpretation.

In light of our findings, it is pertinent to acknowledge the contributions of prior research in metric learning [15] and similarity learning [6, 8, 21], which have explored the application of machine learning (ML) techniques to automatically generate similarity measures. These methods, including those based on qualitative similarity assessments [7, 8], offer potential benefits for both broad and specific applications. They can provide general similarity measures applicable across various tasks, or tailor similarity assessments to the nuanced perceptions of a small group or even an individual clinician, assuming an adequate dataset is available [6]. Given the insights from our experiments, we advocate for increased research focus on developing more sophisticated computational approaches for assessing similarity [19]. This emphasis is crucial due to its significance in both human cognitive processes and the advancement of machine learning technologies.

5 Conclusions

In conclusion, our study underscores a significant discrepancy between widely-used similarity metrics, specifically Cosine and Euclidean distances, and the

evaluative processes of humans. This divergence is particularly pronounced in the alignment of AI-driven similarity measures with human perception, an essential factor in critical decision-making areas like medical diagnosis.

The minimal correlation between these mathematical models and the assessments of medical professionals accentuates the difficulty of aligning AI-generated similarity metrics with human intuition. Although Cosine distance slightly outperforms Euclidean distance, the difference is statistically insignificant, suggesting a slight but not meaningful preference for Cosine distance in the retrieval of clinically similar cases. However, this minor preference is overshadowed by the larger issue: both metrics fall short in accurately reflecting the nuanced complexities of human judgment on case similarity. This gap highlights the pressing need for developing similarity metrics that are more aligned with human cognition, possibly through the application of sophisticated ML techniques that incorporate qualitative evaluations.

Despite the limitations of our study, such as its small sample size and limited scope, it stresses the importance of rethinking the current approaches to similarity assessment in AI. We propose a collaborative effort that combines the insights of medical professionals, cognitive scientists, and AI researchers to bridge the gap between machine and human perceptions of similarity. Looking ahead, the goal should not only be to improve the accuracy of AI but also to ensure that AI-generated decisions and recommendations are intuitively meaningful and valuable to human users, particularly in critical sectors like medical diagnosis.

References

1. Balcan, M.F., Blum, A.: On a theory of learning with similarity functions. In: Proceedings of the 23rd international conference on Machine learning. pp. 73–80 (2006)
2. Cabitza, F., al.: Never tell me the odds. investigating pro-hoc explanations, as instances of frictional ai, in medical decision making. Artificial Intelligence in Medicine (2024)
3. Cabitza, F., Campagner, A., Famiglini, L., Gallazzi, E., La Maida, G.A.: Color shadows (part i): Exploratory usability evaluation of activation maps in radiological machine learning. In: Machine Learning and Knowledge Extraction. CD-MAKE 2022. Lecture Notes in Computer Science, vol 13480. pp. 31–50. Springer (2022)
4. Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L.G., Challa, F., Sulejmani, A., Seghezzi, M., Carobene, A.: The importance of being external. methodological insights for the external validation of machine learning models in medicine. Computer Methods and Programs in Biomedicine **208**, 106288 (2021)
5. Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., Berjano, P.: The elephant in the record: on the multiplicity of data recording work. Health informatics journal **25**(3), 475–490 (2019)
6. Cao, Q., Guo, Z.C., Ying, Y.: Generalization bounds for metric and similarity learning. Machine Learning **102**, 115–132 (2016)
7. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11**(3) (2010)

8. Cheng, W., Hüllermeier, E.: Learning similarity functions from qualitative feedback. In: Advances in Case-Based Reasoning. ECCBR 2008. Lecture Notes in Computer Science, vol 5239. pp. 120–134. Springer (2008)
9. Colla, D., Mensa, E., Radicioni, D.P., Lieto, A.: Tell me why: Computational explanation of conceptual similarity judgments. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. IPMU 2018. Communications in Computer and Information Science, vol 853. pp. 74–85. Springer (2018)
10. Díaz-Agudo, B., Jimenez-Diaz, G., Jorro-Aragoneses, J.L.: User evaluation to measure the perception of similarity measures in artworks. In: Case-Based Reasoning Research and Development. ICCBR 2021. Lecture Notes in Computer Science, vol 12877. pp. 48–63. Springer (2021)
11. Diedenhofen, B., Musch, J.: cocor: A comprehensive solution for the statistical comparison of correlations. PloS one **10**(4), e0121945 (2015)
12. Holyoak, K.J.: Analogy and relational reasoning. The Oxford handbook of thinking and reasoning pp. 234–259 (2012)
13. Kahneman, D., Sibony, O., Sunstein, C.R.: Noise: a flaw in human judgment. Hachette UK (2021)
14. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (2018)
15. Kulis, B., et al.: Metric learning: A survey. Foundations and Trends® in Machine Learning **5**(4), 287–364 (2013)
16. Moser, B.: On representing and generating kernels by fuzzy equivalence relations. Journal of machine learning research **7**(12) (2006)
17. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al.: Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning **10**(1-2), 1–141 (2017)
18. Pagliani, P., Chakraborty, M.: A Geometry of Approximation: Rough Set Theory: Logic, Algebra and Topology of Conceptual Patterns. Springer (2008)
19. Parimbelli, E., Marini, S., Sacchi, L., Bellazzi, R.: Patient similarity for precision medicine: A systematic review. Journal of biomedical informatics **83**, 87–96 (2018)
20. Prion, S., Haerling, K.A.: Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. Clinical Simulation in Nursing **10**(10), 535–536 (2014)
21. Rahnama, J., Hüllermeier, E.: Learning tversky similarity. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1238. pp. 269–280. Springer (2020)
22. Rao, S., Verma, A.K., Bhatia, T.: A review on social spam detection: challenges, open issues, and future directions. Expert Systems with Applications **186**, 115742 (2021)
23. Schoenborn, J.M., Weber, R.O., Aha, D.W., Cassens, J., Althoff, K.D.: Explainable case-based reasoning: A survey. In: AAAI-21 Workshop Proceedings (2021)
24. Scholkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2018)
25. Towne, W.B., Rosé, C.P., Herbsleb, J.D.: Measuring similarity similarly: Lda and human perception. ACM Transactions on Intelligent Systems and Technology **8**(1), 1–28 (2016)
26. Tversky, A.: Features of similarity. Psychological review **84**(4), 327 (1977)