How Accurate Does It Feel? – Human Perception of Different Types of Classification Mistakes

Andrea Papenmeier
andrea.papenmeier@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Dagmar Kern
dagmar.kern@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Daniel Hienert
daniel.hienert@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Yvonne Kammerer
kammerer@hdm-stuttgart.de
Stuttgart Media University
Stuttgart, Germany

Christin Seifert
christin.seifert@uni-due.de
University of Duisburg-Essen
Essen, Germany

ABSTRACT

Supervised machine learning utilizes large datasets, often with ground truth labels annotated by humans. While some data points are easy to classify, others are hard to classify, which reduces the inter-annotator agreement. This causes noise for the classifier and might affect the user's perception of the classifier's performance. In our research, we investigated whether the classification difficulty of a data point influences how strongly a prediction mistake reduces the "perceived accuracy". In an experimental online study, 225 participants interacted with three fictive classifiers with equal accuracy (73%). The classifiers made prediction mistakes on three different types of data points (easy, difficult, impossible). After the interaction, participants judged the classifier's accuracy. We found that not all prediction mistakes reduced the perceived accuracy equally. Furthermore, the perceived accuracy differed significantly from the calculated accuracy. To conclude, accuracy and related measures seem unsuitable to represent how users perceive the performance of classifiers.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Human computer interaction (HCI)**.

KEYWORDS

Accuracy; Perception; Annotations; Ground Truth

1 INTRODUCTION

Artificial intelligence (AI), and particularly machine learning-based methods, are key to many data-driven applications in various fields, e.g., financial risk assessment [13], hiring [29], medical diagnosis [14], music recommendations [62], or social media [2]. Throughout the years, machine learning systems have been optimized to increase their performance and produce better results. So far, however, the evaluation has mainly focused on the system perspective: Machine learning systems are often trained and evaluated on data that contains noise. But what if the quality of the underlying data is not as good as expected and does not reflect the real world? That would result in machine learning systems that are nearly perfect but might be perceived as performing poorly by humans. Consequences might be mistrust in the system [8, 9, 11], a bad user experience [15], or, even worse, severe consequences that affect people's life, e.g., introducing or replicating racist or cultural bias [34, 37, 39].

With the upcoming interest in a more socio-technical view on AI (e.g., [3, 4, 50, 63]), data work is recently gaining attention. Undervalued data quality causes cascading events that – often unnoticeably – impair the performance of machine learning systems [47]. Given a dataset including ground truth labels, supervised machine learning systems learn the underlying function that maps data to labels. Those labels are often retrospectively added to the data by human annotators. However, this might be challenging due to classification difficulties, such as ambiguity in the data [54] or disagreement amongst annotators. A low inter-annotator agreement [10] causes label noise in the data [1, 42, 49]. Furthermore, even experts cannot always annotate data free of doubt [1] and might create a controversial instead of a real ground truth. Researchers address label noise by considering, for example, individual annotator votes for a data label instead of a majority vote [45].

However, in some cases, several different labels are quite plausible, and prediction mistakes (according to the ground truth based on one of these labels) might not be perceived as mistakes by humans [54]. For example, a human annotator might label an image as "monastery", while a machine learning system predicts the label "church" [54]. A user would probably accept both labels. Similarly, the news headline "Keeping A Clear Mindset During School" (taken from the News Category Dataset [35]) was labeled with the topic "college", although "education" could also be an acceptable label. Ultimately, this means that system performance does not always correspond to how users perceive a system: It can be better than assumed [54] but also worse [19].

In this paper, we follow the line of research on user perception of the performance of machine learning systems [19, 23, 24, 54] and investigate to what extent different levels of classification difficulty of single data points affect how users perceive the performance of a classifier. To analyze the effects of classification difficulty, we first set up a crowdsourcing task (N = 54) to identify data points that are easy-to-classify, difficult-to-classify, and impossible-to-classify for humans for a topic classification dataset. We used a dataset that contains ground truth labels for user-formulated requirements on a new laptop and jacket. For example, *"it needs to have lots of pockets and a hood too"* is easy to classify, while *"I am going to buy a new cheap one"* is an impossible-to-classify sentence for which both labels are plausible. We then created two subsets for experimentation: One mixed dataset containing an equal number of easy-to-classify, difficult-to-classify, and impossible-to-classify data points and a clear dataset with only easy-to-classify data points. Subsequently, we manually produced the output of three fictive classifiers that made topic predictions for each data point with an accuracy of 73% and two fictive classifiers with a calculated accuracy of 100%. In an experimental online study (N = 225), we confronted participants with the predictions of one out of the five classifiers and asked them to assess the classifier's accuracy – we call this the "perceived accuracy".

Our findings revealed a significant difference in perceived accuracy between all three classifiers with 73% calculated accuracy: The more difficult it is for humans to correctly classify a data point, the less harmful is a prediction mistake on that data point. As a result, traditional performance measures (e.g., accuracy, precision, recall, F1, ROC AUC) do not accurately reflect the level of perceived accuracy. We further found that the dataset composition plays an important role: A perfect classifier of the mixed dataset was perceived as significantly less accurate than a perfect classifier on the clear dataset, although both had a calculated accuracy of 100%. These findings have important implications for how we should evaluate machine learning systems. Instead of focusing solely on accuracy, we suggest including user-centered factors such as classification difficulty in evaluation metrics.

2 RELATED WORK

In the following section, we contextualize our research in related work. First, we provide a short introduction to performance measures for machine learning systems that we will use to compare our results with. Then, we illustrate a common technical approach to deal with different prediction mistakes. How ground truth labels are gathered and how they contribute to noise in machine learning systems is discussed in 2.3. The following subsection shows how humans perceived systems' accuracies and prediction mistakes. The related work section concludes with an overview of how machine learning systems can be evaluated from a more human-centered perspective.

2.1 Performance of Machine Learning Systems

The performance of classifiers in supervised machine learning can be assessed with different measures, such as accuracy, sensitivity, specificity, precision, recall, F1-score, ROC-curve, and ROC AUC [6, 52]. In binary classification, those eight measures can be defined based on the four possible outcomes of a two-class problem[1]: true positive predictions (tp) and true negative predictions (tn), i.e., examples correctly classified as positive or negative, and their incorrectly classified counterparts false positives (fp) and false negatives (fn). Accuracy is defined as the ratio of correct predictions ($tp + tn$) and total number of predictions ($tp+tn+fp+fn$). Accuracy considers all prediction mistakes equally, independent of the specific type of mistake. Other measures show a class-centric view: Precision is the ratio of correctly predicted positive examples and all examples predicted as positive ($\frac{tp}{tp+fp}$), i.e., comparing the number of examples that were correctly assigned to the positive class to the total number of examples that were initially assigned to the positive class. Recall compares the number of correctly predicted positive examples to the number of positive examples, including those incorrectly predicted as negative ($\frac{tp}{tp+fn}$). In medical applications, sensitivity and specificity are commonly used, where sensitivity and recall are equivalent, and specificity is defined as the ratio of examples correctly predicted as negative and all examples predicted as negative ($\frac{tn}{tn+fp}$). Precision, recall/sensitivity, and specificity assess the effectiveness of the algorithm with respect to a single class [52]. The F1-score is the harmonic mean of precision and recall [52]. All six measures produce values in the range of 0 to 1.

The ROC (receiver operating characteristic) curve [6] is a graph showing the classification performance when varying the class discrimination threshold, i.e., the threshold on the continuous classification score above which the positive class is assigned. The ROC plot shows the relation of the true positive rate (or sensitivity) and false positive rate (1 - specificity). A random classifier, i.e., a classifier that randomly guesses the label, lies on the diagonal in ROC space. A classifier that is better than random shows a ROC curve above the diagonal. AUC takes the area under the ROC curve as a performance measure. Its value ranges between 0 and 1, with values > 0.5 indicating that the classifier performs better than random guessing.

2.2 Cost-Sensitive Learning

In the performance measures discussed in the previous section, prediction mistakes of the same class have the same influence on the final score. That is, it does not matter on which particular example within a class the prediction was incorrect. However, these measures are oblivious to other characteristics, such as what consequences follow from the individual mistake or how plausible a class label appears to a human.

Cost-sensitive learning acknowledges that different prediction mistakes cause different consequences [12, 25]. For example, in medical diagnosis, falsely assuming a patient has cancer and consequently requesting further tests is a less severe mistake than falsely assuming a patient is healthy and consequently discontinuing further treatment. Each type of prediction outcome (tp, tn, fp, fn)

[1] Here we use the common class labels "positive" and "negative", originating from information retrieval where an example can be relevant or irrelevant, i.e., "positive" or "negative". The definition of measures extends to any two-class problem, regardless of the class label names.

receives a "cost" that will be added to a grand total during the evaluation process. Instead of optimizing for one of the performance measures described in the previous section, cost-sensitive systems will try to reduce the sum of all costs associated with the prediction outcomes. Cost-sensitive learning is applicable especially in areas where different prediction mistakes severely impact a personal or societal level, including applications in medicine [60], hiring [67], or credit evaluation [64].

Besides penalizing types of prediction outcomes (tp, tn, fp, fn), other cost distributions are possible. Turney [55] presents a comprehensive overview of cost types that could be included in the training process of machine learning systems. Amongst others, Turney differentiates four types of conditional costs of prediction mistakes, with one being individual costs per example [55], such as the amount of money lost per ignored fraud case in fraud detection.

2.3 Ground Truth Labels and Label Noise

For classification tasks, such as topic prediction, medical diagnosis, or fraud detection, a labeled dataset is needed to train and test supervised machine learning classifiers. Each data point in a labeled dataset is assigned to at least one target label, which constitutes the ground truth. The ground truth may be collected together with the data points, e.g., reviews with star ratings [33], or retrospectively annotated, e.g., objects in a picture [7]. The annotation of a dataset is a complex task fulfilled by human annotators [36] who do not always agree on the target label. The disagreement in the labeling process can have various reasons. Annotators' interpretation of a label concept can evolve throughout the annotation process [26] or annotators may be inattentive – not paying close attention or accidentally selecting the wrong label [45]. Other reasons for disagreement could be ambiguity in the task instructions, in the label interpretation, or in the data itself [10]. Due to the ambiguity, some data points are difficult to label [10, 54], leading to uncertainty of the annotators.

The most common approach to dealing with disagreement between annotators for determining the final ground truth label is the majority vote for a single data point [49]. In cases of disagreement, collecting additional annotations for a data point does not necessarily lead to a more reliable majority vote. Although Gurari and Graumann [21] showed that it is possible to predict annotator agreement and collect more annotations for data points with high disagreement probability, Pavlick et al. [42] found that despite having a high number of annotations for a single data point, the disagreement ratio usually remains constant. However, regardless of how much disagreement was recorded, the final ground truth often collapses all annotations into a single label [10]. This aggregation creates "label noise" [1, 49, 69], i.e., unreliable ground truth data, on which machine learning systems are trained.

There is extensive research to address the label noise problem from a system perspective (see [1, 16] for an overview). Recent research focuses on considering the full label distribution to reflect human perceptual uncertainty (e.g., in [43, 61]). Gordon et al. [19] introduce their disagreement deconvolution approach to align machine learning classification metrics more closely with user-centric performance measures. Considering the disagreement in annotators' opinions, especially in social computing tasks, they suggest comparing classifier predictions to individual, stable opinions from each annotator. Their algorithm allows computing disagreement-adjusted versions of any standard classification measure, taking each annotators' primary label (stable opinion) as individual ground truth values. Gordon et al. evaluated their algorithm with three social computing tasks and two classic machine learning tasks. Especially for the latter, their approach drastically reduced reported performance (as calculated by classical performance measures). Resnick et al. [45] follow a similar approach: They consider individual human annotators rather than a majority vote of all annotators and introduce a new variable that describes "how many labels would be needed to get the same expected score that the classifier got" [45]. They suggest using this score to determine whether a system performs well enough.

2.4 Human Perception of Accuracy and Prediction Mistakes

Research found that low perceived accuracy might affect user experience and, especially, trust in such systems [11]. Yin et al. [68] found significant effects of stated accuracy on people's trust in the system. In their experiment, participants reported more trust in systems that initially claimed to have a high level of accuracy. But this priming effect decreased when users experienced a lower accuracy in practice. Other research found that people stop trusting an algorithm after observing prediction mistakes, even when the algorithm overall outperforms human predictions [8, 11]. This might be explained by the fact that when making judgments under uncertainty, individuals often rely on mental short-cuts or rules of thumb (i.e., heuristics) rather than undertaking a thorough and rational analysis [57]. One such heuristic is the availability heuristic [56], in which individuals evaluate the probability of cases or events based on their mental availability, i.e., "by the ease with which relevant instances come to mind" [56, p. 207]. Thus, if multiple cases or events are disproportionally available, the availability heuristic can lead to biased probability judgments (i.e., over- or underestimation), also known as availability bias [59]. Underestimating a classifier's performance does not only reduce trust [68] but can also result in users ignoring a system recommendation [9]. Nourani et al. [38] added explanations to their machine learning system and researched their effect on user trust and perceived accuracy. Different types of explanations had different effects on how participants perceived the system's accuracy, showing that multiple factors can influence users' perception. Roy et al. [46] stated that if users have a chance to control the system, the self-reported satisfaction remained constant even when system accuracy is relatively low.

Investigating single predictions rather than the aggregated view of a classifier's performance, Kocielnik et al. [24] tested whether a high-recall system (avoiding fn) is more accepted than a high-precision system (avoiding fp). In their use case, a machine learning system automatically extracted and scheduled meeting appointments from emails. Although other researchers have argued (but not shown) that, in general, false positives (fp) are less accepted by users than false negatives (fn) [23, 28], Kocielnik et al. found that for their application, avoiding false negatives was more important to the users. The perceived accuracy of their system was lower for the high-precision system than for the high-recall system. Even though no general rule can be established, it shows that users weigh different outcomes differently, depending on the associated consequences. Although not providing empirical evidence, Lipton [30] argues that users might be inclined to accept systems that show behavior similar to theirs, i.e., systems which make mistakes only on data points that are difficult to classify for humans. Interestingly, in a clinical text labeling scenario, Levy et al. [27] found that domain experts notice if system label recommendations are inadequate but are still likely to accept them for lack of suitable alternatives. In an annotation setting, Tsipras et al. [54] also concentrate on how single classifier predictions are perceived by users. By asking annotators how reasonable they think a prediction is, they compared the performance of different systems on the ImageNet dataset [7]. In their experiment, participants decided whether they agree with predicted labels for multi-object images. Their results showed that non-expert annotators often judged dataset labels as valid even when the prediction was incorrect, which means that the perceived accuracy of the systems was higher than the calculated accuracy. They conclude that just focusing on measuring accuracy is not sufficient to understand system performance in their case on the object recognition task.

2.5 Human-Centered Evaluation of Machine Learning Systems

While machine learning measures mainly focus on system performance, human-centered machine learning (HCML) investigates how the users are affected by such systems [22]. Accordingly, research in HCML focuses on users' evaluations of machine learning systems, including aspects such as users' reliance [31] on the system, or trust [40, 68] in the system, the perceived explainability [51], interpretability [44], or fairness [20, 32] of the system, and the experienced or perceived accuracy of the system [15, 19, 23, 54]. In a recent HCML survey paper, Kaluarachchi et al. [22] provide a comprehensive overview of user studies with machine learning systems. To investigate machine learning applications from a human-centered perspective, standard HCI methods such as observations, interviews, and questionnaires are used or have been adapted to machine learning scenarios. For example, Gero et al. [17] used the think-aloud method and questionnaires to investigate participants' mental models of an AI agent. Other researchers have employed observation studies and interviews to investigate user experience in machine learning settings [5, 65]. While in-person user studies in labs (like in [66]) are rather rare, crowdsourcing tasks are commonly used to collect feedback from the users. Common tasks include labeling text or image data (e.g., in [31]) or assessing given classifications (e.g., in [27, 40, 48, 51]). Often, the classification accuracy is systematically varied to avoid unpredictable behavior by the system [31, 40, 66].

In the context of perception of automated systems, Kay et al. [23] present a survey instrument that helps to predict how acceptable users will find the accuracy of a system. In their user study, they applied an adapted version of the extended Technology Acceptance Model questionnaire [58] to assess their newly introduced measure "acceptability of accuracy". They argue that, compared with the popular F1-score, their approach correlates more with the user's real-life acceptance of a system's accuracy. While the "acceptability of accuracy" is an important addition in the toolbox of HCML evaluation metrics, it does not explain what caused the acceptance or rejection of a system.

3 METHOD

The major goal of the present work is to investigate the role of different types of prediction mistakes on human's perceived accuracy of machine learning systems. Our research is driven by the following research question:

RQ Do different types of prediction mistakes of a classifier equally impact users' perceived accuracy of that classifier?

The literature has shown that creating ground truth labels for training and evaluation of machine learning systems is challenging (e.g., due to ambiguity) [10, 45, 54]. Consequently, not all predictions that are deemed incorrect by the ground truth are perceived to be incorrect by users [54]. It remains unclear to what extent different levels of ambiguity of individual data points impact the users' perception of a classifier's accuracy as a whole. We therefore focused on three different levels of data point ambiguity from the user's point of view: data points that are easy, difficult, and impossible to classify for users.

We expected that the more difficult it is for users to identify a data point's label (classification difficulty), the fewer prediction mistakes made by a system are noticed by a user. As a consequence, the viewer prediction mistakes are noticed by a user, the higher is the perceived accuracy of the system. To be more precise, we hypothesized that prediction mistakes on easy-to-classify data points lead to a significantly lower perceived accuracy than prediction mistakes on difficult-to-classify data points [**H1**]. We also expected that prediction mistakes on difficult-to-classify data points lead to a significantly lower perceived accuracy than those on impossible-to-classify data points [**H2**]. Consequently, if the same number of prediction mistakes leads to a significantly different perceived accuracy for at least one level of classification difficulty (easy, difficult, impossible), the perceived accuracy should also differ significantly from the standard measure of accuracy (calculated as the number of correct predictions divided by the number of all data points) for at least one level of classification difficulty [**H3**]. If human annotators disagree on the label of a data point, even correct predictions on impossible-to-classify data points could be perceived as prediction mistakes. We therefore hypothesized that a classifier with 100% correct predictions will have a perceived accuracy of significantly lower than 100% if it contains some impossible-to-classify data points [**H4**].

To test our hypotheses, we first collected information on the classification difficulty for a set of data points in a crowdsourcing task (see Section 3.1 "Task and Materials"). Based on the results, we curated two datasets to use them with our five fictive classifiers (see Section 3.2 "Experimental Conditions"). In an experimental online user study, we evaluated how participants perceive the accuracy of the five classifiers (see Sections 3.3 "Procedure - 3.4 "Measures"). The experiment had received clearance by the ethics board of our institution. All materials and questionnaires as well as the briefing and debriefing text are available online[2].

[2] https://git.gesis.org/papenmaa/chi22_perceivedaccuracy

3.1 Task and Materials

For experimenting with data points at different levels of classification difficulty, we needed a dataset with a proper (i.e., objectively correct) ground truth that we can enrich with information about the classification difficulty of data points. We, therefore, chose the VACOS_NLQ dataset [41] that provides 3,560 product descriptions of laptops and jackets collected from English native speakers. During the collection of the VACOS_NLQ dataset, each participant was prompted to describe a jacket and a laptop. The ground truth is therefore given by design, not retrospectively annotated, resulting in a proper ground truth even if the texts themselves are ambiguous. Another advantage of the VACOS_NLQ dataset is that the classification task (assigning product categories to product description) is binary and simple, which allowed us to recruit participants without expert knowledge. The concepts of a "laptop" and a "jacket" are well-known and easily distinguishable for laypersons. Moreover, laptops and jackets both have unique characteristics (e.g., hood, zipper, technical hardware) but also share some common characteristics (e.g., price, color), yielding the potential for clear sentences as well as ambiguous sentences.

The product descriptions of the VACOS_NLQ dataset contain on average 3.7 sentences, which gives enough context to clearly identify the product in most cases. To obtain data points at various levels of classification difficulty, we split the descriptions into single sentences and manually pre-selected 129 sentences (62 jacket sentences, 67 laptop sentences). All pre-selected sentences mention one or two product characteristics and have a length similar to the mean length of all sentences in the dataset (M = 13.59 words, SD = 5.98). With the consent of the VACOS_NLQ dataset authors, the pre-selected sentences and the selection criteria are available online[3]. On the platform Prolific[4], we recruited 54 crowd workers (36 female (f), 18 male (m), 0 diverse (d)) with a mean age of M = 34.70 years (SD = 12.85) to classify sentences as either describing a jacket or describing a laptop. Each crowd worker labeled between 20 and 30 sentences. The sentences were drawn at random from the pre-selected pool of 129 sentences. Each sentence was classified by at least 10 workers. Additionally, two attention checks were included in the classification task. All workers passed the attention checks and subsequently received financial compensation of 1.20 GBP for an average completion time of 10 minutes.

We created three sets of sentences that span the whole range of classification difficulty:

- **Set 1: easy sentences.** Sentences that are easy to classify for a human, i.e., sentences for which 100% of the annotators in the crowdsourcing task decided for the correct label (as defined in the ground truth). Example:

  "it needs to have lots of pockets and a hood too" (ground truth label: jacket).

  Of the 129 pre-selected sentences, 35 fall in this category (16 jacket sentences, 19 laptop sentences).
- **Set 2: difficult sentences.** Sentences that are difficult to classify for a human but one class is more likely than the other, i.e., sentences for which around 70-80% of the crowd workers decided for the correct label. Example:

  "easy to care for, that is, to clean and to store" (ground truth label: jacket).

  We identified 11 sentences for this set (6 jacket sentences, 5 laptop sentences).
- **Set 3: impossible sentences.** Sentences that are impossible to classify for a human because both classes are equally likely, i.e., sentences for which around 40-60% of the crowd workers decided for the correct label. Example:

  "I am going to buy a new cheap one" (ground truth label: laptop).

  The crowdsourcing task revealed 13 suitable sentences for this set (4 jacket sentences, 9 laptop sentences).

A one-way ANOVA confirmed that the sets have a sufficient difference (p < .001) in average classification accuracy of annotators (i.e., how many annotators assigned the correct label). Post-hoc tests using two-sided Mann-Whitney U-tests showed that all sets are significantly different from each other (p < .001 for all three comparisons). Importantly, the one-way ANOVA did not show a significant difference between sets for number of characters (p = .780), number of words (p = .780), or lexical density[5] (p = .709). We, therefore, conclude that the three sets differ sufficiently in their classification difficulty for humans, which is, however, not provoked by a difference in lexical form or readability.

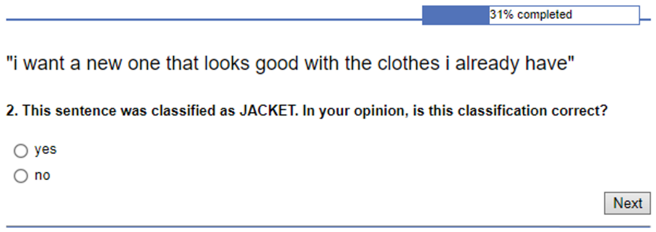
3.2 Experimental Conditions

To evaluate the effect of prediction mistakes for different sentence types, we constructed two datasets from the sentence sets. The **clear dataset** consists of 30 sentences taken only from the set of easy cases. The **mixed dataset** contains a mixture of sentences, with 10 easy sentences, 10 difficult sentences, and 10 impossible

[3] https://git.gesis.org/papenmaa/chi22_perceivedaccuracy
[4] www.prolific.co
[5] Defined as the number of content words (verbs, nouns, adjectives, adverbs) divided by the total number of words in a sentence [18].

Table 1: Experimental conditions of the user study.

Condition	Dataset	Sentence Composition	Mistakes	Calculated Accuracy
(A) *8M_easy*	mixed	10 easy, 10 difficult, 10 impossible	8, on easy	0.73
(B) *8M_difficult*	mixed	10 easy, 10 difficult, 10 impossible	8, on difficult	0.73
(C) *8M_impossible*	mixed	10 easy, 10 difficult, 10 impossible	8, on impossible	0.73
(D1) *0M_mixed*	mixed	10 easy, 10 difficult, 10 impossible	0	1.0
(D2) *0M_clear*	clear	30 easy	0	1.0

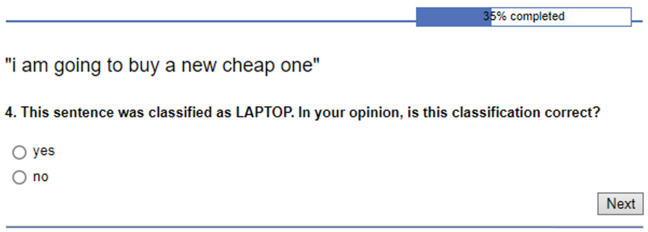


"i want a new one that looks good with the clothes i already have"

2. This sentence was classified as JACKET. In your opinion, is this classification correct?

○ yes
○ no

Figure 1: An easy-to-classify sentence with a correct prediction from the fictive classifier.



"i am going to buy a new cheap one"

4. This sentence was classified as LAPTOP. In your opinion, is this classification correct?

○ yes
○ no

Figure 2: An impossible-to-classify sentence with an incorrect prediction from the fictive classifier.

sentences. We manually created the output of five fictive classifiers (independent variable) to systematically inject prediction mistakes. We create (A) a system that makes 8 prediction mistakes on the 10 easy-to-classify sentences, (B) a system misclassifying 8 out of 10 difficult-to-classify sentences, and (C) a third system that misclassifies 8 out of 10 impossible-to-classify sentences on the mixed dataset. All three fictive systems (A, B, C) misclassify 8 out of 30 sentences in total and therefore have the same calculated accuracy of 73%. The prediction mistakes were made equally on both classes, hence four on sentences describing laptops and four on sentences describing jackets. Additionally, we introduced a perfect system that makes no mistakes, yielding a calculated accuracy of 100% (D1) on the mixed dataset as well as (D2) on the clear dataset. Table 1 shows the resulting five experimental conditions. We decided against training a machine learning system because we wanted to inject structured mistakes, ensuring that all mistakes are made on the same sentence set, e.g., only on the simple-to-classify sentences for condition A. Real-life machine learning systems are likely to make mistakes across all three sentence sets, which would make it difficult to investigate the differences between the sentence sets in our experiment.

3.3 Procedure

We set up an online user study. After giving informed consent, participants received the task description:

> *In this experiment, you will read 30 sentences from product descriptions. The sentences were written by users who were asked to describe a laptop or a jacket they want to buy online. Those sentences have been classified into "jacket descriptions" or "laptop descriptions". Your task is to assess whether the classification is correct.*

Participants did not receive information about who or what made the class predictions to avoid priming effects from participants' expectations about a classification system or about an expert. Figures 1 and 2 show examples of the participants' task in the user study. A single task consists of a sentence from the dataset, the fictive classifier's prediction, and a question asking the participants to indicate their agreement or disagreement with the prediction. We chose an interactive task to (1) measure the participants' agreement with the predictions on sentence level, (2) ensure they carefully read the sentences and predictions, and (3) enforce them to form an opinion about the sentence's class.

All participants passed a training phase in which they assessed the predictions of four sentences: two laptop sentences and two jacket sentences. One sentence of each class was easy to classify and the other one difficult to classify. Afterward, they were informed that they have passed the training and would now start with the main tasks. In the main phase, participants completed 30 tasks with two additional attention checks. The order of sentences was randomized in all conditions to prevent any order effects. Participants in conditions A, B, C, and D1 all interacted with the same 30 sentences from the mixed dataset, whereas participants in condition D2 all saw the same 30 sentences from the clear dataset. Participants did not receive feedback on their classification performance. Following the main phase, participants reported their perceived accuracy, subjective rating of the classifier's performance, perceived agreement, and answered three open questions (see Section 3.4 "Measures"). The participants also answered questions about their demographic background (age, gender, and domain knowledge of laptops and of jackets on a 5-point scale). Finally, the participants were debriefed.

3.4 Measures

We measured the following dependent variables in the user study:

(1) **Calculated accuracy**: accuracy of the fictive classifier with respect to the ground truth. Measured as:

$$\frac{\text{number of correct predictions by classifier}}{\text{total number of data points}}$$

(2) **Human accuracy**: accuracy of the participants with respect to the ground truth. Measured as:

$$\frac{\text{number of correct decisions by participant}}{\text{total number of data points}}$$

where the number of correct decisions is identified by observing agreement (answering "yes", see Figure 1) with correct classifier predictions and disagreement (answering "no") with incorrect classifier predictions.

(3) **Perceived accuracy**: participants' perception of the fictive classifier's accuracy. Collected after interacting with the fictive classifier by asking "What do you think: In how many cases was the correct product category (LAPTOP / JACKET) displayed?". Measured as:

$$\frac{\text{participants' perceived number of correct classifications}}{\text{total number of data points}}$$

(4) **Calculated agreement**: agreement of participants with the fictive classifier. Measured as:

$$\frac{\text{number of agreements (“yes” responses)}}{\text{total number of data points}}$$

(5) **Subjective rating**: participant's assessment of the fictive classifier's performance on a 7-point scale (1 = “very poor” to 7 = “excellent”)

Additionally, after interacting with the systems, we used open questions to ask participants to reflect on their decision-making and perception:

Q1 “In which cases did you have difficulties with your decision?”,
Q2 “In which cases was it easy to come to a decision?”,
Q3 “Why do you think those sentences were incorrectly classified?” (Only if the participant indicated that the classification system made at least one prediction mistake)

3.5 Participants

For our user study, we recruited 235 participants via Prolific[6]. Only residents of the US, UK, and Ireland being English native speakers were allowed to participate in the experiment. Additionally, we used the Prolific's prescreening options to exclude participants with literacy difficulties. After the experiment, we excluded 10 participants from our data due to low-effort responses, e.g., gibberish responses to the open questions or failed attention checks. Ultimately, we keep N = 225 valid responses. Participants (171 f, 53 m, 1 d) were randomly assigned to one of the five conditions (N = 45 in each condition). On average, participants were M = 35.70 years old (SD = 13.53 years). Their average domain knowledge of jackets (M = 3.97, SD = 0.94) and laptops (M = 3.95, SD = 0.91) was comparable, measured on a 5-point scale. The one-way ANOVA did not reveal significant differences between groups with respect to age (F(5,220) = 1.76, p = .139), gender (F(5,220) = 0.97, p = .424), or domain knowledge of jackets (F(5,220) = 2.42, p = .050) or laptops (F(5,220) = 0.62, p = .648). We paid all participants a financial compensation of 1.30 GBP for an average completion time of 12 minutes. The compensation was based on the minimum wage of the UK.

4 RESULTS

To answer our research question, we report differences of perceived accuracy between conditions ([**H1**]), explore how participants perceived the difficulty of the classification task, compare the perceived accuracy with the calculated accuracy ([**H2, H3**]), and investigate the implications of the dataset composition for the perceived accuracy ([**H4**]).

We used one-way ANOVAs for single-variable comparisons of conditions (e.g., comparing the perceived accuracy between conditions) with two-sided Mann-Whitney U-tests (with Bonferroni correction) as post-hoc tests. Likewise, when comparing dependent samples, we used Wilcoxon's two-sided signed-rank tests with Bonferroni adjustments. All statistical results are interpreted with α = .05 (two-sided).

4.1 Perceived Accuracy

To investigate whether all prediction mistakes are perceived equally by users we asked participants to assess the performance of the classifier subsequent to the 30 interactions. We measured performance in two ways: participants' subjective performance rating on a polarity profile and the perceived accuracy on a numeric scale. For the subjective ratings, we provide means and standard deviations as well as the post-hoc test results of comparing the conditions in Table 2. A one-way ANOVA showed a significant difference in subjective ratings (F(5,220) = 3.56, p = .031) between conditions A, B, and C, even though the calculated accuracy was the same in all three conditions (73%). Participants of condition C rated performance as significantly better than participants in B and A. However, participants in A and B did not significantly differ in their subjective ratings. Concerning perceived accuracy, Table 3 reports the means and standard deviations for the perceived accuracy per condition. The perceived accuracy (measured between 0 and 30, according to the number of correctly classified instances) provides a more fine-grained view on the systems' performance than the subjective rating (measured between 1 and 7). Similar to the results of the subjective ratings, a one-way ANOVA revealed a significant difference in perceived accuracy (F(5,220) = 93.18, p < .001). The results of the post-hoc test showed a significant difference between all three conditions: The classifier in A that makes obvious prediction mistakes only on easy-to-classify sentences was perceived to be accurate in significantly less instances than the classifiers in B and C. The classifier in C, making prediction mistakes on impossible-to-classify sentences, was perceived to be accurate in significantly

Table 2: Means and standard deviations of the subjective ratings of all classifiers with 73% calculated accuracy. Comparison of conditions using two-sided Mann-Whitney U-tests with Bonferroni correction. The asterisk (*) denotes significance at α = .05.

Condition	Subjective Rating	Comparison p-values		
(A) *8M_easy*	5.29 (1.15)	n/a	.449	.008*
(B) *8M_difficult*	5.49 (0.88)	.449	n/a	.026*
(C) *8M_impossible*	5.87 (1.05)	.008*	.026*	n/a

Table 3: Means and standard deviations of the perceived accuracy of all classifiers with 73% calculated accuracy. Comparison of conditions using two-sided Mann-Whitney U-tests with Bonferroni correction. The asterisk (*) denotes significance at α = .05.

Condition	Perceived Accuracy	Comparison p-values		
(A) *8M_easy*	0.66 (0.15)	n/a	.008*	< .001*
(B) *8M_difficult*	0.73 (0.13)	.008*	n/a	.030*
(C) *8M_impossible*	0.78 (0.14)	< .001*	.030*	n/a

[6] www.prolific.co

Table 4: The calculated agreement between participants and fictive classifiers per sentence set per condition.

	Calculated agreement			
Condition	**All**	**Set 1, easy**	**Set 2, difficult**	**Set 3, impossible**
(A) *8M_easy*	0.64 (0.07)	0.21 (0.03)	0.91 (0.10)	0.80 (0.13)
(B) *8M_difficult*	0.79 (0.09)	1.00 (0.01)	0.58 (0.18)	0.78 (0.15)
(C) *8M_impossible*	0.86 (0.09)	0.99 (0.03)	0.90 (0.13)	0.67 (0.18)

more instances than both the classifier in A and B. This shows that all classifiers with 73% calculated accuracy were perceived to have a significantly different performance, meaning that the type of prediction mistakes had a significant effect on individuals' perceived accuracy of the classifier.

4.2 Perceived Mistakes

We collected insights into how participants identified prediction mistakes and how they perceived such mistakes. The answers to the open questions Q1 and Q2 (see Section 3.4 "Measures") give an indication of why sentences in the present dataset were difficult to classify for participants. Participants mentioned three sources of difficulty: (1) Vague formulations (e.g., "good material")'. To deal with vagueness, participants developed their own rules throughout the classification task: *"By the end I had decided that I would probably use 'quality' for a jacket"* (P121). (2) Common attributes (e.g., "brand" or "color"): *"on[c]e the word mac was mentioned which could have referred to a mac computer or mac style of jacket"* (P23). (3) Missing context. Some participants commented that the sentences were either too short or not descriptive enough for them to come to a clear decision: *"Where there was no context, it wasn't clear what the sentence was referring to"* (P27). As a consequence of those three sources of difficulties, participants were unsure about the product described in the sentence. However, by the design of our experiment, they were forced to actively agree or disagree with the system's prediction. Participants mentioned that they rather agreed with the system when in doubt: *"Many of the phrases could have been applied to either so I answered affirmatively unless it was obvious that the wording couldn't apply to the category offered"* (P38), and *"There was a few I put yes as just because I felt that it could be either so I would stick with the orig[i]nal classification"* (P121).

To verify this finding, we report the observed agreement of participants with the classifiers per sentence set per condition in Table 4. On the impossible-to-classify sentences (set 3), the classifier in condition C made only two correct predictions. However, the agreement of participants in C on set 3 was 67%, which is higher than the 21% observed in A on set 1. A similar trend can be seen in B on set 2, where the classifier likewise made only two correct predictions. Here, participants' average agreement with the classifier was 58%. As set 2 and 3 contained sentences with higher classification difficulties, participants were more often in doubt about the correct class label and chose to agree with the classifier, leading to a higher agreement with incorrect predictions.

In the open question Q3, participants described why they think the class labels were incorrectly predicted for those sentences. Participants stated that the classifier's prediction mistakes can be ascribed to the same difficulties that humans face (see open questions Q1 and Q2), showing that they expect the classifiers to base their decisions on the same grounds as they did themselves. Some participants also described that identifying prediction mistakes was in itself a difficult task: *"It was just my opinion and not definitely incorrect"* (P37). Another participant concluded that the mistake could have been on their side, showing uncertainty about their own decision: *"it could have been for either so there are some i may have got wrong"* (P100). One participant also criticized the binary classification scheme: *"I felt that it was MORE like one over the other"* (P103). These findings suggest that some participants did not perceive the decisions as binary.

4.3 Traditional Performance Measures

Currently, the performance of classification systems is often evaluated against the ground truth using a range of different measures. Table 5 displays five performance measures (accuracy, precision, recall, F1, ROC AUC) and compares their calculated score with the perceived accuracy score per condition. For all measures (accuracy, precision, recall, F1, ROC AUC), the calculated score differs significantly from the perceived accuracy in all conditions except condition B (where the perceived accuracy is at 73%). We conclude

Table 5: Comparison of traditional measures with perceived accuracy per condition using two-sided Wilcoxon signed-rank tests, asterisk (*) denoting significance at $\alpha = .05$. Showing means and p-values adjusted with Bonferroni correction.

Condition	**Perceived**					**Calculated**					
	Accuracy	**Accuracy**		**Precision**		**Recall**		**F1**		**ROC AUC**	
(A) *8M_easy*	0.66	0.73	<.001*	0.71	<.001*	0.71	<.001*	0.71	<.001*	0.73	<.001*
(B) *8M_difficult*	0.73	0.73	.999	0.71	.667	0.71	.667	0.71	.667	0.73	.667
(C) *8M_impossible*	0.78	0.73	.007*	0.69	<.001*	0.79	.023*	0.73	.007*	0.74	.007*
(D1) *0M_mixed*	0.82	1.00	<.001*	1.00	<.001*	1.00	<.001*	1.00	<.001*	1.00	<.001*
(D2) *0M_clear*	0.95	1.00	<.001*	1.00	<.001*	1.00	<.001*	1.00	<.001*	1.00	<.001*

that neither calculated accuracy nor any of the examined traditional measures are a suitable representation of the perceived accuracy.

4.4 Dataset Composition

Participants also interacted with classifiers that made no mistakes. This was the case in condition D1 on the mixed dataset, containing 30 sentences that are easy, difficult, and impossible to classify, and in condition D2 on the clear dataset with 30 easy-to-classify sentences. To test whether the mere presence of difficult- and impossible-to-classify sentences reduces the perceived accuracy, we compared the perceived accuracies reported in D1 and D2. Table 6 shows that the perceived accuracy of D2 was significantly higher than the perceived accuracy of D1. We, therefore, conclude that the perceived accuracy is significantly influenced by the dataset composition. The presence of difficult- and impossible-to-classify sentences reduced the perceived accuracy, even if no prediction mistakes were made. However, in D2, the perceived accuracy of 95% indicates that the classifier was perceived to make 1.5 incorrect predictions on average. It should be noted that during the classification task, participants had to complete two attention checks that looked similar to the sentences. The sentence text stated that the question is an attention check, while the question (usually asking whether they agreed with the class label) instructed participants to answer "no". It is possible that answering "no" to the attention checks influenced participants' perception, resulting in the perceived average prediction mistake rate of 1.5 for the perfect classifier on the clear dataset. However, the attention check mechanism was constant in conditions, which should equally influence the perceived accuracy in all conditions.

We also compared the perceived accuracy to how often participants agreed with the classifier (see Table 7 for the calculated agreement in D1 and D2). While the agreement shows how often participants actively agreed with the classifier's prediction (binary), the perceived accuracy reflects how accurate participants think the classifier was (potentially continuous). Both metrics are measured as X out of 30. In both conditions D1 and D2, the perceived accuracy is significantly lower than the actual agreement ($p < .001$. for both conditions). Furthermore, although there are no prediction mistakes present in either of the two conditions, the calculated agreement of D1 is significantly lower than the one of D2 ($p < .001$).

Table 6: Means and standard deviations of the perceived accuracy of all classifiers with 100% calculated accuracy. Comparison of conditions using the two-sided Mann-Whitney U-test and p-values adjusted with Bonferroni correction. The asterisk (*) denotes significance at $\alpha = .05$.

Condition	Perceived Accuracy	Comparison p-values	
(D1) *0M_mixed*	0.82 (0.12)	n/a	< .001*
(D2) *0M_clear*	0.95 (0.10)	< .001*	n/a

5 DISCUSSION

The present user study was designed to determine whether different types of prediction mistakes of a classifier equally impact users' perceived accuracy of the classifier. In the remainder, we discuss our findings on the perceived accuracy of classifiers that make different types of prediction mistakes. Moreover, we discuss whether existing performance measures are a suitable representation for the accuracy that users believe to have experienced. Finally, we look at the results of investigating the influence of dataset composition on perceived accuracy and place our findings in a broader context of implications for practitioners and the field of human-centered machine learning.

5.1 Not all prediction mistakes influenced the perceived accuracy equally

We first investigated how accurate participants believed the system to be. In line with our expectations, our results show that even though all three classifiers had the same calculated accuracy, each making eight prediction mistakes on 30 sentences, they brought about significantly different levels of perceived accuracies. Specifically, our results confirm hypothesis [**H1**], because the perceived accuracy in condition A (where prediction mistakes were made on easy-to-classify sentences) was significantly lower than in conditions B (prediction mistakes on difficult-to-classify sentences) and C (prediction mistakes on impossible-to-classify sentences). Our results also confirm hypothesis [**H2**], as the perceived accuracy in condition B was significantly lower than in condition C. To understand why not all prediction mistakes influenced the perceived accuracy equally, we asked participants why they think the classifier made mistakes. In their answers, participants reported that their decision is not always binary ("correct" and "incorrect"). They perceived the decision to be more nuanced (one is "more correct" than the other, but both are "correct"). In some cases, they doubted their own judgement when confronted with a prediction mistake. In the open questions, participants also reported to be inclined to agree with the classifier's prediction when in doubt, i.e., when the difficulty of classification for a human increases. The tendency of agreeing when being in doubt also shows in the results of the calculated agreement. For impossible-to-classify sentences, the agreement with the fictive classifier that made eight incorrect and two correct predictions was at 67%. These findings show that the classifier's (incorrect) predictions impaired participants' decisions for the correct class label, as one would expect to see an agreement of 20%. This means that in cases were sentences were difficult or impossible to classify, participants tended to agree with the classifier when being in doubt – even if the predictions were incorrect. A similar observation has been made by Levy et al. [27], who found

Table 7: Means and standard deviations of the calculated agreement of all classifiers with 100% calculated accuracy. Comparison of conditions using the two-sided Mann-Whitney U-test and p-values adjusted with Bonferroni correction. The asterisk (*) denotes significance at $\alpha = .05$.

Condition	Calculated Agreement	Comparison p-values	
(D1) *0M_mixed*	0.89 (0.08)	n/a	< .001*
(D2) *0M_clear*	0.99 (0.03)	< .001*	n/a

that participants accepted inappropriate predictions in lack of more suitable alternatives. Interestingly, participants justified their behavior by challenging the definition of “correct”. They indicated that for some sentences, both class labels were equally likely and should therefore both be considered correct.

5.2 The calculated measure of accuracy does not accurately reflect perceived accuracy

In hypotheses **H1** and **H2**, we expected and confirmed a significant difference in perceived accuracy between conditions A, B, and C. Since the classifiers in all three conditions make eight prediction mistakes on 30 predictions, hence having a calculated accuracy of 73%, we hypothesized that the perceived accuracy differs significantly from the calculated accuracy in at least one condition [**H3**]. Our results indicate that hypothesis [**H3**] can be accepted, as the calculated accuracy differed significantly from the perceived accuracy in four out of five conditions. Only in condition B (prediction mistakes on difficult-to-classify sentences), no significant difference between calculated and perceived accuracy was shown. We further see in our results that the perceived accuracy in condition A (prediction mistakes on easy-to-classify sentences) was significantly lower (66%) than the calculated accuracy (73%). This indicates an underestimation of the classifier’s performance. Contrarily, in condition C (prediction mistakes on impossible-to-classify sentences), participants over-estimated the performance of the classifier, with the perceived accuracy (78%) being significantly higher than the calculated accuracy (73%). These findings again support the conclusion that the type of prediction mistake (with respect to classification difficulty) influences how users react to a classifier, potentially resulting in under- or overestimation of the actual accuracy. A possible explanation for this observation is the availability bias [56, 59]: Prediction mistakes that were easy to recognize as such (e.g., mistakes on easy-to-classify sentences) are mentally more available than prediction mistakes that were difficult to classify. Clear mistakes might be prevalent when forming an opinion on the classifier’s performance, hence having a stronger influence on perceived accuracy.

For other traditional measures used to evaluate the performance of a system (precision, recall, F1, ROC AUC), we also found significant differences between the calculated score and the perceived accuracy in all conditions except condition B. It should be noted that although accounting for some characteristics of a dataset such as class imbalance (F1) or for a type of outcome, such as fp in precision or fn in recall, all those measures treat all prediction mistakes equally. Furthermore, none of these measures considers how a human perceives individual data points, e.g., with respect to their classification difficulty. We therefore conclude that none of the discussed measures is a suitable representation for perceived accuracy.

5.3 The composition of the dataset itself influences the perceived accuracy

We hypothesized that even a classifier that does not make any prediction mistakes will have a perceived accuracy that is significantly lower than 100% if the dataset contains some impossible-to-classify data points [**H4**]. The statistical analysis of the perceived accuracy in D1 (82%) and D2 (95%) reveals a significant difference, which confirms hypothesis [**H4**]: The mere presence of difficult- or impossible-to-classify sentences reduced the perceived accuracy. Interestingly, participants also thought that the classifiers in both conditions made significantly more mistakes than what they indicated during the interaction (by agreeing or disagreeing with the prediction). This discrepancy might have been caused by sentences where participants were in doubt and therefore had agreed with the prediction, but kept in mind that they were unsure. This uncertainty could have played a role in the retrospective judgement of the classifier’s performance.

5.4 Implications

The implications of our findings are twofold. First, a metric that accounts for classification difficulty is needed to accurately reflect the users’ perception of a classifier’s performance. This measure could either be used during evaluation to draw a more accurate picture of how the performance of a classifier is perceived, or during training to optimize a classifier’s behavior for perceived accuracy. In both cases, a “perception weight” should be assigned to each data point, quantifying the reduction in perceived accuracy that this data point brings about if misclassified. In traditional measures such as accuracy, F1, or ROC AUC, all mistakes have the same weight. The field of cost-sensitive learning [53] has introduced the concept of cost matrices, that could be leveraged to account for individual weights [55] and reduce the gap between system-oriented and user-oriented evaluations [30]. A cost-sensitive evaluation metric with individual “perception weights” might better reflect how a classifier is perceived by the users. While a new metric would help to understand the user’s view on a system, optimizing the system behavior for perceived accuracy should be considered with care, as it holds the potential for manipulating users by fostering overestimation of the classifier’s performance, possibly leading to over-reliance, inappropriate user trust, and deception. Previous works have made first steps towards human-centered measures for machine learning systems, but have either focused on an aggregated view rather than on differences between individual data points [23, 45] or present a theoretical concept that was not tested in user studies [19].

Second, our findings on dataset composition indicate that the dataset itself plays an important role in how a user perceives a classifier’s performance. If practitioners want to understand how their system will be perceived by users, it is not enough to only focus on how many mistakes a system makes. They need to also consider the dataset composition: The mere presence of classification difficulty reduces the perceived accuracy, potentially leading to a significant underestimation of the performance. As a consequence, users could lose trust in the system’s capabilities. Dietvorst, Simmons, and Massey [9] found that an underestimation can lead to ignorance of the system’s suggestions, even in case that the system outperforms human accuracy. In addition to utilizing methods to enhance ground truth robustness (see [1, 16, 43, 61]), we encourage practitioners to also investigate a dataset’s composition concerning the classification difficulties of its data points. It is important for practitioners to understand how a system will be perceived by users when in production, which is why we suggest including data point classification difficulty (from a user's point of view) in the assessment of machine learning systems in practice.

6 LIMITATIONS AND FUTURE WORK

The study is subject to several limitations. First, by asking for the agreement or disagreement on each prediction, participants were forced to form an opinion for each sentence. Although this should make it easier for participants to assess the accuracy of the system at the end of the interaction, it might differ from how users would use a decision-support system in the real world. Replicating our study without asking for an explicit agreement or disagreement for each sentence could deliver additional insights into the opinion-forming process of users and the generalizability of our findings.

Second, some participants described the classification decision as being non-binary. In our study, however, participants were forced to give a binary decision ("yes" or "no") rather than indicating their level of agreement. Based on our findings, we suggest that future studies consider using interval scales to indicate the level of agreement rather than a binary decision.

Third, we have used only one use case in our study. To broaden the picture of how classification difficulty impacts the perceived accuracy in various domains, future research should investigate additional textual datasets, e.g., spam detection or news articles. Those use cases offer data points with varying classification difficulties, as well as different type of mistakes: Authors of spam emails continuously find new ways of disguising their messages (e.g., writing "biitcoiin" instead of "bitcoin") that do not appear in the training data, leading to obvious mistakes of a model on easy-to-classify data points. The transferability of our findings to other data types, e.g., images or audio data, should also be investigated. However, future studies should ensure that the ground truth was, similar to our setup, already gathered during data collection and should avoid datasets that are only retrospectively annotated.

Fourth, by not disclosing to the users where the predictions stemmed from, we aimed to avoid priming the judgement of our participants. Introducing either a system or a human expert that produces the predictions could result in over-reliance or under-reliance [11]. It could be interesting to repeat the experiment with an altered task introduction, explicitly mentioning either a human expert or a machine learning system.

Finally, our study also revealed that participants tended to agree with the system when in doubt, i.e., when the difficulty of decisions increased. To develop a full picture of the agreement behavior, future studies are needed that explore the influencing factors of decisions taken under uncertainty.

7 CONCLUSION

In this paper, we investigated how the individual classification difficulty of data points influences perceived accuracy of a classifier in a binary classification task. We conducted an online user study with 225 participants who received 30 predictions of one out of five fictive classifiers and collected the perceived accuracy from participants. We investigated three systems with equal accuracy (73%) that made mistakes on different types of data points (easy-to-classify, difficult-to-classify, impossible-to-classify). Additionally, we compared perfect classification performance at 100% accuracy on a dataset with only easy-to-classify cases to a dataset with mixed data points. Our findings show that (1) mistakes do not equally impact perceived accuracy (e.g., mistakes on impossible-to-classify data points result in a higher perceived accuracy than mistakes on easy-to-classify data points), (2) the perceived accuracy can significantly differ from calculated accuracy, and (3) the composition of the dataset itself (in terms of the classification difficulty of the data points) has an impact on perceived accuracy. We suggest that predictions on user acceptance of a system should therefore not be based on traditional accuracy metrics such as precision, recall, or the F1-score, as they are not an accurate representation of how users perceive the accuracy of a classification system.

ACKNOWLEDGMENTS

This work was partly funded by the DFG, grant no. 388815326; the VACOS project at GESIS.

REFERENCES

[1] Görkem Algan and Ilkay Ulusoy. 2021. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems* 215 (2021), 106771. https://doi.org/10.1016/j.knosys.2021.106771

[2] Oscar Alvarado and Annika Waern. 2018. *Towards Algorithmic Experience: Initial Efforts for Social Media Contexts.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173860

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[6] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 7 (1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA.* IEEE Computer Society, Miami, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[9] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (2018), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[10] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-18).* Association for the Advancement of Artificial Intelligence, Zürich, Switzerland, 12–20. the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-18) ; Conference date: 05-07-2018 Through 08-07-2018.

[11] Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors* 44, 1 (2002), 79–94. https://doi.org/10.1518/0018720024494856 arXiv:https://doi.org/10.1518/0018720024494856 PMID: 12118875.

[12] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) *(IJCAI'01).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 973–978.

[13] Sophie Emerson, Ruairí Kennedy, Luke O'Shea, and John O'Brien. 2019. Trends and applications of machine learning in quantitative finance. In *8th international conference on economics and finance research (ICEFR 2019).* Social Science Research Network, Lyon,France, 0–9.

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[15] Stephen H. Fairclough, Alexander J. Karran, and Kiel Gilleade. 2015. *Classification Accuracy from the Perspective of the User: Real-Time Interaction with Physiological Computing.* Association for Computing Machinery, New York, NY, USA, 3029–3038. https://doi.org/10.1145/2702123.2702454

[16] Benoit Frenay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

[17] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. *Mental Models of AI Agents in a Cooperative Game Setting.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376316

[18] Timothy Robin Gibson. 1993. *Towards a discourse theory of abstracts and abstracting.* Department of English Studies, University of Nottingham, Nottingham, United Kingdom. 306–399 pages.

[19] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[20] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[21] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3511–3522. https://doi.org/10.1145/3025453.3025781

[22] Tharindu Kaluarachchi, Andrew Reis, and Suranga Nanayakkara. 2021. A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning. *Sensors* 21, 7 (2021), 1–29. https://doi.org/10.3390/s21072514

[23] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. *How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy.* Association for Computing Machinery, New York, NY, USA, 347–356. https://doi.org/10.1145/2702123.2702603

[24] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[25] Matjaz Kukar and Igor Kononenko. 1998. Cost-Sensitive Learning with Neural Networks. In *13th European Conference on Artificial Intelligence, Brighton, UK, August 23-28 1998, Proceedings.*, Henri Prade (Ed.). John Wiley and Sons, Brighton, United Kingdom, 445–449.

[26] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3075–3084. https://doi.org/10.1145/2556288.2557238

[27] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3411764.3445522

[28] Xiao Li, Ye-Yi Wang, and Alex Acero. 2008. Learning Query Intent from Regularized Click Graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 339–346. https://doi.org/10.1145/1390334.1390393

[29] Cynthia C. S. Liem, Markus Langer, Andrew Demetriou, Annemarie M. F. Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph. Born, and Cornelius J. König. 2018. *Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening.* Springer International Publishing, Cham, 197–253. https://doi.org/10.1007/978-3-319-98131-4_9

[30] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[31] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[32] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. *Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[33] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/2766462.2767755

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. https://doi.org/10.1145/3457607

[35] Rishabh Misra. 2018. *News Category Dataset.* ResearchGate. https://doi.org/10.13140/RG.2.2.20331.18729

[36] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. https://doi.org/10.1145/3411764.3445402

[37] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press, New York, United States.

[38] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. AAAI Press, Palo Alto, USA, 97–105.

[39] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books, New York, USA.

[40] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. (2019). https://sites.google.com/view/xai2019/home IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019, XAI 2019 ; Conference date: 11-08-2019 Through 11-08-2019.

[41] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. 2021. Dataset of Natural Language Queries for E-Commerce. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 307–311. https://doi.org/10.1145/3406522.3446043

[42] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics* 7 (11 2019), 677–694. https://doi.org/10.1162/tacl_a_00293

[43] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human Uncertainty Makes Classification More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, New Jersey, USA, 9616–9625.

[44] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[45] Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. 2021. Survey Equivalence: A Procedure for Measuring Classifier Accuracy Against Human Labels. arXiv:2106.01254 [cs.LG]

[46] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. *Automation Accuracy Is Good, but High Controllability May Be Better.* Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290605.3300750

[47] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[48] Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. *Studying the Effects of Cognitive Biases in Evaluation of Conversational Agents.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376318

[49] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 614–622. https://doi.org/10.1145/1401890.1401965

[50] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[51] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. *No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[52] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Germany, 1015–1021.

[53] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. 2010. Cost-sensitive learning methods for imbalanced data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, New Jersey, USA, 1–8. https://doi.org/10.1109/IJCNN.2010.5596486

[54] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, virtual, 9625–9635. http://proceedings.mlr.press/v119/tsipras20a.html

[55] Peter Turney. 2000. Types of cost in inductive concept learning. (2000). https://arxiv.org/abs/cs/0212034 Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning, (2000), Stanford University, California, 15-21.

[56] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5, 2 (1973), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

[57] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[58] Viswanath Venkatesh and Fred D. Davis. 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46, 2 (2000), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926 arXiv:https://doi.org/10.1287/mnsc.46.2.186.11926

[59] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. *Designing Theory-Driven User-Centric Explainable AI*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[60] Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer. 2018. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM transactions on computational biology and bioinformatics* 15, 6 (2018), 1968–1978.

[61] Jing Wang and Xin Geng. 2019. Classification with Label Distribution Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, Palo Alto, USA, 3712–3718.

[62] Xinxi Wang and Ye Wang. 2014. Improving Content-Based and Hybrid Music Recommendation Using Deep Learning. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/2647868.2654940

[63] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence. Ed. by PELILLO, M. and SCANTAMBURLO* 2021 (2021), 224.

[64] Jin Xiao, Xu Zhou, Yu Zhong, Ling Xie, Xin Gu, and Dunhu Liu. 2020. Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. *Knowledge-Based Systems* 189 (2020), 105118. https://doi.org/10.1016/j.knosys.2019.105118

[65] Ying Xu and Mark Warschauer. 2020. What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376416

[66] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[67] Shuo Yang, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. 2017. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. *Knowledge-Based Systems* 136 (2017), 37–45. https://doi.org/10.1016/j.knosys.2017.08.017

[68] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. *Understanding the Effect of Accuracy on Trust in Machine Learning Models*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[69] Xingquan Zhu and Xindong Wu. 2004. Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. *Artif. Intell. Rev.* 22, 3 (Nov. 2004), 177–210.