The Dissertation Committee for Wanxue Dong
certifies that this is the approved version of the following dissertation:

Assessing Workers' Decision Quality with Scarce Ground Truth Data

Committee:

Maytal Saar-Tsechansky, Supervisor

Maria De-Arteaga

Tomer Geva

Anitesh Barua

Assessing Workers' Decision Quality with Scarce Ground Truth Data

by

Wanxue Dong

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2023

Acknowledgments

I would first like to thank you my entire family for unwavering love and support in the PhD journey. I could not have done this without you always being here for me.

My advisor, Maytal, deserves my most heartfelt thanks. You introduced me the joy of doing research, guided me with insightful advice, taught me how to be a good researcher and teacher, and imparted to me many valuable life licenses.

My thanks and appreciation also go out to my collaborators, Maria and Tomer, and many fellow and friends, faculty members, staff in the IROM at McCombs for helping me in navigating the ups and downs of my research.

Lastly, I want to express my deepest gratitude to my heavenly Father, dear Lord, for bringing out the best in me, bringing my loved ones and brothers and sisters in Christ to my life, and for all You blessed me with. You are my strength, strength like no other.

. . .

Assessing Workers' Decision Quality with Scarce Ground Truth Data

Publication No. ___________________

Wanxue Dong, Ph.D.
The University of Texas at Austin, 2023

Supervisor: Maytal Saar-Tsechansky

Accurately assessing workers' decision quality is fundamental for management, and the efficiency of expert and crowd-sourcing markets. This paper establishes novel ML and AI methods to accurately evaluate workers' decision accuracy and bias with scarce ground truth (GT or gold standard GS) data, and to further improve accuracy assessment through costly effectively acquiring GT if given an acquisition budget. Without the proposed methods, assessing workers' decision quality typically requires GT data to compare with workers' noisy decisions. However, GT is often prohibitively costly to acquire for even a small fraction of each worker's decisions. For example, physicians may determine a diagnosis and initiate a treatment, yet the correct decision, such as the one that can be established by a panel of physicians. Consequently, in practice, there is often poor transparency regarding physicians' decision quality. In my dissertation, I collaborating with my coauthors developed the groundwork for achieving scalable and inexpensive assessments of workers' decision accuracy and bias. The empirical results show that the decision accuracy assessment with very limited GT improves the best available approach by 60% to 93%; my bias assessment produces either comparable to or outperforms the commonly used existing approach; my cost-effective GT acquisition strategy applied in Amazon Mechanical Workers' accuracy assessment achieves the same performance only using 1/3 of the GT or improve the assessment by 24%. All proposed methods have significant implications in many impactful domains including health care, fraud detection, fact checking, and online labor markets. The methods proposed in this dissertation address the problem of estimating workers' decision accuracy and bias from historical data with scarcely available ground truth, and achieve the state of the art performance. This dissertation lays the groundwork towards increasing transparency in workers' (sources') decision quality.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Assesing Workers' Decision Quality with Scarce Ground Truth Data

My dissertation research develops novel ML methods to evaluate workers (decision markers) decision quality when there is limited ground truth and existing methods fail to achieve good performance that can be relied on in practice. My research develops new methods that aim to reliably assess (i) experts' decision accuracy (ii) experts or crowd-sourcing workers' societal bias, and (iii) to improve experts' accuracy assessment through costly effectively acquiring ground truth data.

The Chapter 2 and 3 combined establish a machine learning-based framework towards assessing experts' decision accuracy motivated by the goal to accurately and reliably estimate experts' decision accuracies, such as the accuracy of physicians' diagnoses, when ground truth on the correct decisions is scarce, and existing methods, which rely on ground truth, thereby fail[1]. Given experts make non-trivial and consequential decisions, experts' decision accuracy is a fundamental aspect of their judgment quality and is thereby es-

[1]A joint work with Maytal Saar-Tsechansky and Tomer Geva, "A Machine Learning Framework for Assessing Experts' Decision Quality."

sential to both effectively manage experts' resources as well as for consumers' choices who seek experts' advice. In spite of the crucial role of such assessments, experts' decision accuracies are rarely known because of the scarcity of ground truth necessary to achieve these assessments by existing approaches. My work developed innovative machine-learning methods that overcome this challenge, and achieve state-of-the-art performance.

Specifically, my dissertation research developed a novel machine-learning algorithm to estimate experts' decision accuracy by effectively leveraging both abundant historical data on experts' past (noisy) decisions and scarce decision instances with ground truth (GT). This work conducted extensive empirical evaluations of the method's performance relative to alternatives using both benchmark data sets, and a purposefully compiled dataset on human workers' decisions. Given the applied nature of the goals and contexts considered in this research, estimating the benefits of the method entails extensive evaluations that consider a wide array of practical scenarios. My evaluations establish that the method achieves state-of-the-art performance. This is the first work to posit and address the problem of estimating experts' decision accuracies from historical data with scarcely available ground truth, and it is the first to offer comprehensive results on the accuracies that can be achieved across settings. Overall, given the consequences of (in)correct decisions in fields such as healthcare and security, making technology available to ascertain decision accuracy – reliably, cheaply, and at scale – is an important step towards invaluable decision quality evaluation.

Given ground truth or gold standard is costly to acquire, given an acquisition budget, it is valuable to develop an AI method that can cost effectively acquire the labels of instances that can improve the assessment of experts' decision accuracies or labelers biases the most. In Chapter 3, I address this research challenge, which is to cost-effectively acquire GT (or GS) labels for improving experts' accuracy assessment. Ultimately, for a given acquisition budget, the proposed algorithm aims to acquire labels for particularly informative instances that will improve the assessment of experts decision accuracy the most[2].

Chapter 5 considers a different dimension of workers' decision quality, decision bias. It introduces a new ML-based method, leverages very limited GT data to assess relative (societal) biases in human-generated labels/sources. Societal biases encoded in human decisions (assessments or labels) have been highlighted as an important source of algorithmic unfairness. Thus, assessing workers' decision biases is crucial for assessing the usability of human-generated labels for training ML models, and mitigating the risk of encoding decision makers' biases in algorithmic predictions. Yet, the most prominent metric that relies on statistical parity, the Selection Rate (SR), is not a reliable method because it does not consider the relationship with a GT or gold standard to assess bias. This work develops a novel and principled machine learning method to accurately assess the relative extent of bias contained in labels produced by different labelers (or different sources, more broadly), when

[2]This work is closely advised by Maytal Saar-Tsechansky.

gold standard labels are scarce given that they are costly or difficult to acquire. This work provides theoretical guarantees and empirically demonstrate that the method outperforms the commonly used alternative, SR, which may be misleading when humans make decisions (intentionally or unintentionally) that aim to game the assessments. The proposed approach lays the groundwork towards reliable bias assessment in labeling and offers an important building block towards mitigating algorithmic bias stemming from biased labels[3].

For furture research, I plan to build on my dissertation research to improve human-AI collaborations that rely heavily on correct assessment of human decision accuracies and biases. Presently, most prior work assumes such assessment can be reliably produced from historical data; however, as discussed above, in many critical domains, such as medicine, the historical data rarely include ground truth. I therefore aim to explore how integration of the methods I developed can impact our ability to better leverage human-AI complementarities.

1.2 Dissertation Guide

Chapter 2 describes a ML-based framework to estimate experts' decision accuracy by effectively leveraging both abundant historical data on experts' past (noisy) decisions and scarce decision instances with ground truth (GT).

[3] The work closely follows Wanxue Dong, Maria De-Arteaga and Maytal Saar-Tsechansky, "A Machine Learningbased Framework towards Assessment of Labelers' Biases."

Chapter 3 extends the study in Chapter 2 including an advanced methodology to assess experts' decision accuracy. The method proposed in this chapter ensures that with respect of the number of the ground truth, it can produce reliable estimation of the experts decision accuracy.

Chapter 4 introduces a cost-effectively sampling strategy to acquire GT if given a limit acquisition budget.

Chapter 5 describes a new ML-based method, leveraging very limited GT (or GS) data to assess relative (societal) biases in human-generated labels/sources.

Chapter 2

Assessing Experts' Decision Accuracy with Scarce Ground Truth

2.1 Introduction

Across key domains, human expert assessments and crowd annotations are essential for labeling data to train machine learning models, and constitute a pathway through which human's biases are learned by algorithms. Once deployed, biased Machine Learning (ML) algorithms can have significant impact in human's lives in many realms, including healthcare, recruitment, promotion, and colleague admission, among others. In this research, we explore how to leverage scarce GT decisions (labels) to assess biases in human-generated labels. We propose a machine learning-based framework to produce a relative assessment of the extent of bias contained in labels produced by different labelers or sources, when GT labels are costly or difficult to acquire and thus available for only a small set of instances. For example, gold-standard labeled instances can be acquired from costly professional fact checkers examining online claims' veracity to constitute a gold-standard when assessing crowdsourced labels. The proposed methodology does not require overlap between the instances assessed by different labelers nor between these and the instances for which GT labels are available. After providing theoretical guarantees, we empirically show that our method outperforms or produces at least comparable results to several existing alternatives to assess biases present in human labels, including a commonly used benchmark relying on statistical parity, which we show may be misleading when humans (intentionally or unintentionally) produce poor quality orderings within protected groups. Our empirical results establish the performances that can be achieved across diverse settings, including settings that involve different data domains, labelers' (sources') biases, class or group distributions, and amounts of GT data. We also show the downstream value of our approach in improving the quality of ML algorithms induced from biased labels. The proposed approach lays the groundwork towards increased transparency in labelers' biases and offers an important building block towards mitigating algorithmic bias stemming from biased labels.

2.2 Related Work

To our knowledge, no prior work has addressed the problem we consider nor offered comprehensive results on computational, scalable and inexpensive estimation of experts' decision accuracies. In this section, we discuss how different streams of prior work relate to the contributions we present here.

The most common practice for assessing worker decision quality when ground truth is scarce has been the use of traditional peer/human evaluations [70], such as peer or committee-based evaluations [191]. A large body of work over several decades has suggested and analyzed human-based approaches, such as human-based, relative performance rating and pairwise ranking [160]; exploring the correlations between workers' reviews to identify inconsistencies [93]; and examining the evidence of rating reliability and validity [219]. However, given experts' time and effort are costly, extensive engagement of such experts to evaluate their peers decisions in a continuous fashion is prohibitive.

2.2.1 Machine Learning-Based Evaluation

Recent machine learning research has considered problems involving human and expert workers. However, most of these works considered problems and settings that differed meaningfully from those we consider here. In particular, a significant stream of work considered the problem of improving the accuracy of data labels obtained from multiple annotators, such as crowd workers [e.g., 40, 41, 181, 234, 216], as well as expert workers [e.g., 229]. Unlike our focus on *costly experts*, most of these works focused on inexpensive workers who perform simple, intuitive tasks and in markets characterized by inexpensive, non-expert workers [118]. Importantly, works in this stream of research have focused on methods that consider repeated labeling, in which multiple workers evaluate the same data instance and where the likely ground truth is inferred by aggregating multiple labels [e.g., 45, 234, 229, 40, 41, 117, 181, 192, 221, 216].

Other works relate to our research because they consider predicting or assessing workers' current or future performance, but consider settings in which relevant ground truth is always available, or consider other challenges than assessing experts' decision accuracy. For example, [125] considered predicting future work performance based on the workers' performance history in a different domain; and [124] developed a method for predicting workers' skill-set-specific reputation scores in a dynamic setting. [33] propose a scalable approach for technical-skill *testing* of workers, involving scalable generation of effective test tasks/questions, based on which workers' technical skills are assessed and for which ground truth is known. Other methods considered learning predictive models from noisy (e.g., human) labels (or decisions) but did not develop methods to assess decision makers' decision accuracy with limited ground truth. For example, [21] and [49] aimed to improve model learning by removing mislabeled instances. Several other works considered the cost-effective acquisition of noisy labels (typically produced by imperfect human labelers) from which to learn accurate predictive models (e.g., [98], [85], [80]).

Recent works [115, 204] aimed to improve model learning from noisy labels and estimate labelers' accuracies, simultaneously. While these methods did not consider how to bring to bear limited ground truth to assess workers' decision accuracy, they can apply to estimate workers' accuracies in our setting.[1] [204] showed that their approach is superior to the one proposed by [115], and we thus empirically compare our approach to it. Specifically, [204] proposed minimizing the loss for models that accommodate a cross-entropy loss function and included a regularization term based on labelers' estimated

[1]Both works also consider the use of repeated labeling, but this scenario is not applicable in our expert setting.

accuracies. Our approach is distinct from the method proposed by [204] by two key elements. First, our approach is designed to leverage scarce ground truth, and the method by [204] does not take advantage of such data. Second, our approach is model/domain-agnostic; it allows using the model induction algorithm most suitable for the underlying expert data domain. In contrast, [204] consider models with a cross-entropy loss function, and the method is thus only applicable to data domains where such models are suitable. Consequently, the worker assessment produced in [204] does not yield competitive performance in the setting we consider in this paper: we show that even for settings where our approach has the least relative advantage, with a minimal number of ground truth labels, our method yields superior assessments of experts' accuracies.

Finally, related work on which we build [83, 84] proposed the problem of *ranking* expert workers according to the quality of their decisions in the absence of ground truth decisions. However, this work did not address the problem of estimating workers' decision accuracies and considered different settings than the settings we focus on here. In particular, ranking is a fundamentally different task than estimating workers' accuracies, and achieving it serves different practical goals. While ranking aims to position workers relative to others within a cohort, unlike an estimation of an expert's absolute decision accuracy, ranking cannot be used to establish whether a given worker meets a certain performance requirement or expectation, to optimally assign workers to tasks, or to determine whether there are practically meaningful gaps between workers' decision accuracies, which are integral to inform retention and compensation decisions. The method we develop here offers novel means to reliably estimate experts' decision accuracies, and it produces state-of-the-art estimates unmatched by existing alternatives.

2.2.2 Experts' Decision-Making Errors

Research regarding the causes for experts' errors spans over multiple decades and covers various aspects of experts' decision making. Prior literature has identified that different experts have inherently differential overall expertise [231, 211]; thus, different experts exhibit different accuracy rates. Many of the experts' decision errors are outcomes of inherent and contextual factors. Inherent factors reported in the literature are based on the expert's individual abilities and affect the quality and accuracy of experts' decisions [57, 99]. These factors include the expert's ability to perceive large, meaningful, and easily-neglected patterns [57, 99]; the ability to think fast and to effectively characterize or represent a problem [57, 99]; the ability to make decisions without requiring conscious initiation or sufficient time to think through the situation [57]; and a prolonged experience through practice and education [99]. These factors generally result in experts' making fewer errors than novice or less talented performers [57, 99]. Nevertheless, such skill, vigilance, and conscientiousness were found to be essential but not sufficient to prevent errors because of experts' cognitive biases and limitations [52]. In effect, the extent to which experts are inherently prone to cognitive biases and limitations may increase the likelihood of errors stemming from ineffective use of information and from the way experts generate mental models from such information [190]. Perhaps not unexpectedly, experts were shown to be prone to suffer from many of the cognitive limitations that affect humans, more broadly [105]. [38] offer a comprehensive review of such limitations. For example, experts were found to rely on mental heuristics rather than fully using available information [106] and to be prone to various biases, including confirmation bias [168], anchoring bias [209], and availability bias [208].

Contextual factors were also reported to affect experts' errors [190]. Among the contextual (e.g., task-related) factors that increase the likelihood of experts' errors, the literature identifies rapid response time requirements [52, 119]; task complexity [19]; financial incentives [30]; lack of appropriate technology or instrumentation [55]; limited access to information and analysis or being provided ambiguous information [18, 119, 237]; exposure to extraneous information [52]; variations in task demands (e.g., requested by supervisors); and social/organizational influences [52], including whether other experts were involved in the decision process [237], lack of feedback, the extent to which the goal is well defined, and the need to collaborate with other individuals [119]. In addition, expert errors were found to be driven by psychological-specific reasons caused by the surroundings, such as fatigue [143], distractions, excessive workload, and time pressure [56, 58, 89, 170], as well as their own emotional state [69]. Examples include judges that were observed to issue harsher decisions just before their lunch break [42] and physicians working night shifts who were significantly affected and were more likely to neglect some portion of standard procedures [195]. All of these factors result in large variations in experts' decision errors [129, 220, 170, 201].

Crucial to this work are both the data availability and the subsequent ability to infer an expert's decision accuracy. The fundamental phenomenon of an expert's inherent ability, which is invariant across decisions and that affects the expert's decision accuracy across instances,is unknown. This is a fundamental aspect of experts' decision performance and is not observed. Furthermore, key contextual factors, which may vary over time, such as fatigue, distractions, social influences, hunger, or emotional states, can further compound an expert's performance. Thus, given an expert's unknown, inherent ability to yield correct decisions, research has documented that the expert's performance can decline due to contextual factors. Importantly, such contextual factors are, in practice, rarely documented so as to be associated with the relevant decisions that experts make. In addition, such information is difficult to recover retrospectively, or it otherwise may require intrusive and expensive collection procedures. Together, an expert's (unknown) inherent ability and any unobserved contextual factors that compound it are such that do not allow to reliably predict across contexts the event of an error in a given instance. The approach we develop here does not aim to do so, and ,thus, does not rely on the availability of contextual information to produce estimations of experts' decision accuracies.

2.2.3 Limited Ground Truth

We consider common expert settings in which ground truth about the correct decision is costly to acquire and thus scarce. As such, our work is distantly related to research on model induction from scarce ground truth in the machine learning literature. However, this literature does not consider our problem, and often consider data with meaningfully different properties. Specifically, weakly supervised learning [235] consider the task of inducing model arising in contexts with limited ground truth. For example, *incomplete supervision* assumes a small amount of correctly labeled data is available along with abundant *unlabeled* data. Model learning in such settings has been typically handled using semi-supervised learning approaches [25] or by active learning-based approaches which accommodate acquisition of additional labels [187]. *Inaccurate supervision* considers the case where available labels are not all correct. This problem typically is handled by methods that aim to learn a model from the given noisy data and then use the model to correct or eliminate incorrect labels [21]. Another distantly related problem is the *cold start* problem in recommender systems [185], where there are limited data about items' ratings, users' characteristics, or users' past preferences. This problem is often handled in practice by using simple models that are less likely to overfit the data.[2]

[2]Few-shot learning [217] is another related stream of work which considers inferences from limited training data and selecting the most likely class from a set of "query" classes, even if the relevant class has not been observed in the training data.

Our work differs meaningfully from the above streams of works which do not consider or can apply directly to address our problem of estimating experts' decision accuracies. Specifically, a predominant element of our methodology is to build on models' inferences to produce accurate assessments of experts accuracies, which work on learning from limited ground truth as not considered. However, inference based on such approaches can be used to infer the likely ground truth, and, based on which, experts can be evaluated. In Section 3.2.3.1 we report comparisons of our approach to such alternatives. We show that direct use of such methods to infer the ground truth, and which does not address the challenges in our problem setting, significantly under perform the approach we develop here.

2.3 Problem Formulation

We consider a set of K expert workers $W = \{W_1, ..., W_K\}$, where each routinely makes multiple decisions and where decisions made by different workers are drawn from the same distribution. For example, workers may be auditors who decide whether a given tax return claim is fraudulent or radiologists who decide whether a patient's image exhibits a certain malady. (Henceforth, we use the terms *expert workers*, *workers*, and *experts* interchangeably). We consider a challenging setting that arises often in practice, where each decision instance, such as a particular patient's diagnosis, is made by a single expert, so that the sets of decisions made by each expert are mutually exclusive. For a given expert worker, W_k, historical data about n_{w_k} past decisions are available, and where instance feature values arriving from distribution $\mathcal{X}$ is available and given by $S_{\mathrm{W}_k} = \{X_i^k, \hat{Y}_i^k\}_{i=1}^{n_{\mathrm{W}_k}}$. For each decision instance i, historical data include the worker's decision $\hat{Y}_i^k \in \{0, 1\}$ (e.g., whether or not the patient has a tumor) along with a feature vector $X_i^k \sim \mathbb{P}(\mathcal{X})$, reflecting feature values for the decision instances, such as various lab blood-test results and symptoms. Note that X_i^k does not necessarily correspond to the full set of holistic information that was available to worker W_k, which may be either structured or unstructured or both. Rather, it includes a set of feature values that are retrospectively available and may include either a subset or a superset of the information available to the expert worker.

For each worker W_k, we seek to assess the worker's decision accuracy, given by $q_{\mathrm{W}_k} = (\sum_{i=1}^{n_{\mathrm{W}_k}} I[Y_i^k == \hat{Y}_i^k])/n_{\mathrm{W}_k}$, where Y_i^k is the ground truth (correct) decision, and I is the truth function, such that $I[\cdot] = 1$ if $(\cdot)$ is true, and $I[\cdot] = 0$, otherwise.

Table 2.1: Key Notations

Notation	Description
B_k	A single base model, mapping : $X^k \rightarrow \hat{Y}^k$, which is trained on worker W_k's decision instances S_{W_k}
$B_j(X_i^k)_z$	Base model B_j's probability estimate that X_i^k maps to class z
$Conf_{X_i^k}$	The confidence in the ensemble model M's prediction for instance X_i^k
DQ_{W_k}	The decision quality score for the worker W_k; it corresponds to the ordinal ranking of workers.
$f : DQ \rightarrow q$	A learned mapping between a worker's DQ to the worker's decision accuracy
$GT = \bigcup_{k=1}^{K} GT_k$	The union of decision instances with ground truth information of all workers in W
M	Ensemble model M
n_{W_k}	Number of decisions made by expert worker W_k
q_{sw_i}	Decision accuracy of a synthetic worker's decision set S_{sw_i}
q_{W_k}	True decision accuracy for worker W_k
$S_{\mathrm{W}_k} = \{X_i^k, \hat{Y}_i^k\}_{i=1}^{n_{\mathrm{W}_k}}$	Worker W_k's decision data
S_{sw}	Decision data reflecting synthetic worker sw
$W = \{W_1, ..., W_k\}$	Set of expert workers to be evaluated

In this work, we consider a challenge arising in many expert environments, where ground truth information, such as decisions produced by a panel of experts, are costly and can thus be acquired for only a scarce subset of decisions made by each expert worker. Specifically, the set of decisions with ground truth for worker W_k is given by $GT_k = \{X_i^k, Y_i^k\}_{i=1}^{t_{\mathrm{W}_k}}$, where for all instances $X_i^k \in GT_k$: $X_i^k \sim \mathbb{P}(\mathcal{X})$, $GT_k \subseteq S_{\mathrm{W}_k}$, and where ground truth data are scarce, that is, $|GT_k| \ll |S_{\mathrm{W}_k}|$. Table 2.1 summarizes key notations used throughout the paper.

2.4 Machine-Learning-Based Decision Quality Estimation (MDE)

In this section, we outline our approach for addressing the problem above: the **M**achine-learning-based **D**ecision quality **E**stimation (MDE) method with only limited GT.

The MDE approach is a machine-learning-based approach to estimate experts' accuracy that exploits the large amount of data available on the experts' decisions, along with (scarce) ground truth information. MDE is detailed in Algorithm block MDE.

When ground truth is scarce, MDE aims to effectively leverage the large number of noisy decisions by expert workers, along with scarce ground truth information, to infer expert workers' decision accuracies. In principle, one can trivially compute the rate of correct decisions for each expert worker based on the accuracy rate for the expert worker's past decisions with ground truth. However, when ground truth data are known for only a handful of each worker's decisions, the accuracy of this trivial assessment is poor. Meanwhile, for settings in which no ground truth information is available and expert workers' true accuracies are unknown, prior work proposed a Decision Quality (DQ) score and showed that ranking decision makers by their respective DQ scores yields a ranking similar to the workers' ranking based on their true (and unknown) decision accuracy rates [84]. However, and importantly, the DQ scores do not correspond to decision accuracy estimates—that is, they do not reflect a worker's rate of correct decisions. Thus, in this setting, computing an expert's frequency of correct decisions based on ground truth instances yields a poor estimate of the expert worker's decision accuracy, and prior work's DQ scores yield a good ranking but do not reflect expert workers' accuracy rates. In light of these challenges, the first element of our approach, MDE, offers a computational framework that allows us to effectively leverage both the DQ scores and the limited ground truth to produce estimates of expert workers' decision accuracies.

In particular, MDE relies on two key notions. First, workers' DQ scores can be computed without ground truth and have been shown to correlate with workers' true accuracies. Consequently, if the true accuracy, q, of some workers was somehow known, it would be possible to produce a set of (DQ, q) pairs, from which it is possible to induce a mapping between a worker's DQ score and the worker's decision accuracy, $f : \text{DQ} \rightarrow q$. Such mapping could be subsequently applied to infer the decision accuracies of workers whose true decision accuracies are unknown.

The second notion that MDE builds on aims to overcome the challenge of producing the (DQ, q) pairs from which a mapping between a worker's DQ score and accuracy can be learned. In particular, to induce a correct mapping, both the q values (representing accuracies) should be correct, and sufficient (DQ, q) pairs should be available from which to reliably learn the mapping. However, in our setting, a worker's true decision accuracy, q, is unknown. (As discussed, there are also no existing approaches that can reliably estimate the worker's decision accuracy, given scarce ground truth.)

To address this challenge, we propose an approach to exploit the available scarce ground truth in a novel way to produce a large set of (DQ, q) pairs, where q is the true accuracy, rather than a noisy estimate. Specifically, we propose an approach that includes three elements: (1) Our approach first coalesces historical decision instances with ground truth from all the experts to compile a data set of ground truth instances; this data set is then used to generate a large number of *"synthetic workers"* with known, predetermined q values (decision accuracies); (2) our approach then produces DQ scores for all the synthetic workers and learns a mapping $f : DQ \rightarrow q$ from the (DQ, q) pairs; and (3) the mapping can then apply to infer any given expert's decision accuracy from the expert's DQ score. In the following subsections, we discuss and outline each of these elements in turn.

2.4.0.1 Producing decision data for synthetic workers with predetermined accuracies.

MDE first compiles a data set of ground truth decision instances that is the union of all decision instances for which ground truth is available from all experts: $GT = \bigcup_{k=1}^{K} GT_k$. GT is then used to produce a set of semi-synthetic decision data sets, $S_{sw} = \{S_{sw_i}\}_{i=1}^{m}$, where each S_{sw_i} corresponds to an individual *synthetic worker*'s decision set and reflects q_{sw_i}, a predetermined decision accuracy rate ($0.5 \leq q_{sw_i} \leq 1$).[3] Importantly, each set S_{sw_i} contains all the instances in GT, and the predetermined decision accuracy q_{sw_i} is produced by flipping the (correct) labels of $1-q_{sw_i}$ proportion of instances, drawn uniformly at random from GT, thereby creating a $1 - q_{sw_i}$ proportion of incorrect decisions. In this procedure, we specify that synthetic workers' accuracies (q_{sw_i} values) will differ by a fixed (small) interval *intv* within the range $[0.5, 1]$.[4] For a formal presentation of this procedure, see lines 2-6 in Algorithm MDE.

2.4.0.2 Generating (DQ, q) pairs and a score-accuracy mapping.

Having the synthetic workers' decision data available allows us to compute the DQ score for each synthetic worker's decision data set S_{sw_i} using the REQ method. The REQ method [84] computes DQ scores based on the weighted rate of agreement between the expert's decisions and the decisions inferred by an ensemble model M and where each (dis)agreement is weighted

[3]Thus, MDE produces semi-synthetic decision data, reflecting accuracy rates expected to arise in practice. Because we focus on expert workers, we consider settings where experts exhibit a higher accuracy rate than can be produced by a random choice. We thus simulate semi-synthetic data sets with accuracies in the range $[0.5, 1]$.

[4]In the experiments that follow, we use the default value of $intv = 0.005$. As a result, the number of synthetic workers, N, is equal to 101, so that the entire range $[0.5, 1]$ is densely covered by synthetic workers' predetermined accuracies (q_{sw_i} values). Note that the choice of intervals is not intended to replicate the distribution of real workers' accuracies, which is unknown in our setting; rather, the choice of (DQ, q) pairs with dense q values aims to provide a dense coverage of the range of possible accuracies of real workers so as to facilitate accurate induction of the mapping from DQ to accuracy (q), as described in the following subsection.

by the corresponding confidence in the ensemble's prediction.

Specifically, the ensemble's inferred decision, $M(X_i)$ for a decision instance X_i, is produced from the prediction of an ensemble of base models $\{B_j\}_{j=1}^{K}$, where each base model B_j was trained on individual (real) worker decision data S_{W_j} to produce a mapping $B_j : X \rightarrow \hat{Y}$.

More formally, the ensemble's inferred decision for each instance $X_i^k \in S_k$ is given by: $M(X_i^k) = \arg\max_z(\sum_{j=1, X_i^k \notin S_j}^{K} B_j(X_i^k)_z)$, where $B_j(X_i^k)_z$ denotes base model B_j's probability estimate that X_i^k belongs to the decision class z. $X_i^k \notin S_j$ indicates that the sum does not include estimations of a base model B_j if X_i^k is a member of the data set S_j from which B_j was induced.

Ultimately, a DQ score for a decision data set S_k is given by:

$$DQ_k = \frac{\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^+} Conf_{X_i^k}}{(\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^+} Conf_{X_i^k}) + (\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^-} Conf_{X_i^k})} \tag{2.1}$$

In Equation 2.1, the sets $s_k^+ \subset S_k$ and $s_k^- \subset S_k$ denote the set of a worker's decisions that agrees and that disagrees, respectively, with ensemble model M inferred labels. (Note that Eq. 2.1 could be used for either real workers or synthetic workers; therefore, S_k could be a real worker's decision set S_{W_k} or a synthetic worker's decision set S_{sw_k}.) $Conf_{X_i^k}$ denotes ensemble M's confidence in inferring the decision X_i^k's, given by $Conf_{X_i^k} = \sum_{j=1, X_i^k \notin S_j}^{K} B_j(X_i^k)_{M(X_i^k)}$, where $B_j(X_i^k)_{M(X_i^k)}$ denotes B_j's probability estimate that X_i^k maps to the class inferred by the ensemble M, and where

$X_i^k \notin S_j$ indicates that the sum does not include estimations of a base model B_j if X_i^k is a member of the data set S_j from which B_j was induced. Thus, the confidence $Conf_{X_i^k}$ reflects a weighted count of votes of the base models toward model M_j's class prediction (X_i^k), where each vote is weighted by the corresponding base model's probability estimation.

We note that, different from the problem settings in [84], scarce ground truth decisions are available in our problem settings, and they could be advantageous in improving base models' induction because they allow for replacing noisy labels with correct labels during the base models' training. Therefore, we slightly modify the REQ procedure that was described above. Specifically, before we induce the base models, we copy each worker's decision data S_{W_k} into $S_{\mathrm{W}_k}^{copy}$. We then replace each noisy decision $\hat{Y}_i^k$ in $S_{\mathrm{W}_k}^{copy}$ with the corresponding ground truth decision Y_i^k when it is available. We then use $S_{\mathrm{W}_k}^{copy}$ (rather than S_{W_k}) as training data when inducing each base model B_k.

In the experiments that follow, the base models were produced, by default, using a Random Forest algorithm with 100 trees. Note, however, that base models can be induced using any classification algorithm that produces class probability estimates and that is most advantageous for the specific domain and available features. For example, if the inputs provided for each decision instance are unstructured images, rather than tabular data, it is possible to use Convolutional Neural Network (CNN) or Vision Transformers to train the base models. Lines 7–10 in Algorithm MDE MDE detail how the REQ base models are trained. Procedure "Produce DQ Score" in lines 17–24 in

Algorithm MDE MDE detail how the DQ scores are calculated.

Together, the synthetic workers' predetermined accuracies and the DQ scores result in a data set of DQ-accuracy pairs, $\{DQ_{\text{SW}_i}, q_i\}_{i=1}^{N}$, from which a mapping $f : DQ \to q$ can be learned (Lines 11–13 in Algorithm MDE MDE). In principle, any regression algorithm can be applied to learn this mapping and can be selected based on cross-validation performance. In our implementation, we used a simple linear regression, informed by the analysis in [84], from which it ensues that a linear relationship exists between a worker's DQ score and the worker's true decision accuracy rate.

2.4.0.3 Inferring real workers' decision accuracies.

The DQ score for each (real) expert worker W_k's historical data, S_{W_k}, is computed. Subsequently, the mapping f is applied to produce MDE's assessment of each (real) worker's decision accuracy. (See lines 14-16 in Appendix A Algorithm MDE and the procedure, "Produce Assessment," in lines 25–29 in Algorithm MDE.)

To reduce the variance of our estimation, the steps outlined in Sections 4.1.1.–4.1.3. can be repeated using different random seeds in Section 4.1.1. Specifically, in each repetition, the decision data of a given synthetic expert is simulated by inverting a different set of instances drawn uniformly at random. As a result of these repetitions, we produce C different sets of DQ-accuracy pairs, from which C different mappings, $\{f_c\}_1^C$, are learned. The final assessment of an expert's decision accuracy is then given by the average assessment produced by the C mappings. Predicted quality values $\hat{q}$ are then truncated to the range $[0.5, 1]$.

In sum, MDE includes four main elements that are advantageous towards assessing experts' decision accuracies, given scarce ground truth: (1) the use of both ground truth and non-ground truth instances; (2) using an ensemble of base models to create ranking-based DQ scores; (3) the novel use of ground truth data to create synthetic workers, which enables the generation of (DQ,q) pairs; and (4) learning a mapping function from DQ scores to accuracies (q) that is used to assess the accuracies of expert workers. In the following section, we present our full method, MDE-HYB, which extends MDE and is designed to produce accurate assessments given any context and availability of ground truth instances (either scarce or abundant).

2.5 Results

Recall that, MDE is designed to be complementary to EAR particularly when ground truth is scarce. MDE-HYB leverages both methods to yield robust performance across settings. Tables 2.2 and 2.3 shows a comparison between MDE and EAR. As expected, MDE is significantly superior to EAR when ground truth is scarce. Similarly, when the number of ground truth instances increases, EAR yields better performance for the Audit and AMT datasets, both of which are characterized by lower predictability. When ground truth is abundant and predictability is low, EAR is more advantageous, given it does not rely on learning from noisy data. These results demonstrate that MDE and EAR are indeed often complementary, and it is therefore beneficial to leverage both methods across contexts, as done by MDE-HYB (chapter 3).

Table 2.2: Comparison between MDE and EAR with AMT Real Workers

GT PER WORKER	MDE	EAR	MDE-HYB IMPROV
5	**0.060**	0.096	37.2%**
10	**0.06**	0.068	11.7%**
15	**0.059**	0.056	-7%††
20	**0.06**	0.046	-28.5%††
25	**0.059**	0.042	-43.1%††
30	**0.059**	0.038	-56.2%††
50	**0.06**	0.027	-119%††
100	**0.059**	0.015	-292%††

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE IMPROV shows the improvement of MDE over EAR. ** MDE is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the EAR is significantly better than MDE ($p < 0.05$). †: ($p < 0.1$).

Table 2.3: Comparison between MDE and EAR with Low and High Quality Workers

		Low Quality			High Quality		
DATASET	GT PER WORKER	MDE	EAR	MDE IMPROV	MDE	EAR	MDE IMPROV
Audit	5	**0.041**	**0.142**	71.0%**	**0.062**	**0.119**	47.7%**
	10	**0.041**	**0.106**	61.5%**	**0.051**	**0.090**	43.4%**
	15	**0.043**	**0.090**	52.6%**	**0.047**	**0.073**	35.5%**
	20	**0.037**	**0.078**	52.1%**	**0.046**	**0.059**	23.2%**
	25	**0.039**	**0.068**	43.4%**	**0.042**	**0.054**	21.5%**
	30	**0.036**	**0.062**	41.7%**	**0.040**	**0.050**	21.4%**
	50	**0.035**	**0.047**	25.3%**	0.039	**0.037**	-5.23%
	100	0.035	**0.034**	-2.7%	0.035	**0.026**	-31.6%††
	300	0.034	**0.019**	-75.3%††	0.032	**0.014**	-125%††
Movie	5	**0.023**	**0.132**	82.5%**	**0.029**	**0.120**	75.6%**
	10	**0.017**	**0.103**	83.2%**	**0.022**	**0.083**	73.3%**
	15	**0.019**	**0.090**	79.3%**	**0.019**	**0.071**	73%**
	20	**0.016**	**0.076**	79.3%**	**0.017**	**0.062**	71.9%**
	25	**0.014**	**0.071**	79.7%**	**0.016**	**0.055**	70.6%**
	30	**0.015**	**0.065**	76.9%**	**0.016**	**0.049**	68.1%**
	50	**0.013**	**0.050**	73.3%**	**0.014**	**0.039**	62.8%**
	100	**0.013**	**0.035**	63.0%**	**0.013**	**0.026**	50.8%**
	300	**0.012**	**0.018**	34.8%**	**0.012**	**0.014**	11.1%**
Spam	5	**0.016**	**0.133**	87.7%**	**0.017**	**0.121**	85.6%**
	10	**0.015**	**0.103**	85.4%**	**0.016**	**0.083**	81.2%**
	15	**0.015**	**0.086**	82.2%**	**0.015**	**0.069**	77.8%**
	20	**0.015**	**0.078**	80.8%**	**0.015**	**0.059**	75.2%**
	25	**0.015**	**0.065**	76.6%**	**0.015**	**0.054**	72.4%**
	30	**0.015**	**0.062**	75.6%**	**0.015**	**0.046**	68.2%**
	50	**0.014**	**0.044**	67.8%**	**0.014**	**0.036**	60.5%**
	100	**0.014**	**0.027**27	46.0%**	**0.013**	**0.020**	33.5%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE IMPROV shows the improvement of MDE over EAR. ** MDE is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the EAR is significantly better than MDE ($p < 0.05$). †: ($p < 0.1$).

Chapter 3

Assessing Experts' Decision Accuracy Irrespective of the Number of Ground Truth

3.1 Machine-learning-based Decision Quality Estimation-Hybrid (mde-hyb)

This advanced approach, the **M**achine-learning-based **D**ecision quality **E**stimation-**Hyb**rid (MDE-HYB), ensures reliable estimates of accuracy irrespective of the number of ground truth. The MDE-HYB first produces and uses two complementary estimates of experts' decision accuracies, each relying on different information sets and processes that can be advantageous under different circumstances. The first estimate, the MDE in Chapter 2, exploits the large amount of data available on the experts' decisions, along with (scarce) ground truth information. Yet, as more ground truth becomes available, an estimation that relies exclusively on ground truth decisions can yield an optimal estimation. Hence, this approach incorporates a second estimate that is simply the frequency of correct decisions computed exclusively from ground truth data, which we henceforth refer to as the **E**stimated **A**ccuracy **R**ate (EAR). Our method, MDE-HYB, evaluates the error rates of the two estimates (MDE and EAR), and if one of the estimates is deemed superior, it selects that estimate. Otherwise, MDE-HYB infers experts' decision accuracies as a linear combination of the two estimates. In the following sections, we outline each of the elements of our approach.

3.1.1 MDE-HYB: Balancing Estimations from Noisy Labels and from Ground Truth

The MDE method in Chapter 2, is designed to leverage inferences from big, noisy decision data, in addition to scarce ground truth data. It aims to be advantageous particularly when the scarce ground truth per expert ($\frac{|GT|}{|W|}$) cannot yield reliable assessments when it is used exclusively. However, as more ground truth data are available for each expert, an exclusive reliance on ground truth can yield optimal assessments. In particular, an alternative estimation, EAR, corresponds to estimating the decision accuracy of expert W_k, based on the rate of accurate decisions among the set GT_k of decisions with ground truth; EAR is given by:

$$\hat{q}_k^{\text{EAR}} = (\sum_{i=1}^{|GT_k|} I[Y_i == \hat{Y}_i])/|GT_k| \tag{3.1}$$

In general, in different domains and given different numbers of ground truth instances per expert, either approach – EAR or MDE – may yield a more reliable assessment. Thus, leveraging either approach may be more appropriate in different contexts so as to produce a more reliable estimation than is possible by relying exclusively on one approach, across contexts. We build on this notion and propose a method that would always produce results that are at least as good as, or superior to, the alternative, regardless of the quantity of ground truth and the context. To this end, the proposed method, MDE-HYB, evaluates the accuracy of the assessments produced by MDE and by EAR in any given context; MDE-HYB then either selects the assessment approach that is superior or infers experts' accuracies using a linear combination of both estimates. The key challenge we address is that determining the accuracy of the assessments produced by MDE and EAR in a given context is non-trivial, given that workers' accuracies are unknown. Algorithm block MDE-HYB MDE-HYB outlines the pseudo code for producing MDE-HYB's estimations.

Specifically, MDE-HYB first applies MDE and EAR separately to produce assessments for each worker (lines 2–3 in Algorithm MDE-HYB). Then, to determine the accuracy of each approach's assessments, MDE-HYB generates an estimation of the distribution of errors produced by each method (i.e., MDE and EAR). If either MDE or EAR is estimated to have a statistically significant and meaningfully lower assessment error, then experts' accuracies are inferred based on this superior assessment approach. Otherwise, when there is no evidence that this approach is superior to the other in a given context, MDE-HYB assesses an expert's accuracy as a linear combination of the assessment produced by MDE and EAR.

Because properties of the distribution of MDE's errors cannot be computed in closed form, MDE-HYB estimates MDE's error distributions by a form of bootstrapping. Specifically, we draw R different samples from GT, and from each sample, MDE-HYB internally simulates decision data for P additional synthetic workers, with predetermined (known) accuracies. This step results in decision data for $R * P$ additional synthetic workers. MDE and EAR are then applied to estimate the accuracies of these synthetic workers. Comparing the assessments of MDE and EAR to the predetermined decision accuracies of these synthetic workers allows us to produce a distribution of assessment errors for each approach.

Specifically, to create variations across R different samples of synthetic workers, each sample size t is produced by drawing instances at random from the set of all available ground truth instances, GT, such that $t \leq |GT|$.[1] From each of the samples, we then simulate decision data for P different synthetic workers, each with a different decision accuracy $q_{r,p}$. Each synthetic worker's accuracy $q_{r,p}$ is drawn uniformly from the estimated range of the real workers' accuracies $[\hat{q}_{lower}, \hat{q}_{upper}]$. Specifically, the range $[\hat{q}_{lower}, \hat{q}_{upper}]$ reflects the 99% confidence interval of the real workers' estimated accuracies, estimated as the average assessment produced by MDE and EAR and given by $\{(\hat{q}_k^{\text{EAR}}+\hat{q}_k^{\text{MDE}})/2\}_1^K$ (lines 14–15 in Algorithm MDE-HYB). The synthetic workers' decision data is then simply generated by flipping the (correct) decisions of a $(1 - q_{r,p})$ proportion of the decision instances in the corresponding sample. Finally, for each of the R samples, we apply MDE and EAR separately to produce decision accuracy assessments for P synthetic workers. Because the true accuracy of each synthetic worker is known, the estimation errors for MDE and EAR can be directly computed. The entire procedure for generating the error distributions

[1]In the experiments reported here, t is either 20% of the workers' average decision data set size or t=$|GT|$ if the former is larger than $|GT|$. Other sample sizes can be used so as to create diverse samples.

for both MDE and EAR is detailed in line 4 and lines 13–30 in Algorithm MDE-HYB.

We now have the error distribution for both EAR and MDE so as to assess whether one approach yields a superior error to the other. Specifically, we examine whether the difference in errors is greater than d, where d reflects a meaningful difference, given the relevant context, in practice.[2] We do so via two 2-sample one-tailed t-tests, comparing the error means of the two approaches. Specifically, in the first test, the null hypothesis is given by H_0^1 : $(\mu_{\text{MDE}} - \mu_{\text{EAR}}) \leq d$, with an alternative hypothesis of $H_a^1 : (\mu_{\text{MDE}} - \mu_{\text{EAR}}) > d$; in the second test, the null hypothesis is $H_0^2 : (\mu_{\text{EAR}} - \mu_{\text{MDE}}) \leq d$, with an alternative hypothesis of $H_a^2 : (\mu_{\text{EAR}} - \mu_{\text{MDE}}) > d$.[3] (See lines 5–6 in Algorithm MDE-HYB.)

Finally, if one of the two (but not both) null hypotheses is not supported, MDE-HYB uses the superior approach to infer experts' decision accuracies. Alternatively, if both null hypotheses cannot be rejected, and if the two error means are comparable in the relevant context, MDE-HYB assesses an expert's accuracy as a linear combination of both MDE's and EAR's assessments. In particular, based on the hypotheses tests' p-values, the method which we are more confident that will have a smaller error rate, will have a higher weight

[2]In the empirical results we report, $d = 0.01$; that is, a 1% difference in diagnosis accuracy has significant consequences in rare disease diagnosis.

[3]Note that we use the two tests because in either case, the alternative hypothesis states that the mean error either of MDE or of EAR is meaningfully larger than that of the other, but not the other way around. In addition, in the empirical results reported below, $\alpha = 0.02$.

in the linear combination. (See lines 7–12 in Algorithm MDE-HYB.)

3.2 Results

In this section, we report the results of empirical evaluations, comparing our approach's performances relative to each alternative under different settings; we also report the results of ablation studies that evaluate the relative contributions of key elements of our proposed method. Consequently, we assess and report MDE-HYB's performance: (1) for different data domains, (2) when ground truth is available for different numbers of decision instances, and (3) when workers exhibit different levels of expertise.

Focusing first on low-quality workers, the top row of Figure 3.1 shows curves of the average MAE achieved by MDE-HYB, along with that achieved by the benchmarks (EAR, GM-GT, and GM-ALL), for settings with scarce ground truth instances, and when workers' decision accuracies range between 61% and 80% (henceforth referred to as low-quality workers). Table 3.1 shows the results of these experiments, along with the improvement achieved by MDE-HYB relative to each of the benchmarks and its statistical significance.[4] Although our focus is on settings that have scarce ground truth instances, note that the tables present results for both scarce and abundant ground truth

[4]Note that because of space constraints, results reported in tables throughout this paper are shown with only two decimal points; as a result, in a few cases, the reported difference between two methods is slightly different than if the two respective (truncated) numbers in the table are subtracted. In addition, given the smaller size of the Spam dataset, in Table 2 and subsequent tables, we cannot produce results for this data for settings in which each of 20 workers has 300 instances with ground truth labels.

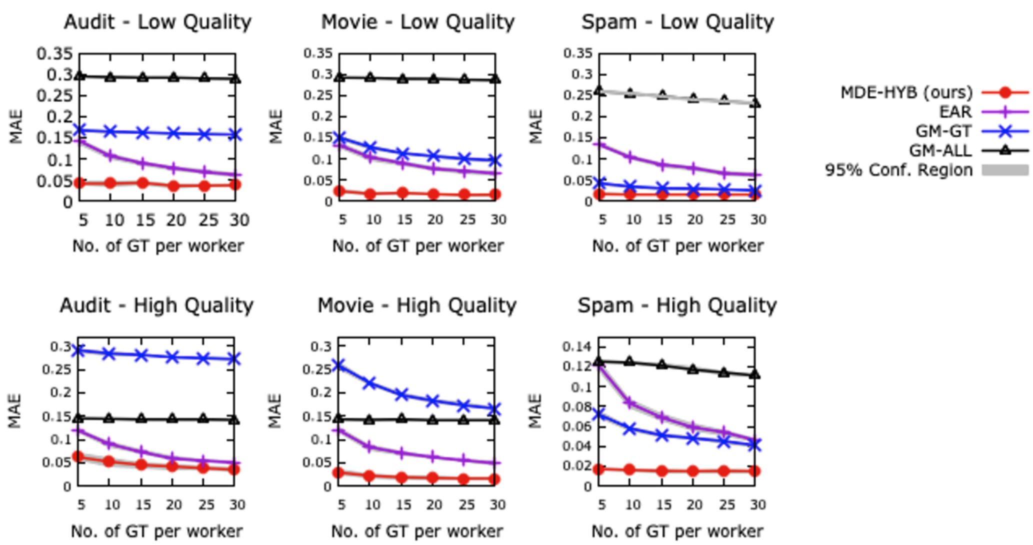
instances to determine whether MDE-HYB may be inferior and thus undesirable when ground truth is abundant.

Then, in the bottom row of Figure 3.1 (high-quality workers) and in Table 3.2, we focus on higher quality workers, whose decision accuracies range between 76% and 95%. As shown, when ground truth data are scarce, MDE-HYB achieves significantly superior estimations of workers' decision accuracies as compared to each of the alternatives, across domains and levels of workers' expertise. For example, assuming five ground truth instances per worker, MDE-HYB achieves between 60.8% and 93.7% higher accuracy, relative to the alternatives, across the three domains. For the Audit dataset, where all methods produced the highest estimation errors, the best alternative, EAR, exhibits an average 14.2% error; the worst alternative, GM-ALL, yields a 29.5% error; and MDE-HYB exhibits an average error of only 4.1%.

Note that, given MDE-HYB's use of inference, its performance relates also to the predictability of a given domain. It exhibits an error between 4.1% and 1.9% for the Audit domain, which has low predictability (AUC of 0.671), and an error between 1.6% and 1.4% for the spam domain, for which the AUC is 0.987. Finally, We also observe similar findings that demonstrate the superiority of MDE-HYB, regardless of the level of predictability, for high quality workers, as shown in the bottom row of Figure 3.1 and in Table 3.2.

Interestingly, as with MDE-HYB, all benchmarks take advantage of ground truth data; and GM-GT and GM-ALL explicitly make use of the *GT* set to induce a global model. Yet, MDE-HYB more effectively exploits inference, ground

Figure 3.1: MDE-HYB's Performance Relative to Benchmarks



MAE measure for experts' accuracy estimation errors (mean measured across 50 repetitions) for our MDE-HYB approach and the baseline approaches. Results are reported given a varying number of ground truth instances, different datasets, and workers' quality levels. The grey shaded region shows the 95% confidence bound for each method. Note that in many cases the confidence bounds are very narrow and are therefore not visually observable.

truth data, and experts' noisy decisions, resulting in a consistently advantageous performance across settings that is unmatched by any of the alternatives. Furthermore, it is important to note that the GM-ALL baseline uses a set of data that is identical to the set used by MDE-HYB. The fact that GM-ALL was the weakest baseline in the majority of cases and always produced inferior results to MDE-HYB highlights an important point: The benefit of our MDE-HYB approach is not simply from bringing to bear both noisy labels and ground truth; rather, it results from the non-trivial and meaningful manner by which it brings noisy labels and ground truth to bear, allowing our approach to leverage imperfect information from noisy labels to yield an accurate estimation of experts' accuracies.

Table 3.1: MDE-HYB and Benchmarks Performance Measured by MAE for Low-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.041**	0.142	71.0%**	0.167	75.4%**	0.295	86.1%**
	10	**0.041**	0.106	61.5%**	0.164	75.2%**	0.293	86.1%**
	15	**0.043**	0.090	52.6%**	0.162	73.7%**	0.292	85.4%**
	20	**0.036**	0.078	54.1%**	0.160	77.5%**	0.292	87.7%**
	25	**0.036**	0.068	47.3%**	0.158	77.2%**	0.290	87.6%**
	30	**0.037**	0.062	41.2%**	0.157	76.7%**	0.289	87.4%**
	50	**0.043**	0.047	9.8%**	0.152	71.9%**	0.285	85.0%**
	100	**0.034**	0.034	0.0%	0.145	76.3%**	0.275	87.5%**
	300	**0.019**	0.019	0.0%	0.122	84.2%**	0.236	91.8%**
Movie	5	**0.023**	0.132	82.5%**	0.150	84.6%**	0.292	92.1%**
	10	**0.017**	0.103	83.2%**	0.127	86.4%**	0.291	94.1%**
	15	**0.019**	0.090	79.3%**	0.113	83.6%**	0.289	93.6%**
	20	**0.016**	0.076	79.3%**	0.106	85.2%**	0.289	94.6%**
	25	**0.014**	0.071	79.7%**	0.100	85.6%**	0.287	95.0%**
	30	**0.015**	0.065	76.9%**	0.096	84.4%**	0.286	94.8%**
	50	**0.013**	0.050	73.3%**	0.084	84.1%**	0.282	95.2%**
	100	**0.015**	0.035	56.7%**	0.074	79.3%**	0.270	94.4%**
	300	**0.011**	0.018	40.5%**	0.055	80.2%**	0.223	95.1%**
Spam	5	**0.016**	0.133	87.7%**	0.042	60.8%**	0.259	93.7%**
	10	**0.015**	0.103	85.4%**	0.034	56.2%**	0.254	94.1%**
	15	**0.015**	0.086	82.2%**	0.030	49.7%**	0.248	93.8%**
	20	**0.015**	0.078	80.8%**	0.029	48.2%**	0.241	93.8%**
	25	**0.015**	0.065	76.6%**	0.027	42.8%**	0.236	93.5%**
	30	**0.015**	0.062	75.6%**	0.025	39.8%**	0.231	93.4%**
	50	**0.014**	0.044	67.8%**	0.021	31.9%**	0.208	93.2%**
	100	**0.014**	0.027	45.7%**	0.015	0.3%	0.149	90.3%**

Experts' accuracy estimation errors. Values shown are mean absolute error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Our results also underscore a key aspect of the performance of our approach: MDE-HYB's performance is robust across settings that involve different levels of availability of ground truth. Note that, given a sufficiently large number of instances of ground truth information, EAR is guaranteed to converge to the correct decision accuracy of a given worker. However, Figure 3.1 and Tables 3.1 and 3.2 show that all the methods' estimations improve with more ground truth; yet MDE-HYB consistently either exhibits significantly superior accuracies or is otherwise comparable to the best alternative. Importantly, recall that MDE-HYB aims to estimate workers' decision performances under scarce ground truth. However, it can be safely deployed to yield state-of-the-art performance, regardless of the number of available ground truth instances. In fact, when ground truth is abundant, both MDE-HYB and EAR, in particular, achieve comparable and highly accurate estimations.

Overall, MDE-HYB exhibits robust performance, consistently producing either the best, or at least comparable, estimations of experts' decision accuracies relative to the alternatives; these results hold across domains, across the number of ground truth instances, and across the workers' level of expertise.

Table 3.2: MDE-HYB and Benchmarks Performance Measured by MAE for High-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.062**	0.119	47.7%**	0.291	78.5%**	0.145	56.8%**
	10	**0.052**	0.090	43.0%**	0.284	81.9%**	0.144	64.2%**
	15	**0.045**	0.073	37.9%**	0.281	83.9%**	0.143	68.4%**
	20	**0.042**	0.059	29.1%**	0.276	84.8%**	0.143	70.6%**
	25	**0.038**	0.054	30.0%**	0.274	86.2%**	0.143	73.6%**
	30	**0.035**	0.050	30.8%**	0.272	87.2%**	0.142	75.4%**
	50	**0.036**	0.037	2.6%	0.265	86.3%**	0.14	74.0%**
	100	**0.026**	0.026	0.0%	0.252	89.5%**	0.135	80.5%**
	300	**0.014**	0.014	0.0%	0.211	93.2%**	0.116	87.6%**
Movie	5	**0.029**	0.120	75.6%**	0.259	88.6%**	0.143	79.5%**
	10	**0.022**	0.083	73.3%**	0.221	89.9%**	0.142	84.4%**
	15	**0.019**	0.071	73.0%**	0.195	90.1%**	0.143	86.5%**
	20	**0.017**	0.062	71.9%**	0.183	90.5%**	0.141	87.6%**
	25	**0.016**	0.055	70.6%**	0.173	90.6%**	0.141	88.5%**
	30	**0.016**	0.049	68.1%**	0.166	90.5%**	0.141	88.8%**
	50	**0.014**	0.039	63.7%**	0.147	90.4%**	0.139	89.8%**
	100	**0.015**	0.026	44.1%**	0.128	88.4%**	0.133	88.9%**
	300	**0.009**	0.014	30.6%**	0.095	90.2%**	0.109	91.4%**
Spam	5	**0.017**	0.121	85.6%**	0.072	75.8%**	0.125	86.1%**
	10	**0.016**	0.083	81.2%**	0.058	73.2%**	0.124	87.4%**
	15	**0.015**	0.069	77.8%**	0.051	70.0%**	0.121	87.3%**
	20	**0.015**	0.059	75.2%**	0.048	69.3%**	0.117	87.5%**
	25	**0.015**	0.054	72.4%**	0.045	66.7%**	0.114	87.0%**
	30	**0.015**	0.046	68.2%**	0.041	64.7%**	0.111	86.9%**
	50	**0.015**	0.036	59.3%**	0.034	57.6%**	0.101	85.5%**
	100	**0.012**	0.020	38.3%**	0.022	43.8%**	0.072	82.8%**

Experts' accuracy estimation errors. Values show mean absolute error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better or otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

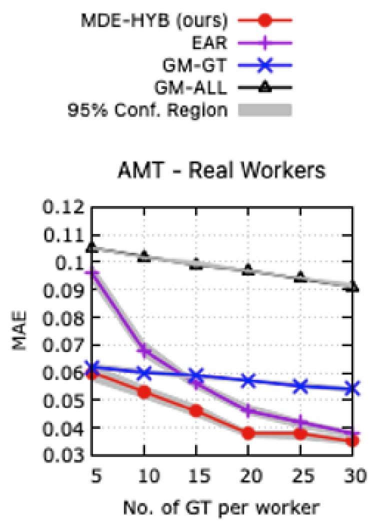
3.2.1 Evaluation on Purposely Compiled Human Workers' Decision Dataset

We applied MDE-HYB to evaluate the decision accuracy of human workers, recruited via Amazon Mechanical Turk (AMT), to determine the sentiments expressed in product reviews. Figure 3.2 and Table 3.3 show performance comparisons of MDE-HYB and the benchmarks. Recall that, because of the cost of acquiring workers' decisions, these data likely include a smaller number of decision instances for each worker than is available from workers' histories in many settings in practice. As a result, this factor may undermine the effectiveness of machine learning models induced from the data. Nevertheless, as we show below, these results establish the robustness of our approach and corroborate the conclusions drawn from the results reported previously. In particular, the results establish that MDE-HYB yields state-of-the-art performance, yielding consistently and statistically significant better estimations than the alternatives, or otherwise estimations that are comparable to any of the existing alternatives.

3.2.2 Ablation Studies

The empirical evaluations demonstrate that MDE-HYB takes advantage of data-driven inference from different experts, as well as the calibration of experts' accuracies, and it relies both on experts' noisy decisions and on ground truth in a manner that is unmatched by the use of this information by benchmark methods. In this section, we report on our ablation studies as we aim

Figure 3.2: Performance with AMT Real Workers



MAE measure for experts' accuracy estimation errors. The grey shaded region shows the 95% confidence bound for each method.

Table 3.3: MDE-HYB and Benchmarks Performance Measured by MAE with Real Workers

GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
5	**0.06**	0.096	37.2%**	0.062	3.6%*	0.105	42.7%**
10	**0.053**	0.068	22.4%**	0.06	12.0%**	0.102	47.9%**
15	**0.046**	0.056	17.9%**	0.059	22.4%**	0.099	54.1%**
20	**0.038**	0.046	18.1%**	0.057	33.3%**	0.097	60.7%**
25	**0.038**	0.042	9.1%**	0.055	31.8%**	0.094	59.9%**
30	**0.035**	0.038	7.1%**	0.054	34.5%**	0.091	61.5%**
50	**0.027**	**0.027**	-0.2%	0.047	41.4%**	0.081	66.2%**
100	**0.015**	**0.015**	0.0%	0.031	50.6%**	0.054	71.9%**

Experts' accuracy estimation errors. Values show mean absolute error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better or otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

to establish the relative benefits of key elements of our approach. Specifically, we intend to establish the benefits from learning and aggregating the inference from each individual expert's decision data and, separately, the benefits of learning both from each expert's noisy decisions and from ground truth data. In addition, we study whether MDE-HYB indeed benefits from its ability to adaptively bring to bear both MDE and EAR methods. We first explore whether MDE-HYB's inference of $M(X_i^j)$ is beneficial. In particular, we explore the benefit that stems from MDE-HYB's inference's being based on an ensemble of base models, each induced from a *different expert's data*; this approach allows us to account for the idiosyncratic uncertainties of different base models while aggregating their inferences.

As an alternative, we consider a variant of our approach, which follows Algorithms MDE and MDE-HYB, except that in this variant, $M(X_i^k)$ is inferred from a single model, induced from $\{S_{W_k}\}_1^K$. That is, it is inferred from all experts' noisy decision instances. We refer to this variant as MDE-HYB-SingleModel-All, or MDE-HYB-SM-ALL. A second variant aims to explore whether greater benefit can be achieved by inferring $M(X_i^k)$ using a model induced exclusively from instances with ground truth, thereby avoiding learning from the experts' noisy decisions altogether. Given the scarcity of ground truth in our setting, inducing base models from only a handful of ground truth instances (e.g., five) would not be feasible; hence, for this variant, we also replace the ensemble with a single model, induced only from $GT = \bigcup_{k=1}^{K} GT_k$. We refer to this variant as MDE-HYB-SingleModel-GroundTruth, or MDE-HYB-SM-GT. Finally, for both variants, given that $Conf_{X_i^k}$ can no longer be the summation of the base models' probabilities, $Conf_{X_i^k}$ is simply the single model's estimated probability for the predicted class.

Table 3.4: MDE-HYB and Variants Performance measured by MAE for Low-quality Workers

DATASET	GT PER WORKER	MDE-HYB	MDE-HYB-SM-GT	MDE-HYB IMPROV	MDE-HYB-SM-ALL	MDE-HYB IMPROV
Audit	5	**0.041**	0.163	74.9%**	0.294	86.0%**
	10	**0.041**	0.159	74.3%**	0.293	86.1%**
	15	**0.043**	0.156	72.7%**	0.292	85.4%**
	20	**0.036**	0.155	76.8%**	0.291	87.6%**
	25	**0.036**	0.153	76.4%**	0.291	87.6%**
	30	**0.037**	0.151	75.9%**	0.289	87.3%**
	50	**0.043**	0.147	70.9%**	0.285	85.0%**
	100	**0.034**	0.137	75.0%**	0.274	87.5%**
	300	**0.019**	0.11	82.6%**	0.235	91.8%**
Movie	5	**0.023**	0.145	84.1%**	0.293	92.1%**
	10	**0.017**	0.123	86.0%**	0.291	94.1%**
	15	**0.019**	0.109	82.9%**	0.291	93.6%**
	20	**0.016**	0.099	84.3%**	0.289	94.6%**
	25	**0.014**	0.093	84.4%**	0.288	95.0%**
	30	**0.015**	0.089	83.1%**	0.287	94.8%**
	50	**0.013**	0.078	82.8%**	0.283	95.3%**
	100	**0.015**	0.065	76.6%**	0.27	94.4%**
	300	**0.011**	0.045	76.0%**	0.223	95.1%**
Spam	5	**0.016**	0.034	51.2%**	0.273	94.0%**
	10	**0.015**	0.029	47.3%**	0.267	94.4%**
	15	**0.015**	0.025	38.3%**	0.262	94.1%**
	20	**0.015**	0.023	34.1%**	0.251	94.0%**
	25	**0.015**	0.022	29.6%**	0.245	93.8%**
	30	**0.015**	0.021	26.4%**	0.238	93.6%**
	50	**0.014**	0.017	15.7%**	0.214	93.4%**
	100	0.014	**0.011**	-25.8%††	0.149	90.3%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over a variant. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the other method is significantly better than MDE-HYB ($p < 0.05$). †: ($p < 0.1$).

For settings with scarce ground truth, Figure 3.3 shows the MDE-HYB's performance relative to that of each of the variants described, for experts

Table 3.5: MDE-HYB and Variants Performance measured by MAE for High-quality Workers

DATASET	GT PER WORKER	MDE-HYB	MDE-HYB-SM-GT	MDE-HYB IMPROV	MDE-HYB-SM-ALL	MDE-HYB IMPROV
Audit	5	**0.062**	0.283	77.9%**	0.144	56.5%**
	10	**0.052**	0.276	81.3%**	0.143	64.1%**
	15	**0.045**	0.272	83.4%**	0.144	68.6%**
	20	**0.042**	0.268	84.3%**	0.143	70.6%**
	25	**0.038**	0.265	85.7%**	0.143	73.5%**
	30	**0.035**	0.262	86.7%**	0.142	75.4%**
	50	**0.036**	0.254	85.7%**	0.14	74.1%**
	100	**0.026**	0.238	88.9%**	0.135	80.5%**
	300	**0.014**	0.192	92.5%**	0.111	87.0%**
Movie	5	**0.029**	0.252	88.3%**	0.143	79.5%**
	10	**0.022**	0.211	89.4%**	0.144	84.5%**
	15	**0.019**	0.186	89.6%**	0.143	86.5%**
	20	**0.017**	0.172	89.8%**	0.142	87.7%**
	25	**0.016**	0.162	90.0%**	0.142	88.5%**
	30	**0.016**	0.154	89.8%**	0.142	88.9%**
	50	**0.014**	0.134	89.5%**	0.139	89.8%**
	100	**0.015**	0.113	87.0%**	0.133	88.9%**
	300	**0.009**	0.079	88.1%**	0.103	90.9%**
Spam	5	**0.017**	0.058	69.6%**	0.131	86.7%**
	10	**0.016**	0.046	66.0%**	0.128	87.8%**
	15	**0.015**	0.041	62.7%**	0.124	87.7%**
	20	**0.015**	0.038	61.0%**	0.121	87.9%**
	25	**0.015**	0.035	57.9%**	0.117	87.3%**
	30	**0.015**	0.033	55.7%**	0.116	87.4%**
	50	**0.015**	0.026	43.8%**	0.102	85.7%**
	100	**0.012**	0.016	24.2%**	0.074	83.1%**

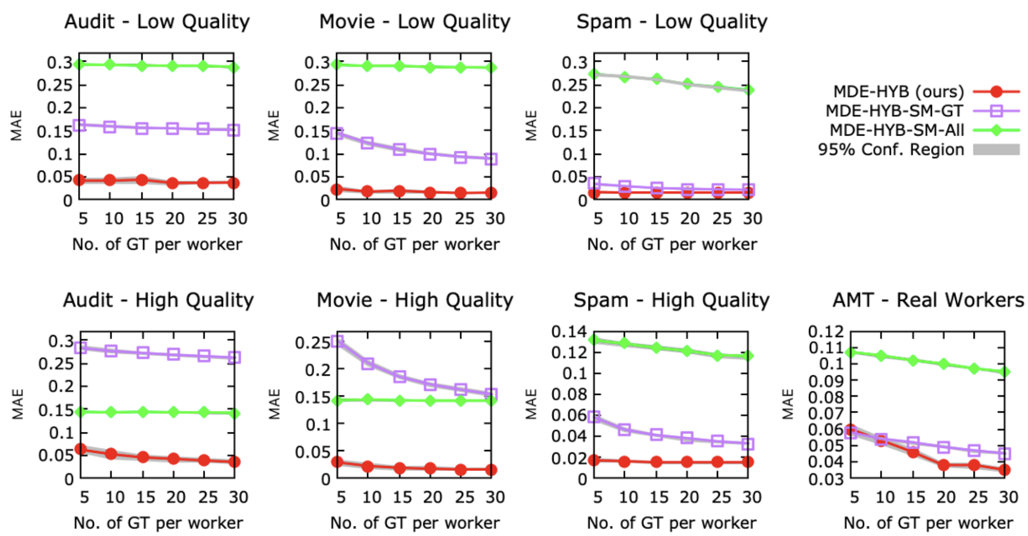
Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over a variant. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

of either low quality (top row) or higher quality (bottom row), or for the AMT workers (bottom row, far right). Tables 3.4 to 3.6 provides detailed numerical results and statistical significance tests for these settings, as well as for a setting with abundant ground truth data. The results show that inferring $M(X_i^k)$ from a single model that is induced either exclusively from GT (MDE-HYB-SM-GT) or from both GT and S (MDE-HYB-SM-ALL) almost always yields a substantially and statistically significant worse performance than that of MDE-HYB.

Of particular interest is MDE-HYB's use of noisy labels. Of the two variants of our approach, MDE-HYB-SM-ALL—which uses all of the experts' (noisy) decisions, as well as ground truth labels, to induce an ensemble model—often yields the worst performance, even compared to when the model is induced exclusively from ground truth. In contrast, MDE-HYB yields a superior performance relative to both variants across settings. These results establish that MDE-HYB's inference by modeling the uncertainties of different expert models is advantageous, relative to learning an ensemble from the entire data; in addition, the results establish MDE-HYB's use of noisy data is instrumental towards its better performance.

The results reveal how MDE-HYB's learning of individual experts' decision patterns, which accounts for uncertainties in each expert's base model inferences (reflected in $conf_{X_i^k}$), enables MDE-HYB to more effectively leverage experts' noisy decisions than is possible with a model that simply is induced from experts' noisy decisions. In addition, the variant of our approach that infers $M(X_i^k)$ using a single model, induced exclusively from the set of ground truth instances GT, similarly does not match MDE-HYB's performance; this underscores that MDE-HYB's particular use of experts' noisy decisions renders these decisions instrumental in producing a better performance than is possible when learning comes exclusively from instances with ground truth labels.

Figure 3.3: Evaluating Variants of MDE-HYB



MAE measure for experts' accuracy estimation errors (mean measured across 50 repetitions) for our MDE-HYB approach and for variants of it. Results are reported with varying numbers of ground truth instances, different datasets, and different workers quality levels. The grey shaded region shows the 95% confidence bound for each method. Note that in many cases the confidence bounds are very narrow and are therefore not visually observable.

Next, recall that MDE-HYB is designed to leverage both MDE and EAR to yield a robust performance across settings. In particular, the MDE component of MDE-HYB is designed to be complementary to EAR, particularly when ground truth is scarce, which is the context on which we focus in this work.

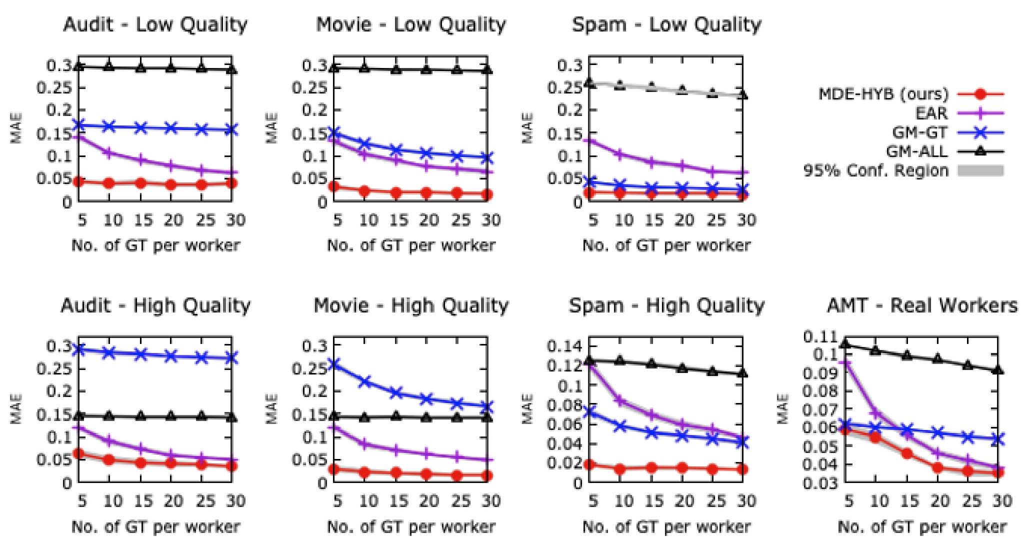
Given this complementary design, MDE-HYB aims to adaptively rely on MDE or EAR to yield a reliable state-of-the-art performance across settings. Our main results show that MDE-HYB often is superior to EAR, particularly when ground truth is scarce; hence, we establish that MDE-HYB relies on MDE in its superior performance. We also question, then, whether MDE-HYB benefits from EAR in achieving its performance, such as when ground truth is abundant. To answer this question, we compared MDE-HYB and MDE alone, and we report the results of these experiments in Tables 3.7 and 3.8. Our results show that, as expected, MDE-HYB in most settings yields a similar performance to that of MDE, particularly when ground truth is scarce. However, when more ground truth is available, MDE-HYB most often offers superior accuracies to those achieved by MDE. In particular, MDE-HYB offers higher estimation accuracy than that produced by MDE when ground truth is more abundant, as in the Audit and AMT datasets, which also are characterized by lower predictability. In these contexts, MDE-HYB can produce better estimations than is possible with MDE alone. It does so by effectively adapting so that it relies more on EAR, which uses ground truth exclusively; thus, the results are unaffected by the predictability of the data domain. Overall, our results demonstrate that MDE-HYB's performance relies on both the MDE and EAR components and that it does so by effectively adapting so that it relies on either or both, across contexts.

Tables 3.7 and 3.8 shows a comparison between MDE-HYB and its component, MDE, for domains with low- and high-quality workers, and for the

AMT data of real workers, respectively. When more ground truth becomes available, MDE-HYB most often obtains superior or at least equivalent results to MDE. MDE-HYB especially significant for the Audit and AMT datasets which are characterized by lower predictability. In these contexts, MDE-HYB is able to produce better estimatiosn than possible with
textscmde alone by effectively adapting to rely more on the assessments by its EAR component, which is unaffected by low predictability.

3.2.3 Additional Evaluations

Figure 3.4: MDE-HYB's Performance when LogitBoost Is Used for Inference by MDE-HYB



MAE of experts' accuracy estimation (measured across 50 repetitions) for our MDE-HYB approach (using LogitBoost) and the baseline approaches. Results are shown for varying numbers of ground truth instances, different datasets, and different workers' quality levels. The grey shaded region shows the 95% confidence bound for each method.

3.2.3.1 Experiments with different learners.

An important property of MDE-HYB is that the inference element of our approach—that is, the inference of $B_j(X_i^k)$—is model-agnostic. This property is important, given that different techniques' inductive biases are more advantageous for learning models in different domains. Consequently, our approach can be applied to assess experts' accuracies using any modeling technique that is most suitable for learning the underlying experts' domain (e.g., for determining fraud or for medical diagnoses from medical records). In general, for a given domain, the performance of alternative inductive techniques for inferring $B_j(X_i^k)$ can be evaluated over set S to select the one that yields the best performance (e.g., AUC).[5] In the main results (section 6.1), we reported MDE-HYB's performance using a random forest algorithm with 100 trees to induce each expert's base model; here, we demonstrate MDE-HYB's performance when $B_j(X_i^k)$ is inferred with additive logistic regression using LogitBoost [77]. For scarce ground truth instances, Figure 3.4 shows the MDE-HYB's performance relative to the three benchmarks (EAR, GM-GT, and GM-ALL) for low- and high-quality expert workers and for AMT real workers. In addition, Tables 3.9 and 3.10 show all our results, including results for settings that have abundant ground truth, and statistical significance tests' results. All results indicate that MDE-HYB—when using LogitBoost to infer $B_j(X_i^k)$—exhibits the same benefits previously established: MDE-HYB often yields substantially better es-

[5]Note that this evaluation is possible because the *ranking* of different models by their performance on noisy data (the data in our setting) also correctly reflects these models' relative performances on correctly labeled data [49].

timations than ones that are possible with the benchmarks, and otherwise, at least comparable ones.

Table 3.6: MDE-HYB and Variants Performance measured by MAE for AMT Real Workers

GT PER WORKER	MDE-HYB	MDE-HYB-SM-GT	MDE-HYB IMPROV	MDE-HYB-SM-ALL	MDE-HYB IMPROV
5	0.060	**0.058**	-3.8%	0.107	43.9%**
10	**0.053**	0.054	2.0%	0.105	49.2%**
15	**0.046**	0.052	11.5%**	0.102	55.3%**
20	**0.038**	0.049	22.6%**	0.100	61.9%**
25	**0.038**	0.047	19.5%**	0.097	61.2%**
30	**0.035**	0.045	21.8%**	0.095	62.8%**
50	**0.027**	0.038	27.5%**	0.084	67.6%**
100	**0.015**	0.023	35.7%**	0.032	52.9%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over a variant. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Table 3.7: Comparison between MDE-HYB and MDE with Low and High Quality Workers

DATASET	GT PER WORKER	Low Quality MDE-HYB	Low Quality MDE	Low Quality MDE-HYB IMPROV	High Quality MDE-HYB	High Quality MDE	High Quality MDE-HYB IMPROV
Audit	5	**0.041**	**0.041**	0.0%	**0.062**	**0.062**	0.0%
	10	**0.041**	**0.041**	0.0%	0.052	**0.051**	-0.6%
	15	**0.043**	**0.043**	0.0%	**0.045**	0.047	3.8%
	20	**0.036**	0.037	4.1%*	**0.042**	0.046	7.6%**
	25	**0.036**	0.039	6.8%**	**0.038**	0.042	10.8%**
	30	0.037	**0.036**	-1%	**0.035**	0.040	12%**
	50	0.043	**0.035**	-20.9%††	**0.036**	0.039	7.5%*
	100	**0.034**	0.035	2.6%	**0.026**	0.035	24%**
	300	**0.019**	0.034	43%**	**0.014**	0.032	55.5%**
Movie	5	**0.023**	**0.023**	0.0%	**0.029**	**0.029**	0.0%
	10	**0.017**	**0.017**	0.0%	**0.022**	**0.022**	0.0%
	15	**0.019**	**0.019**	0.0%	**0.019**	**0.019**	0.0%
	20	**0.016**	**0.016**	0.0%	**0.017**	**0.017**	0.0%
	25	**0.014**	**0.014**	0.0%	**0.016**	**0.016**	0.0%
	30	**0.015**	**0.015**	0.0%	**0.016**	**0.016**	0.0%
	50	**0.013**	**0.013**	0.0%	**0.014**	**0.014**	2.3%
	100	0.015	**0.013**	-17.2%††	0.015	**0.013**	-13.6%††
	300	**0.011**	0.012	8.7%**	**0.009**	0.012	21.9%**
Spam	5	**0.016**	**0.016**	0.0%	**0.017**	**0.017**	0.0%
	10	**0.015**	**0.015**	0.0%	**0.016**	**0.016**	0.0%
	15	**0.015**	**0.015**	0.0%	**0.015**	**0.015**	0.0%
	20	**0.015**	**0.015**	0.0%	**0.015**	**0.015**	0.0%
	25	**0.015**	**0.015**	0.0%	**0.015**	**0.015**	0.0%
	30	**0.015**	**0.015**	0.0%	**0.015**	**0.015**	0.0%
	50	**0.014**	**0.014**	0.0%	0.015	**0.014**	-3%
	100	**0.014**	**0.014**	-0.6%	**0.012**	0.013	7.2%*

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the variant MDE. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the MDE is significantly better than MDE-HYB ($p < 0.05$). †: ($p < 0.1$).

Table 3.8: Comparison between MDE-HYB and MDE with AMT Real Workers

GT PER WORKER	MDE-HYB	MDE	MDE-HYB IMPROV
5	**0.060**	**0.060**	0.0%
10	**0.053**	0.060	12.1%**
15	**0.046**	0.059	23.3%**
20	**0.038**	0.060	36.2%**
25	**0.038**	0.059	36.5%**
30	**0.035**	0.059	40.5%**
50	**0.027**	0.060	54.2%**
100	**0.015**	0.059	74.5%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the variant MDE. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$).

Table 3.9: MDE-HYB's Performance Relative to Benchmarks (LogitBoost used for inference by MDE-HYB, GM-GTand GT-ALL remained unchanged) for Low-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.043**	0.142	69.5%**	0.167	74.0%**	0.295	85.3%**
	10	**0.038**	0.106	64.0%**	0.164	76.9%**	0.293	87.0%**
	15	**0.040**	0.090	55.0%**	0.162	75.0%**	0.292	86.2%**
	20	**0.036**	0.078	54.0%**	0.160	77.5%**	0.292	87.6%**
	25	**0.036**	0.068	47.4%**	0.158	77.2%**	0.290	87.6%**
	30	**0.038**	0.062	38.7%**	0.157	75.8%**	0.289	86.8%**
	50	0.048	**0.047**	-2.1%	0.152	68.3%**	0.285	83.0%**
	100	**0.033**	0.034	2.9%	0.145	77.1%**	0.275	87.9%**
	300	**0.018**	0.019	4.3%*	0.122	84.8%**	0.236	92.2%**
Movie	5	**0.031**	0.132	76.3%**	0.150	79.1%**	0.292	89.3%**
	10	**0.023**	0.103	77.7%**	0.127	82.0%**	0.291	92.1%**
	15	**0.019**	0.090	79.3%**	0.113	83.6%**	0.289	93.6%**
	20	**0.019**	0.076	75.5%**	0.106	82.5%**	0.289	93.6%**
	25	**0.017**	0.071	76.4%**	0.100	83.4%**	0.287	94.2%**
	30	**0.016**	0.065	75.0%**	0.096	83.2%**	0.286	94.3%**
	50	**0.016**	0.050	68.0%**	0.084	81.2%**	0.282	94.3%**
	100	**0.019**	0.035	46.4%**	0.074	74.6%**	0.270	93.0%**
	300	**0.012**	0.018	32.5%**	0.055	77.2%**	0.223	94.5%**
Spam	5	**0.019**	0.133	85.9%**	0.042	54.8%**	0.259	92.7%**
	10	**0.018**	0.103	82.8%**	0.034	47.1%**	0.254	93.0%**
	15	**0.017**	0.086	80.7%**	0.030	45.8%**	0.248	93.3%**
	20	**0.017**	0.078	78.6%**	0.029	40.9%**	0.241	93.1%**
	25	**0.017**	0.065	74.5%**	0.027	37.5%**	0.236	93.0%**
	30	**0.016**	0.062	74.5%**	0.025	37.6%**	0.231	93.1%**
	50	**0.016**	0.044	64.1%**	0.021	27.3%**	0.208	92.4%**
	100	**0.014**	0.027	47.3%**	0.015	2.0%	0.149	90.6%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the other method is significantly better than MDE-HYB ($p < 0.05$). †: ($p < 0.1$).

Table 3.10: MDE-HYB's Performance Relative to Benchmarks (LogitBoost used for inference by MDE-HYB, GM-GT and GT-ALL remained unchanged) for High-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.063**	0.119	47.6%**	0.291	78.4%**	0.145	56.7%**
	10	**0.049**	0.090	45.4%**	0.284	82.6%**	0.144	65.8%**
	15	**0.043**	0.073	40.4%**	0.281	84.6%**	0.143	69.9%**
	20	**0.041**	0.059	31.1%**	0.276	85.2%**	0.143	71.5%**
	25	**0.039**	0.054	27.8%**	0.274	85.8%**	0.143	72.8%**
	30	**0.035**	0.050	30.8%**	0.272	87.1%**	0.142	75.5%**
	50	0.038	**0.037**	-2.9%	0.265	85.4%**	0.140	72.6%**
	100	**0.026**	0.026	2.1%	0.252	89.8%**	0.135	80.9%**
	300	**0.014**	**0.014**	-0.4%	0.211	93.2%**	0.116	87.6%**
Movie	5	**0.029**	0.120	75.6%**	0.259	88.7%**	0.143	79.5%**
	10	**0.023**	0.083	72.4%**	0.221	89.5%**	0.142	83.91%**
	15	**0.020**	0.071	71.6%**	0.195	89.7%**	0.143	85.8%**
	20	**0.018**	0.062	70.7%**	0.183	90.0%**	0.141	87.1%**
	25	**0.016**	0.055	70.4%**	0.173	90.5%**	0.141	88.3%**
	30	**0.016**	0.049	68.1%**	0.166	90.5%**	0.141	88.8%**
	50	**0.015**	0.039	61.1%**	0.147	89.7%**	0.139	89.0%**
	100	**0.015**	0.026	42.5%**	0.128	88.1%**	0.133	88.5%**
	300	**0.011**	0.014	17.9%**	0.095	88.3%**	0.109	89.9%**
Spam	5	**0.018**	0.121	85.4%**	0.072	75.4%**	0.125	85.8%**
	10	**0.014**	0.083	82.6%**	0.058	75.3%**	0.124	88.3%**
	15	**0.015**	0.069	78.5%**	0.051	71.2%**	0.121	87.6%**
	20	**0.015**	0.059	75.4%**	0.048	69.2%**	0.117	87.6%**
	25	**0.014**	0.054	73.6%**	0.045	68.1%**	0.114	87.7%**
	30	**0.013**	0.046	70.7%**	0.041	67.7%**	0.111	88.0%**
	50	**0.014**	0.036	62.2%**	0.034	59.7%**	0.101	86.5%**
	100	**0.012**	0.020	39.6%**	0.022	44.3%**	0.072	83.2%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB using LogitBoost over an alternative. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$). ††: the other method is significantly better than MDE-HYB ($p < 0.05$). †: ($p < 0.1$).

Table 3.11: MDE-HYB's Performance Relative to Benchmarks (LogitBoost used for inference by MDE-HYB, GM-GT and GT-ALL remained unchanged) for AMT Real Workers

GT PER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
5	**0.059**	0.096	38.0%**	0.062	4.9%**	0.105	43.4%**
10	**0.055**	0.068	20.3%**	0.060	9.7%**	0.102	46.5%**
15	**0.046**	0.056	18.0%**	0.059	22.4%**	0.099	54.1%**
20	**0.038**	0.046	18.4%**	0.057	33.6%**	0.097	60.9%**
25	**0.036**	0.042	13.8%**	0.055	35.3%**	0.094	61.9%**
30	**0.035**	0.038	8.7%**	0.054	35.6%**	0.091	62.1%**
50	**0.027**	0.027	0.2%	0.047	41.6%**	0.081	66.4%**
100	**0.015**	0.015	0.0%	0.031	50.6%**	0.054	71.9%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB using LogitBoost over an alternative. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

3.2.3.2 Comparisons with a model-specific alternative with no ground truth.

As discussed in the literature review, [204] develop a method that is designed to address a different problem and in different settings from the ones we consider here, but that can be applied to infer workers' qualities. As we discussed, the method has several limitations that render its performance in our problem settings noncompetitive, including in particular that it does not exploit the availability of limited ground truth. Hence, we compare this approach and MDE-HYB in settings that are least advantageous to MDE-HYB. Our results, reported in Table 3.12, show that even in this setting, MDE-HYB yields a superior assessment of experts' accuracies.

Table 3.12: Comparison to Tanno et al.'s Alternative Baseline with GT per worker = 5

TASK	MDE-HYB (ours)	TANNO ET AL. Best Architecture	MDE-HYB IMPROV
Audit - Low Quality	**0.041**	0.062	33.6%**
Audit - high Quality	**0.062**	0.162	61.3%**
Movie - Low Quality	**0.023**	0.202	88.6%**
Movie - high Quality	**0.029**	0.04	26.4%**
Spam - Low Quality	**0.016**	0.027	39.6%**
Spam - high Quality	**0.017**	0.035	49.4%**
AMT - Real Workers	**0.06**	0.062	3.87%*

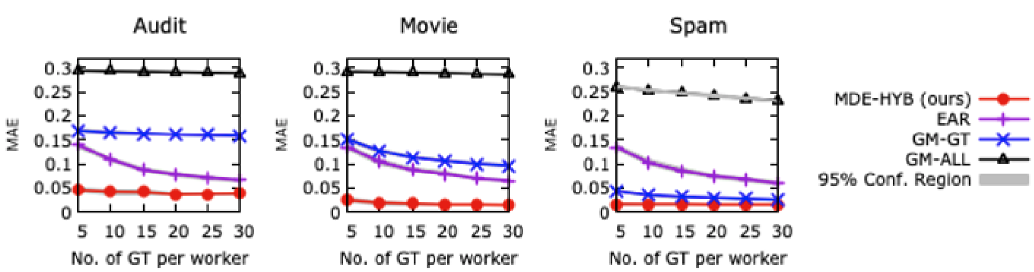
Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over Tanno et al.'s best architecture. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

3.2.3.3 Results for Correlated Experts' Errors.

Experts' error rates may at times be driven by properties of the decision tasks themselves, and in this case, all experts may exhibit a higher rate of error for a given set of instances. For example, such conditions correspond to settings where physicians have a higher likelihood of misdiagnosing a certain (e.g., rare) disease that all physicians have less experience diagnosing. In this section, we consider such settings in which all experts exhibit higher error rates for certain decision instances, relative to other instances. In these experiments, we considered low-quality experts, all of whom exhibited a 0.2 higher likelihood of error for a given subset of decision instances. Thus, given that the experts' overall accuracies ranged between 61% and 80%, they exhibited lower accuracy for the same set of instances, ranging between 41% and 60%. All experts exhibited an increased error rate for instances that have a feature value larger than the 90^{th} percentile of a given continuous feature.[6] The workers' overall accuracy rate remained the same as before. (Hence, workers exhibited increased accuracy for all other instances.)

Our results, shown in Figure 3.5, indicate that, in this setting as well, MDE-HYB either considerably outperforms the alternatives or otherwise exhibits comparable performance to them.

Figure 3.5: MDE-HYB's Performance Relative to Benchmarks given Correlated Experts' Errors



MAE measure for experts' accuracy estimation errors (measured across 50 repetitions) for our MDE-HYB approach and the baseline approaches when decision errors are correlated. Results are shown for a varying number of ground truth instances and different datasets. The grey shaded region shows the 95% confidence level for each method.

3.2.3.4 Illustration of Potential Practical Implications.

We report a simple analysis that illustrates the potential economic and societal effects of using MDE-HYB in a market setting. Specifically, we considered a scenario in which patients wish to receive diagnostic advice from expert

[6]The continuous features of age, star, and word frequency (word_freq_all) were used for the Audit, Movie, and Spam datasets, respectively.

physicians with a diagnostic accuracy of at least 90%, and where patients select experts based on the experts' diagnostic accuracy estimations produced either by MDE-HYB or the best alternative, EAR. We assessed the resulting misdiagnosis rates, and we discuss the subsequent costs in this section. For the purpose of this evaluation, we considered high-quality experts and a setting where MDE-HYB yields moderate improvement relative to (EAR): We used the Audit dataset where MDE-HYB yields the most conservative benefits, given the low predictability in this context. Further, we assumed 30 ground truth instances per worker (i.e., ground truth is not highly scarce), thus benefiting the EAR alternative. As before, we conducted 50 repetitions and used the assessments of the two methods in each experiment to inform patients' choices.

Diagnostic errors are a major patient safety challenge in the United States [194] and are estimated to lead to tens of thousands of deaths in U.S. hospitals alone (with estimates ranging from 40,000 to 80,000) [94]. Further, an additional 40.7% of diagnostic error-related adverse events result in serious disabilities [135]. In terms of economic costs, the estimated inflation-adjusted, 25-year sum of diagnosis-related health care payments has been reported to be $38.8 billion, where diagnostic errors accounted for 35.2% of the total payments [184].

Our analysis results indicate that the average misdiagnosis rate when MDE-HYB is used to select experts drops by 40.7% compared to when they are selected by the best alternative (dropping from a 4.7% misdiagnosis rate with EAR to 2.8% with MDE-HYB). Importantly, a 40.7% drop in misdiagnosis rates corresponds to a significant and practical improvement in the number of patients affected, the loss of lives, and the overall economic loss.

3.2.3.5 Sensitivity Analyses.

We explored the sensitivity of MDE-HYB's performance by varying different aspects of the settings. We report our findings in this section.

Varying the number of ground truth instances per worker. We evaluated MDE-HYB in a setting where different expert workers have a different number of ground truth labels. We found that under this setting, MDE-HYB consistently outperformed the best alternative. Results are reported in Tables 3.13 to 3.15.

Table 3.13: MDE-HYB and Benchmarks Performance measured by MAE when low quality workers have different amount of ground truth

DATASET	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	**0.049**	0.113	56.4%**	0.164	70.0%**	0.293	83.2%**
Movie	**0.029**	0.112	73.6%**	0.123	76.1%**	0.291	89.9%**
Spam	**0.016**	0.114	85.5%**	0.035	52.8%**	0.255	93.5%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

The effect of the subset of the features available on performance. Recall that, based on our problem formulation, the information available as historical data for MDE-HYB may be either a subset or a superset of the feature

Table 3.14: MDE-HYB and Benchmarks Performance measured by MAE when high quality workers have different amount of ground truth

DATASET	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	**0.079**	0.095	17.1%**	0.283	72.1%**	0.144	45.1%**
Movie	**0.037**	0.097	62.1%**	0.213	82.7%**	0.143	74.1%**
Spam	**0.017**	0.093	82.0%**	0.058	71.0%**	0.122	86.2%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Table 3.15: MDE-HYB and Benchmarks Performance measured by MAE when AMT workers have different amount of ground truth

DATASET	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Amazon	**0.054**	0.075	27.5%**	0.061	10.4%**	0.102	46.8%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

set the human decision maker used to arrive at a decision. We thus conducted studies in which 25% of features were removed and could not be used by MDE-HYB. We found that in this setting MDE-HYB continued to produce reliable assessments and remained the method of choice, relative to the best alternative. Importantly, given that MDE-HYB balances the benefits of relying on either MDE or EAR or both, it exhibits robust performance when predictability is limited. We report the results of this analysis in Tables 3.16 and 3.17

Robustness to hyper parameter settings. We also evaluated MDE-HYB's robustness by using different hyper parameters that corresponded to these guidelines. We found that MDE-HYB is robust and provides consistent performance.

3.2.3.6 Evaluations with Other Experts' Error Distribution.

In this paper, we have reported results for different contexts in which experts' errors are driven by different and unknown factors. Specifically, we presented results with human decision makers recruited via AMT, where the underlying drivers of expert errors are both unknown and can vary across experts. In Section 3.2.3.3, we presented results when experts errors are correlated so that *all* experts are more likely to err when evaluating instances with certain properties.

We complemented these previous results with additional evaluations which explore MDE-HYB's performance when each expert exhibits an increased

Table 3.16: Comparison between MDE-HYB and EAR when features are randomly removed by 25% with Low and High Quality Workers

DATASET	GT PER WORKER	Low Quality MDE-HYB	Low Quality EAR	Low Quality MDE-HYB IMPROV	High Quality MDE-HYB	High Quality EAR	High Quality MDE-HYB IMPROV
Audit	5	**0.041**	0.142	72.4%**	**0.062**	0.121	51.7%**
	10	**0.041**	0.106	60.4%**	**0.052**	0.081	40.9%**
	15	**0.043**	0.087	48.7%**	**0.045**	0.072	36.3%**
	20	**0.036**	0.077	51.1%**	**0.042**	0.060	30.2%**
	25	**0.036**	0.071	48.6%**	**0.038**	0.056	28.2%**
	30	**0.037**	0.062	37.8%**	**0.035**	0.047	25.1%**
	50	**0.043**	0.051	6.9%**	**0.036**	0.038	3.31%**
	100	**0.034**	0.035	0.0%	**0.026**	0.027	0.0%
	300	**0.019**	**0.019**	0.0%	**0.014**	**0.014**	0.0%
Movie	5	**0.023**	0.139	80.1%**	**0.029**	0.120	74.9%**
	10	**0.017**	0.106	82.9%**	**0.022**	0.088	74.0%**
	15	**0.019**	0.089	80.0%**	**0.019**	0.071	69.1%**
	20	**0.016**	0.081	79.1%**	**0.017**	0.061	70.3%**
	25	**0.014**	0.071	76.9%**	**0.016**	0.054	67.0%**
	30	**0.015**	0.063	75.7%**	**0.016**	0.050	67.5%**
	50	**0.013**	0.049	69.0%**	**0.014**	0.038	58.5%**
	100	**0.015**	0.034	50.8%**	**0.015**	0.026	41.3%**
	300	**0.011**	0.018	36.1%**	**0.009**	0.013	20.9%**
Spam	5	**0.016**	0.138	87.0%**	**0.017**	0.119	84.0%**
	10	**0.015**	0.099	82.6%**	**0.016**	0.085	78.5%**
	15	**0.015**	0.086	80.8%**	**0.015**	0.068	77.2%**
	20	**0.015**	0.076	79.5%**	**0.015**	0.060	72.6%**
	25	**0.015**	0.070	77.0%**	**0.015**	0.053	69.8%**
	30	**0.015**	0.061	74.9%**	**0.015**	0.046	66.7%**
	50	**0.014**	0.047	68.0%**	**0.015**	0.033	56.0%**
	100	**0.014**	0.027	44.9%**	**0.012**	0.021	40.6%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative EAR. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Table 3.17: Comparison between MDE-HYB and EAR when features are randomly removed by 25% with AMT Real Workers

GT PER WORKER	MDE-HYB	EAR	MDE-HYB IMPROV
5	**0.06**	0.096	37.2%**
10	**0.053**	0.068	22.3%**
15	**0.045**	0.056	19.0%**
20	**0.038**	0.046	18.0%**
25	**0.037**	0.042	10.1%**
30	**0.035**	0.038	7.4%**
50	**0.027**	**0.027**	-0.1%
100	**0.015**	**0.015**	0.0%

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative EAR. MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

error rate, relative to the expert's average error rate when the expert encounters different types of decision instances. This setting reflects contexts, such as when one physician may exhibit a higher error rate when diagnosing older (or younger) patients, while another physician may exhibit an increased error rate when diagnosing patients of a certain ethnicity. We detail the simulation settings and report the results in Tables 3.18 and 3.19. We find that in this context MDE-HYB remains the method of choice and consistently outperforms the best alternative.

3.2.3.7 Additional Benchmarks.

In this section, we report a comparison between MDE-HYB and several additional benchmarks. In particular, we considered a broad set of methods that use the available ground truth and experts' noisy labels in different ways to infer the most likely correct decisions, based on which experts' accuracies can be assessed. As shown, MDE-HYB consistently outperformed these benchmarks. Further, the benchmarks we considered in this supplemental analysis were also less competitive than the best alternatives we have already considered in this paper. In Tables 3.20 and 3.21, we report a comparison between MDE-HYB performance and the performance of additional benchmark approaches. We discuss each of these benchmarks below.

The first benchmark is based on a KNN classifier. This benchmark was trained using all the instances with ground truth labels. Subsequently each expert worker was scored according to the level of agreement between

Table 3.18: New Simulation: MDE-HYB and Benchmarks Performance measured by MAE for Low-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.055**	0.137	59.9%**	0.173	68.2%**	0.294	81.2%**
	10	**0.059**	0.109	45.4%**	0.170	65.0%**	0.293	79.7%**
	15	**0.065**	0.091	28.3%**	0.168	61.1%**	0.292	77.7%**
	20	**0.058**	0.076	24.5%**	0.165	65.1%**	0.291	80.2%**
	25	**0.049**	0.068	27.9%**	0.166	70.5%**	0.290	83.1%**
	30	**0.049**	0.064	23.3%**	0.163	69.6%**	0.289	82.9%**
	50	**0.047**	0.049	3.8%**	0.157	69.7%**	0.285	83.3%**
	100	0.036	0.036	0.0%	0.151	76.2%**	0.275	87.0%**
	300	0.019	0.019	0.0%	0.126	84.8%**	0.235	91.8%**
Movie	5	**0.023**	0.146	84.3%**	0.150	84.7%**	0.292	92.1%**
	10	**0.019**	0.112	83.3%**	0.127	85.3%**	0.291	93.6%**
	15	**0.017**	0.092	81.8%**	0.113	85.2%**	0.290	94.2%**
	20	**0.015**	0.080	81.3%**	0.105	85.8%**	0.289	94.8%**
	25	**0.014**	0.073	80.0%**	0.100	85.5%**	0.288	95.0%**
	30	**0.014**	0.068	79.3%**	0.096	85.2%**	0.286	95.1%**
	50	**0.014**	0.058	75.6%**	0.085	83.4%**	0.282	95.0%**
	100	**0.017**	0.042	57.9%**	0.073	76.2%**	0.270	93.5%**
	300	**0.012**	0.022	44.9%**	0.056	77.7%**	0.222	94.4%**
Spam	5	**0.018**	0.142	87.6%**	0.042	57.7%**	0.257	93.1%**
	10	**0.016**	0.106	84.8%**	0.033	51.8%**	0.251	93.6%**
	15	**0.016**	0.094	82.6%**	0.031	46.6%**	0.246	93.3%**
	20	**0.015**	0.076	80.0%**	0.028	45.5%**	0.240	93.6%**
	25	**0.015**	0.068	77.6%**	0.026	41.2%**	0.235	93.5%**
	30	**0.015**	0.065	76.4%**	0.025	38.3%**	0.229	93.3%**
	50	**0.015**	0.052	70.5%**	0.021	26.1%**	0.207	92.6%**
	100	0.015	0.027	45.4%**	0.014	-7.4%	0.149	90.1%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Table 3.19: New Simulation: MDE-HYB and Benchmarks Performance measured by MAE for High-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	EAR	MDE-HYB IMPROV	GM-GT	MDE-HYB IMPROV	GM-ALL	MDE-HYB IMPROV
Audit	5	**0.074**	0.123	39.6%**	0.295	74.8%**	0.144	48.4%**
	10	**0.067**	0.087	23.2%**	0.29	76.9%**	0.144	53.3%**
	15	**0.062**	0.07	12.1%**	0.287	78.4%**	0.143	56.8%**
	20	**0.058**	0.06	2.8%**	0.283	79.4%**	0.143	59.1%**
	25	**0.046**	0.054	14.0%**	0.281	83.6%**	0.142	67.6%**
	30	**0.041**	0.052	20.8%**	0.278	85.1%**	0.142	70.9%**
	50	0.038	0.039	1.2%	0.272	85.9%**	0.14	72.7%**
	100	0.027	0.027	0.0%	0.258	89.6%**	0.135	80.1%**
	300	0.015	0.015	0.0%	0.215	93.0%**	0.115	86.9%**
Movie	5	**0.028**	0.12	76.9%**	0.259	89.3%**	0.143	80.6%**
	10	**0.023**	0.089	73.8%**	0.22	89.5%**	0.143	83.7%**
	15	**0.019**	0.07	72.7%**	0.196	90.3%**	0.142	86.6%**
	20	**0.017**	0.061	71.5%**	0.183	90.5%**	0.141	87.7%**
	25	**0.016**	0.055	70.7%**	0.173	90.7%**	0.141	88.5%**
	30	**0.015**	0.052	70.4%**	0.166	90.8%**	0.14	89.1%**
	50	**0.016**	0.037	57.5%**	0.147	89.4%**	0.138	88.7%**
	100	**0.016**	0.028	45.1%**	0.128	87.8%**	0.132	88.2%**
	300	**0.009**	0.012	30.7%**	0.095	91.1%**	0.109	92.2%**
Spam	5	**0.019**	0.119	84.3%**	0.072	74.1%**	0.123	84.8%**
	10	**0.016**	0.088	81.6%**	0.059	72.3%**	0.12	86.5%**
	15	**0.016**	0.071	78.2%**	0.052	70.0%**	0.118	86.8%**
	20	**0.015**	0.057	73.2%**	0.048	68.4%**	0.115	86.7%**
	25	**0.015**	0.052	70.9%**	0.045	65.9%**	0.112	86.4%**
	30	**0.015**	0.049	69.8%**	0.042	64.7%**	0.11	86.5%**
	50	**0.017**	0.036	53.7%**	0.034	50.5%**	0.099	82.9%**
	100	**0.013**	0.021	39.4%**	0.022	42.5%**	0.071	82.2%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of experts' accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

the worker's decisions and the KNN (k=3) model predictions. The KNN-based benchmark was chosen given the popularity of using simple methods, that are less likely to overfit the data, when there is limited ground truth data, such as in the case of the well-known "cold-start problem". The results for this benchmark are reported in Tables 3.20 and 3.21 in the column titled KNN.

Another baseline that we evaluated is a baseline that aims to benefit from all the data (both ground truth and noisy labels) by learning a feature vector representation (embedding) function using an Autoencoder that was trained using the features of all data instances (both ground truth and noisy labeled instances). We then applied the trained Autoencoder to generate embeddings for the ground truth instances and classified each (non ground truth) decision instance according to its corresponding embedding vector's Euclidean distance from the (average) vector representation for ground truth instances in each class. Expert workers were scored according to the extent that their decisions corresponded to the classifications made by this classifier. This baseline is inspired from few shot learning papers such as [29] which learnt an embedding function to represent the features, and subsequently used the similarity/distance between query and support instances in their modelling. The performance of this baseline approach is detailed in the column titled "AUTOENCODER" in tables 3.20 and 3.21, for low and high quality workers respectively.

An additional baseline is based on a semi-supervised learning technique. Semi-supervised learning methods are useful in case of "incomplete supervision" tasks [235] when there is only scarce correctly labeled data and most instances are unlabelled. Specifically, we applied the established LABEL-PROPAGATION approach reported by [236]. We used the method to predict the labels of the expert workers' decisions and scored the expert workers according to the extent that their decisions match the model predictions. Results for this approach are reported in tables 3.20 and 3.21 in column "Label Prop.".

Additionally, we implemented a baseline which we refer to as "Similarity-Based Classifier". To implement this baseline, we split the instances with ground truth into two class groups based on their ground truth decision value. Then for each of the two classes, we calculate a feature vector that is based on the (per feature) class mean. Subsequently, we calculate the Euclidian distance between the feature vector of a non-ground truth instance and the two, previously calculated, per-class mean vectors. We then classify the non-ground truth instance to the class with the lowest distance. Lastly, we measure each expert's decision accuracy according to the level of agreement between the experts' decisions and the predicted decision by this approach. The results are presented in tables 3.20 and 3.21 (Column titled "Similar-Based").

Finally, we implemented another baseline approach which we refer to as Distribution-Based Scoring (Dist-Based). This approach posits that the features of each (correct) decision class have a multidimensional normal distribution. Based on this approach we first split the instances with ground truth into two class groups based on the ground truth decision value. Then for each class, we estimate the (per-feature) distribution parameters. Subsequently, we score each non-ground truth instance according to the percentile difference from the feature distribution mean of each class. Next, for each non-ground truth instance, we determine that the predicted decision belongs to the class with the highest score. Finally, we measure each expert's decision accuracy according to the level of agreement between the experts' decisions and the predicted decision by this approach. The results are presented in tables 3.20 and 3.21 (Column titled "Dist-Based").

As observed MDE-HYB consistently and significantly outperforms all the additional baselines reported. As such it remains the method of choice. MDE-HYB superiority is due to its effective use of both ground truth and noisy labels, as well as its dedicated procedures to obtain accurate assessmentx from model-based predictions. In contrast the baseline approaches naively rely on inferring a model which is primarily-based or solely-based on scarce ground truth instances. This naive reliance on comparing model predictions to experts' decisions, without any dedicated procedure to obtain assessments, result in the baseline approaches poor performance.

Table 3.20: MDE-HYB's Performance Relative to Additional Benchmarks for Low-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	KNN	MDE-HYB IMPROV	AUTOEN-CODER	MDE-HYB IMPROV	LABEL-PROP.	MDE-HYB IMPROV	DISTRIB. BASED	MDE-HYB IMPROV	SIMILAR. BASED	MDE-HYB IMPROV
Audit	5	**0.041**	0.182	77.4%**	0.18	77.2%**	0.193	78.8%**	0.168	75.5%**	0.197	79.2%**
	10	**0.041**	0.181	77.5%**	0.178	77.1%**	0.192	78.8%**	0.164	75.1%**	0.197	79.4%**
	15	**0.043**	0.18	76.4%**	0.177	76.0%**	0.191	77.7%**	0.16	73.4%**	0.197	78.4%**
	20	**0.036**	0.177	79.7%**	0.177	79.7%**	0.19	81.1%**	0.158	77.3%**	0.197	81.8%**
	25	**0.036**	0.177	79.7%**	0.177	79.6%**	0.189	80.9%**	0.155	76.8%**	0.196	81.6%**
	30	**0.037**	0.177	79.3%**	0.176	79.2%**	0.188	80.6%**	0.156	76.6%**	0.196	81.4%**
	50	**0.043**	0.173	75.3%**	0.174	75.5%**	0.184	76.8%**	0.153	72.1%**	0.197	78.3%**
	100	**0.034**	0.166	79.4%**	0.168	79.5%**	0.175	80.4%**	0.147	76.6%**	0.197	82.6%**
	300	**0.019**	0.14	86.2%**	0.143	86.5%**	0.147	86.9%**	0.125	84.6%**	0.185	89.6%**
Movie	5	**0.023**	0.192	88.0%**	0.197	88.3%**	0.198	88.3%**	0.139	83.4%**	0.185	87.5%**
	10	**0.017**	0.188	90.8%**	0.195	91.1%**	0.2	91.4%**	0.123	86.0%**	0.178	90.3%**
	15	**0.019**	0.182	89.8%**	0.192	90.3%**	0.197	90.6%**	0.115	83.8%**	0.175	89.4%**
	20	**0.016**	0.183	91.4%**	0.191	91.8%**	0.198	92.1%**	0.111	85.9%**	0.172	90.9%**
	25	**0.014**	0.181	92.0%**	0.188	92.3%**	0.196	92.6%**	0.107	86.5%**	0.17	91.5%**
	30	**0.015**	0.179	91.6%**	0.187	92.0%**	0.196	92.4%**	0.103	85.5%**	0.169	91.1%**
	50	**0.013**	0.175	92.3%**	0.185	92.8%**	0.194	93.1%**	0.095	85.9%**	0.166	91.9%**
	100	**0.015**	0.164	90.7%**	0.177	91.4%**	0.183	91.7%**	0.086	82.3%**	0.162	90.6%**
	300	**0.011**	0.129	91.5%**	0.138	92.1%**	0.133	91.8%**	0.065	83.2%**	0.134	91.9%**
Spam	5	**0.016**	0.082	80.0%**	0.128	87.2%**	0.141	88.4%**	0.046	63.9%**	0.129	87.2%**
	10	**0.015**	0.071	78.7%**	0.122	87.7%**	0.131	88.6%**	0.041	63.4%**	0.123	87.8%**
	15	**0.015**	0.066	76.8%**	0.123	87.5%**	0.124	87.7%**	0.04	61.3%**	0.123	87.6%**
	20	**0.015**	0.06	75.0%**	0.121	87.6%**	0.116	87.0%**	0.039	61.1%**	0.121	87.6%**
	25	**0.015**	0.057	73.2%**	0.118	87.0%**	0.108	85.9%**	0.036	57.9%**	0.118	87.1%**
	30	**0.015**	0.054	71.9%**	0.114	86.7%**	0.102	85.1%**	0.036	58.2%**	0.114	86.8%**
	50	**0.014**	0.043	67.3%**	0.102	86.2%**	0.082	82.8%**	0.032	55.3%**	0.103	86.3%**
	100	**0.014**	0.028	47.7%**	0.073	80.3%**	0.053	72.6%**	0.024	39.0%**	0.076	81.0%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

Table 3.21: MDE-HYB's Performance Relative to Additional Benchmarks for High-quality Workers

DATASET	GT PER WORKER	MDE-HYB (ours)	KNN	MDE-HYB IMPROV	AUTOEN-CODER	MDE-HYB IMPROV	LABEL-PROP.	MDE-HYB IMPROV	DISTRIB. BASED	MDE-HYB IMPROV	SIMILAR. BASED	MDE-HYB IMPROV
Audit	5	**0.062**	0.317	80.3%**	0.312	80.0%**	0.334	81.3%**	0.291	78.5%**	0.344	81.9%**
	10	**0.052**	0.314	83.6%**	0.307	83.2%**	0.335	84.6%**	0.283	81.8%**	0.343	85.0%**
	15	**0.045**	0.313	85.6%**	0.307	85.3%**	0.332	86.4%**	0.278	83.8%**	0.344	86.9%**
	20	**0.042**	0.309	86.4%**	0.306	86.3%**	0.331	87.3%**	0.274	84.7%**	0.342	87.7%**
	25	**0.038**	0.308	87.7%**	0.305	87.6%**	0.33	88.6%**	0.272	86.1%**	0.341	88.9%**
	30	**0.035**	0.306	88.6%**	0.304	88.5%**	0.328	89.4%**	0.27	87.1%**	0.34	89.8%**
	50	**0.036**	0.3	87.9%**	0.301	87.9%**	0.319	88.6%**	0.264	86.3%**	0.341	89.3%**
	100	**0.026**	0.287	90.8%**	0.291	90.9%**	0.304	91.3%**	0.254	89.6%**	0.342	92.3%**
	300	**0.014**	0.241	94.1%**	0.249	94.2%**	0.253	94.3%**	0.216	93.3%**	0.322	95.5%**
Movie	5	**0.029**	0.334	91.2%**	0.343	91.4%**	0.348	91.5%**	0.24	87.7%**	0.32	90.8%**
	10	**0.022**	0.325	93.2%**	0.341	93.5%**	0.348	93.6%**	0.213	89.6%**	0.31	92.8%**
	15	**0.019**	0.319	94.0%**	0.334	94.2%**	0.348	94.5%**	0.199	90.3%**	0.301	93.6%**
	20	**0.017**	0.318	94.5%**	0.334	94.8%**	0.346	95.0%**	0.193	91.0%**	0.3	94.2%**
	25	**0.016**	0.315	94.8%**	0.328	95.0%**	0.345	95.3%**	0.185	91.2%**	0.295	94.5%**
	30	**0.016**	0.311	94.9%**	0.326	95.2%**	0.345	95.4%**	0.178	91.1%**	0.293	94.6%**
	50	**0.014**	0.305	95.4%**	0.322	95.6%**	0.339	95.8%**	0.166	91.5%**	0.288	95.1%**
	100	**0.015**	0.284	94.8%**	0.307	95.2%**	0.318	95.4%**	0.149	90.1%**	0.279	94.7%**
	300	**0.009**	0.222	95.8%**	0.238	96.1%**	0.23	95.9%**	0.113	91.7%**	0.233	96.0%**
Spam	5	**0.017**	0.139	87.5%**	0.223	92.2%**	0.25	93.0%**	0.078	77.5%**	0.221	92.1%**
	10	**0.016**	0.122	87.2%**	0.216	92.8%**	0.226	93.1%**	0.07	77.6%**	0.22	92.9%**
	15	**0.015**	0.112	86.3%**	0.212	92.8%**	0.214	92.8%**	0.066	76.9%**	0.213	92.8%**
	20	**0.015**	0.104	85.9%**	0.208	93.0%**	0.198	92.6%**	0.063	76.9%**	0.21	93.0%**
	25	**0.015**	0.098	84.9%**	0.202	92.7%**	0.188	92.1%**	0.061	75.7%**	0.204	92.7%**
	30	**0.015**	0.092	84.2%**	0.199	92.7%**	0.176	91.7%**	0.061	76.1%**	0.198	92.7%**
	50	**0.015**	0.075	80.6%**	0.18	91.9%**	0.144	89.8%**	0.054	73.1%**	0.179	91.8%**
	100	**0.012**	0.047	73.3%**	0.129	90.4%**	0.091	86.3%**	0.038	67.6%**	0.129	90.3%**

Experts' accuracy estimation errors. Values show Mean Absolute Error (MAE). MDE-HYB IMPROV shows the improvement of MDE-HYB over the alternative; MDE-HYB yields substantially better and otherwise comparable estimations of workers accuracies. ** MDE-HYB is statistically significantly better ($p < 0.05$), *: ($p < 0.1$).

3.3 Limitations and Future Work

We consider common experts' settings in practice, where expert workers' decision accuracy in expectation is higher than that of a random draw. However, similar to all other machine-learning-based methods, our approach may not produce an advantageous performance under unlikely pathological conditions, such as when most experts make a decision entirely at random, or when most experts are adversarial and intentionally invert their decisions. Developing methods to extend MDE-HYB to assess the decision accuracy of non-expert workers that may be adversarial, or that in expectation, may be less accurate than a random draw, provide interesting avenues for future research.

Many machine learning methods do not lend themselves to closed-form analyses. In principle, the more complex the methods and the corresponding settings are, the less likely it is that a closed form representation is possible, or that necessary abstractions can yield meaningful insights toward real world settings. Indeed, our settings involve humans decisions, and our approach involves inductions, empirical measures, and statistical procedures applied to these methods. This complexity precludes a representation of our approach's behavior in closed form. Our results demonstrate the consistent performance of our approach: both that it is better than the alternatives considered and the range of results that can be expected across domains.

Our hope is that our work will motivate future research that builds on our framework and the empirical evaluations we report here, both to advance our understandings of how further improvements can be achieved and to promote the integration of these methods in practice. As the integration of machine learning in practice clearly demonstrates, such integration is essential to advance progress in practice and to identify challenges that arise in particular contexts. Thus, our work suggests several interesting directions for future research.

For example, experts' decisions are costly to acquire, and hence, ground truth in these settings is inherently costly as well. Given a limited budget for ground truth acquisitions (e.g., via a panel of experts), it is valuable to explore whether or when intelligent information acquisition approaches can identify the instances for which to acquire ground truth information, with the goal of meaningfully improving the assessment of workers' decision accuracies. Active learning traditionally has considered selectively acquiring labels when the goal is to improve the generalization performance of a model induced from the acquired sample. Thus, novel acquisition policies are needed to address the goal of improving the assessment of experts' decision accuracies produced by MDE-HYB or by future methods developed for this problem.

We consider experts' prediction accuracy, which is an integral aspect of experts' overall judgment quality. We focus on prediction of discrete outcomes, such as diagnostic decisions. We hope that our work inspires future research to identify new opportunities to assess other key aspects of experts' judgment quality. For instance, while we consider contexts where experts predict a discrete outcome, future work may build on our work to consider predictions of real values.

Similar to most innovations, increasing transparency in experts' markets may have a variety of implications, introducing both further progress and new challenges. We previously discussed key organizational tasks that this technology can inform and advance. Here, we note possible implications that may inspire future work on related challenges. In particular, interesting and related questions that may be explored pertain to identifying particularly productive ways to bring assessments of experts' accuracies to bear in organizational settings. For example, should ongoing feedback be provided to the experts themselves to inform them about their performance, and if so, how?

In addition, depending on the context in which management would bring assessment information to bear, such as to inform managers about decisions regarding workers' retraining, exploring if and how experts' decision performance may be affected by such practices would be useful. Studies have shown that assessments by peers or supervisors are affected by interpersonal relationships and biases [e.g. 23]. In our approach, peers' and supervisors' evaluations are not used, and thus, any interpersonal histories and biases cannot affect the evaluation. At the same time, it would be useful to establish how different ways in which machine-learning-based assessments are brought to bear, can affect experts' performances. For example, given experts' knowledge that their decisions are being evaluated over time, would experts tend to improve their performance? Would workers tend to be more diligent to exhibit a consistent performance over time? Are there conditions or kinds of tasks in which experts' performances might be undermined as a result of assessments?

Chapter 4

Cost-effectively Acquiring Data for Assessing Workers' Decision Accuracy

Previous studies in Chapter 2 (MDE) and 3 (MDE-HYB) assume that the scarce GT (or GS) is available with random sampling. In this study, we hope to achieve better performance with the same amount of scarce ground truth data or to reach the same performance with less GT data. Evidently, the acquisition of a large number of high-quality annotated datasets consume costly manpower, time-consuming, or difficult to obtain, making it unfeasible in fields especially in domains that require high levels of expertise, such as fraud detection, information extraction, medical diagnosis, etc. Therefore, it is valuable to investigate whether Active Learning (AL) type of techniques (select most useful examples to acquire GT label in order to enhance the model quality) can be used to reduce the expenses of acquiring costly GT data while retaining the powerful accuracy assessment of MDE-HYB. We instantiate uncertainty sampling with different measures, analyze the properties of the sampling strategies thus obtained, and compare them to improve the accuracy assessment. Thereby, in the case of given very limited acquisition budget, this AL inspired strategy improves the state of the art performance in assessing workers' decision quality.

4.1 Related Work

The emerge of active machine learning has been more than decades of years [150]. The two main steams of active learning include 1) stream-based selective sampling [7, 96, 147], in which unlabeled instances are presented one by one from the data source, and the learner has to decide whether the instance is informative enough that it should be acquired ground truth or be discarded; 2) the query strategies can be divided into several categories, including the uncertainty based approach [15, 104, 138, 174, 189, 206], diversity-based approach [16, 79, 91, 167], expected model change [76, 182, 188], and hybrid approaches [8, 193, 225]. This work belongs to uncertainty based approaches.

It's more computational efficient to query a batch of instances to acquire ground truth rather than a single one at each iteration, so the method proposed in this study utilizes batch mode strategy to avoid frequent training with little change in the training data.

Active learning has been used in many applications, e.g., drug discovery [178, 51, 47], chemistry optimizations [179], crowd-sourcing markets, and etc. The goal of traditional active learning in the applications is to select the most useful examples which if labeled would significantly boost the learning ability and enhance prediction performance, while this work is to improve the decision accuracy assessment.

Other related works including machine learning-based evaluation towards decision quality evaluation, experts making decision errors, and limited ground truth are listed in Chapter 2.2.

4.2 Cost-Effectively Machine-learning-based Decision quality Estimation (ce-mde)

This **C**ost-**E**ffectively **M**achine-learning-based **D**ecision quality **E**stimation (CE-MDE) is inspired by the query by committee ([189]), which selects the sample on which there will be the most disagreement among a consensus of multiple predictive models. CE-MDE defines a Decision Difficulty score (DD) which used as the weight of each instance to be sampled. The problem setting this approach considers is mostly the same as in Chapter 2 besides $GT = \{GT_k\}_1^K$ is not randomly selected from every expert's decision set but based on the distribution of the DD scores (This chapter shares the same notations as in Chapter 2)

DD score for each instance is given by:

$$DS_{X_i^k} = \frac{\sum_{j=1, X_i^k \notin S_j}^{K} B_j(X_i^k)_{\sim M(X_i^k)}}{\sum_{j=1, X_i^k \notin S_j}^{K} B_j(X_i^k)_{M(X_i^k)}} \tag{4.1}$$

The larger the DD score, the more difficult the instance is. Instead of only selecting most difficult samples, which makes the instances of $\{\{S_{sw_n}^c\}_1^N\}_1^C$ synthetic workers not representative of the real workers (referred to Algorithm MDEfrom line 2 to 6). CE-MDE will select both the difficult instances if $\{X_i^k, Y_i^k\} \in \{S_{\mathrm{w}_j}\}_1^K$ and $DS_{X_i^k} >= DD_{high\_threshold}$ where $DD_{high\_threshold}$ of which is the least difficult instance to be acquired GT label given an acquisition budget, and the easy instances if $\{X_i^k, Y_i^k\} \in \{S_{\mathrm{w}_j}\}_1^K$ and $DS_{X_i^k} <=$

$DD_{low\_threshold}$ to "acquire" PGT label ($M(X_i^k)$) which is given by the ensemble model M. There are two benefits of including both difficult and easy instances for $\{\{S^c_{sw_n}\}_1^N\}_1^C$ synthetic workers' instance sets: 1) $\{\{S^c_{sw_n}\}_1^N\}_1^C$ has larger size for producing mappings to infer workers' accuracy (referred to Algorithm MDEline 2-6 and line 11-13), 2) $\{\{S^c_{sw_n}\}_1^N\}_1^C$ are more representative of real workers' instance sets through because the instances are from different difficulty levels .

4.3 Results

We report the results in Tables 4.1 and 4.2. Our results show that CE-MDE achieves comparable or better performance as the MDE-HYB (assumes randomly sampling GT) if we acquire from 3 to 10 GT per worker for 20 workers. In the evaulation with AMT real workers, CE-MDE acheives comparable performance as the original MDE-HYB only using 1/3 amount of GT (CE-MDE's MAE: 0.045 acquiring 5 GT per worker compared to MDE-HYB's MAE: 0.046 acquiring 15 GT per worker).

We also evaluated other settings in which more GT (more than 10 GT per worker) can be acquired, because CE-MDE is not able to leverage EAR as MDE-HYB, CE-MDE's performance become wrose than MDE-HYB. In conclusion, we suggest that if given very limited amount of budget to buy no more than 10 GT per worker, CE-MDE's acquisition strategy is recommended; if more GT can be acquired, MDE-HYB with random sampling GT will produce stable results.

In the furture work, I hope to effectively leverage the predictions of EAR as MDE-HYB, so that CE-MDE is not restricted to the number of GT acquired.

Table 4.1: CE-MDE and original MDE-HYB Performance comparision

DATASET	GT-PE	CE-MDE	ORIGINAL MDE-HYB	CE-MDE IMPROV
Audit	**3**	0.035	0.041	14.2%
	5	0.035	0.041	14.2%
	10	0.029	0.043	32.9%
	15	0.047	0.043	-8.9%
Spam	**3**	0.016	0.016	0.1%
	5	0.016	0.016	1.5%
	10	0.015	0.015	1.4%
	15	0.015	0.015	2.8%

Table 4.2: CE-MDE and original MDE-HYB Performance comparision with AMT Real Workers

DATASET	GT-PE	CE-MDE	ORIGINAL MDE-HYB	CE-MDE IMPROV
Amazon	**3**	0.049	0.056	12.4%
	5	0.045	0.06	24.4%
	10	0.045	0.053	15.2%
	15	0.046	0.046	-0.3%

Chapter 5

Assessing Labelers' Biases with Scarce Ground Truth (gold standard)

5.1 Introduction

Across key domains, human expert assessments and crowd annotations are essential for labeling data to train machine learning models, and constitute a pathway through which human's biases are learned by algorithms. Once deployed, biased Machine Learning (ML) algorithms can have significant impact in human's lives in many realms, including healthcare, recruitment, promotion, and colleague admission, among others. In this research, we explore how to leverage scarce GT decisions (labels) to assess biases in human-generated labels. We propose a machine learning-based framework to produce a relative assessment of the extent of bias contained in labels produced by different labelers or sources, when GT labels are costly or difficult to acquire and thus available for only a small set of instances. For example, gold-standard labeled instances can be acquired from costly professional fact checkers examining online claims' veracity to constitute a gold-standard when assessing crowdsourced labels. The proposed methodology does not require overlap between the instances assessed by different labelers nor between these and the instances for which GT labels are available. After providing theoretical guarantees, we empirically show that our method outperforms or produces at least comparable results to several existing alternatives to assess biases present in human labels, including a commonly used benchmark relying on statistical parity, which we show may be misleading when humans (intentionally or unintentionally) produce poor quality orderings within protected groups. Our empirical results establish the performances that can be achieved across diverse settings, including settings that involve different data domains, labelers' (sources') biases, class or group distributions, and amounts of GT data. We also show the downstream value of our approach in improving the quality of ML algorithms induced from biased labels. The proposed approach lays the groundwork towards increased transparency in labelers' biases and offers an important building block towards mitigating algorithmic bias stemming from biased labels.

5.2 Related Work

The risks of learning from noisy or biased labels are a well-known concern in machine learning. In the context of crowdsourcing, the quality of the labels obtained has been subject to doubt [171], and the impact of different aggregation mechanisms when multiple labels are available per instance has been studied [44]. Separate lines of works develop algorithms for acquiring [81] and learning from noisy labels [139], with a large body of work studying the robustness of such approaches e.g., [159]. Crucially, these methods typically assume forms of noise that deviate from the scenarios in which multiple labelers share incorrect beliefs, which is particularly plausible when the goal is to assess labelers' bias, as these may be reflective of widely held societal stereotypes. Our research contributes to this body of work by proposing methodology to assess relative biases across labelers without assuming that the majority will be correct nor requiring the modification of the data collection process, which we achieve by leveraging a small disjoint pool of gold-standard labels.

The problem of assessing human bias and decision quality has been a subject of study across disciplines. There are works focusing on evaluating cognitive bias [48, 34, 1, 26] serving different goals and using different methodological approaches than ours, including measuring individual differences in cognitive biases, improving rational thinking and mediating decision biases emotionally or psychologically. Relatedly, there exist works evaluating decision makers' biases among individuals with special traits, for example, alcohol dependence (AD) [161]. Other related work addresses the problem of either ranking or directly assessing experts' overall decision accuracy with scarce gold standards, e.g., [53, 84]. However, these works do not consider assessing labelers'/decision-makers' biases; furthermore, [84] also do not consider how ground truth data can be brought to bear. In the context of crowdsourcing, researchers have estimated decision reliability based on workers' various behavioural and demographic traits, e.g., [114]. Yet, these works evaluate decision quality by centering accuracy, while neglecting the risks of biases that may be contained in human-generated labels and potentially shared among the majority of labelers, and which our work aims to assess.

As part of the approach proposed in this paper, we apply algorithmic fairness methodologies developed in the recent years. Bias mitigation strategies broadly fall under three lines of work: casual fairness [13], individual fairness [17], and group fairness [108]. Our proposed method leverages the fact that ML models are prone to replicating bias contained in training labels, and we thus also integrate algorithmic fairness methodologies to disentangle the bias introduced during the learning process from the bias coming from the human labels themselves. We do so by implementing a group fairness strategy to mitigate bias with respect to the observed labels via a post-processing approach grounded on [92].

5.3 Problem Formulation

We consider a set of K sources of human labels, such as crowd labelers or domain experts, $L = \{L^1, ..., L^K\}$, whose decisions $Y' = \{Y'^1, ..., Y'^K\}$ are encoded in historical data of their decisions. In addition, we consider settings where a small set of gold standard labels, $GS = \{X_l, Y_l\}_{l=1}^m$ is available for instances that may not overlap with any of the labelers' own decision sets. Y is the gold standard label vector, available for the set GS, and likely unavailable for the labelers' instance sets $S = \{S_{L^k}\}_{k=1}^K$, where S_{L^k} indicates labeler L^k's instance set. Figure 5.1 illustrates our settings, including the labelers' decision sets S (left) and a non-overlapping GS data (right).

For a given labeler, L^k, the labeler's assessment for each instance i, $Y'^k_i \in \{0, 1\}$ and its feature vector $X^k_i \sim \mathbb{P}(\mathcal{X})$, are available, such that each

L_i	X1	...	...	Xm	Y'	Y
L1	...	...	...	...	1	?
L1	...	...	...	...	1	?
...	...	...	...	...	0	?
L2	...	...	...	...	0	?
L2	...	...	...	...	1	?
...	...	...	...	...	1	?
...	...	...	...	...	0	?
...	...	...	...	...	1	?
...	...	...	...	...	1	?
Lk	...	...	...	...	1	?
Lk	...	...	...	...	1	?

Labelers' Decision Sets: $S = \{S_{L^1}, S_{L^2}, \ldots S_{L^K}\}$
where $S_{L^k} = \{X_i^k, Y'^k_i\}_{i=1}^{n_{L^k}}$

X1	...	...	Xm	Y'	Y
...	...	...	...	?	1
...	...	...	...	?	0
...	...	...	...	?	1
...	...	...	...	?	0

An Independent Set of Gold
Standard Data: $GS = \{X_l, Y_l\}_1^m$

Figure 5.1: An illustration of labelers' decisions set S (left) and a non-overlapping set with gold-standard labels, GS (right).

labeler has an associated set of instances $S_{L_k} = \{X_i^k, Y'^k_i\}_{i=1}^{n_{L^k}}$, where n_{L^k} is the number of instances labeled by labeler L^k. The sets of instances assessed by different labelers need not overlap but should be drawn from the same class distribution. We seek to produce relative assessment of labelers' decision biases, defined as the labelers' relative ranking by their respective biases, where bias can be the difference in true positive rates (TPRs) across groups (GAP) defined by a sensitive attribute A [46], for example, $A = a, \sim a$ (in Eq.5.1). We consider this measure throughout this paper, but have also found that our approach also applies effectively for different metrics of biases, such as the difference in false positive rates (FPRs) across groups.

$$GAP^{\,k}_{Y'|Y,\,A} = TPR^{\,k}_{Y'|Y,\,a} - TPR^{\,k}_{Y'|Y,\,\sim a} \tag{5.1}$$

We explore how to leverage scarce and costly gold standard data to assess biases in labelers' decisions. For example, labelers may correspond to a group of crowdworkers or other non-experts, tasked with identifying misinformation in online news stories, which has been proposed as a scalable solution to mitigate misinformation [4]. In controlled experiments, labelers biases in this context have been assessed by collecting labels from professional fact-checkers and from crowdworkers for an overlapping pool of cases [4]. Given the experts limited accessibility, this is both costly and not scalable. In this setting, our work could enable the assessments of relative bias of individual labelers or different sources of labels in newly collected crowdsourced labels (so as to improve learning misinformation detection models from the data) using a previously existing pool of professional assessments.

5.4 Methods

This section first introduces the proposed methodology, then briefly provides theoretical reasoning and guarantees, and finalizes with a detailed description of parameter tuning.

5.4.1 Machine-learning-based labelers' Bias Assessment (mba)

The proposed **M**achine-learning-based labelers' **B**ias **A**ssessment (MBA) method leverages a typically problematic property of ML models, which are prone to reproducing biases contained in training labels. The proposed approach first trains models to predict each labeler' assessments, yielding a set of models $\{B^k\}_{k=1}^K$, where each model is a mapping $B^k : X^k \mapsto Y'^k$, induced from labeler L^k's data set, S_{L^k}. We ultimate aim to use the models $\{B^k\}_{k=1}^K$ to infer labelers' relative biases. However, biases contained in the models will have multiple sources; in particular, some biases may be introduced during model training and not correspond to (and thereby might compound) the labeler's biases. Thus, the second stage of the proposed algorithm applies a bias mitigation strategy to counter bias introduced during the learning phase, which assesses disparate deviations of a model's prediction $\hat{Y}$ with respect to the label it is trained to predict, Y'. We do so by proposing a recall-versus-precision ratio (RPR) constraint via post-processing, where we consider the group-specific recall and precision, equivalent to true positive rate and positive predictive value of models' predictions $\hat{Y}$ with respect to labelers' decisions Y' given in a protected group, namely $TPR_{\hat{Y}|Y',A}$ and $PPV_{\hat{Y}|Y',A}$. We then apply the set of post-processed models $\{B^k\}_{k=1}^K$ to make predictions over the set GS. Finally, we estimate the relative assessment of the labelers' decision bias, defined in Equation 5.1 by assessing biases of $\{B^k\}_{k=1}^K$ with respect to Y, i.e., $GAP_{\hat{Y}|Y,A}$.

Figure 5.2: Method Key Steps

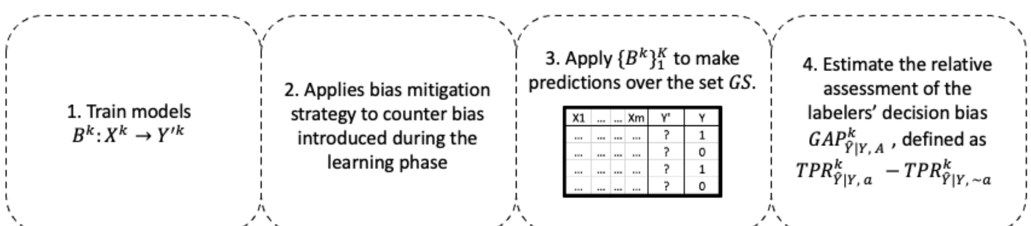


Figure 5.2 shows the four key steps in our approach, and the complete procedure is detailed in Algorithm 1 MBA.

5.4.2 Theoretical Analysis

We now show that, given the correct functional form specification of the labelers' models, i.e., functional form of the relationship between the dependent variable and each independent variable, $f : X \mapsto Y'$, our method can recover the correct relative bias assessments of human labelers.

Theorem 5.4.1. *Given the correct functional form for the labelers models* $(f : X \rightarrow Y')$*, then there exits a ratio* $\frac{TPR^{l}_{\hat{Y}|Y',\,A}}{PPV^{l}_{\hat{Y}|Y',\,A}} = \frac{TPR^{k}_{\hat{Y}|Y',\,A}}{PPV^{k}_{\hat{Y}|Y',\,A}} = c$*, such that if the biases exhibited in labelers l and k' models are folllowing* $GAP^{\,l}_{\hat{Y}|Y,\,A} > GAP^{\,k}_{\hat{Y}|Y,\,A}$*, then the decision biases of this pair of labelers are also following* $GAP^{\,l}_{Y'|Y,\,A} > GAP^{\,k}_{Y'|Y,\,A}$*, where* $GAP^{\,i}_{\hat{Y}|Y,\,A} = TPR^{\,i}_{\hat{Y}|Y,\,a} - TPR^{\,i}_{\hat{Y}|Y,\,\sim a}$ *and* $GAP^{\,i}_{Y'|Y,\,A} = TPR^{\,i}_{Y'|Y,\,a} - TPR^{\,i}_{Y'|Y,\,\sim a}$.

Due to the limited space, we cannot show the entire proof in this paper.

5.4.3 Parameter Selection

In this section, we discuss how we derive the ratio c in Theorem 1 to allow recovery of labelers' biases. Note that the ratio c can be simplified as follows:

$$c = \frac{TPR_{\hat{Y}|Y',\,A}}{PPV_{\hat{Y}|Y',\,A}} = \frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}} = \frac{TP+FP}{TP+FN} = \frac{|\hat{Y}=1,A=a,\sim a|}{|Y'=1,A=a,\sim a|} \tag{5.2}$$

Eq.5.2 reveals the relationship between a labeler model's positive predictions, $\hat{Y} = 1, A = a, \sim a$, and the actual labeler's positive decisions, $Y' = 1, A = a, \sim a$ for a given group $A = a$ or $A =\sim a$. This relationship implies a corresponding desired probability threshold for classification of instances from each protected group.

There are multiple possible values of c that can satisfy the ratio in Eq.5.2, each corresponding to a different probability threshold. We use cross validation (cv) to identify a value c. Once c is determined, we adjust the probability threshold of each model to achieve the ratio c. Note that prior to enforcing the desired threshold on all the labelers' models, each model has an initial threshold for each protected group variable value, given by $\{\pi'_{A=a}, \pi'_{A=\sim a}\} = \{0.5, 0.5\}$. The ultimate threshold pairs, given by $\pi^k_{A=a}$ and $\pi^k_{A=\sim a}$ for labeler L^k, is the averaged across all cv iterations. The procedure for tuning parameter c and identifying the ultimate threshold pair are detailed in Algorithm 2: Find Optimal C.

Once a threshold is identified, each labeler model B^k and the corresponding thresholds pair $\pi^k_{A=a}$ and $\pi^k_{A=\sim a}$, are applied to classify the gold standard instances in GS, based on which the model's prediction biases are computed (8-10 lines in Algorithm 1: MBA), and subsequently ranked.

5.5 Empirical Evaluations

To evaluate our method, we conducted empirical evaluations using simulation studies based on four publicly available datasets: Adult, also known as "Census Income" dataset, Credit dataset from UCI, predicting the default payments of credit card clients [1], Employees Evaluation for Promotion (Employee) dataset from Kaggle[2], and Hospital Readmission Rates dataset from Kaggle [3]. The simulation studies offer controlled settings to allow us to compare the proposed approach with the alternative benchmark, SR, under a variety of settings, including different magnitudes of labelers' decision biases; different class distributions; and different *types* of biases, such as when labelers exhibit correct within-group orderings but have different decision thresholds conditioned on groups, and incorrect within-group orderings driven by the misuse of an interaction variable.

Gold standard labels. We begin by considering a setting where the prevalence of the positive labels is constant across sensitive groups, which yields a scenario where the baseline, SR, may appear to be a sensible choice, given unbiased labels should yield no difference in selection rates across groups. In order to evaluate our method's performance under different class distributions, we consider two scenarios: a positive label prevalence of 20% and 30%, respectively. Note that these two distributions will correspond to settings in which the positive class is smaller, which often arise in practice, e.g., a smaller proportion of candidates would be selected from a large pool of applications. We then select a pool of 400 instances with synthetic gold-standard labels from

[1]https://archive-beta.ics.uci.edu/dataset/350/default+of+credit+card+clients

[2]https://www.kaggle.com/muhammadimran112233/employees-evaluation-for-promotion

[3]https://www.kaggle.com/code/iabhishekofficial/prediction-on-hospital-readmission

each protected group, randomly sampled, as the disjoint set of gold standard data.

Decision Simulation. We run experiments under two types of decision simulations corresponding to two scenarios of interest: "correct within-group ordering" and "incorrect within-group ordering". For the Adult dataset, for example, the "correct within-group ordering" setting means that a labeler infers that women are less likely than others to earn a high income, and thus applies a different threshold for this group, yielding a predefined $TPR_{Y'|Y,\, A=women}$, i.e., true positive rate of labelers' decisions with respect to the gold standard labels within the women group. We assume that labelers correctly assess men, except for random noise that yields an average $TPR_{Y'|Y,\, A=men} = 0.95$. In the "incorrect within-group ordering" setting, we consider labelers' misuse of an interaction term resulting in biased decisions. Specifically, the interaction $sex \times age$ reflects how a labeler relates *age* with *sex*; negative deviations from the true coefficient correspond to a higher degree of bias, e.g., assuming that older women are more likely to earn less, for instance.

It is important to note that even though our theoretical analysis provides guarantees when the functional form specification of the labelers' models is correct, our empirical assessment does not make this assumption. The results show that without knowing the correct functional form, i.e., using a different functional form to simulate labelers decisions and for the labelers' models, our approach remains effective under these settings.

Benchmark. We evaluated our proposed approach relative to the "Selection Rate" (SR) benchmark, which is perhaps the most intuitive and widely considered measure [155] when gold standard labels are unavailable. Specifically, SR estimates a labeler's bias by the difference between the proportion of positive labels the labeler assigns to instances from different groups. For example, the difference of promotion rates among male and female employees.

$$\widehat{GAP}_{sr}^{k} = \sum_{i=1}^{|S_{L^k}|} I[Y'^k = 1|A = a] - \sum_{i=1}^{|S_{L^k}|} I[Y'^k = 1|A =\sim a] \tag{5.3}$$

5.6 Results

In this section, we assess the performance of the proposed approach and compare it with that of the benchmark, SR, under the different settings described in Section 5.5.

Table 5.1 and 5.3 show Spearman's rank-order correlation and their statistical significance of the proposed method, MBA, and of the benchmark SR, for settings where labelers exhibit either correct or incorrect within-group orderings, respectively. Table 5.2 and 5.4 show Pearson correlation coefficients and their statistical significance of the proposed method for the same settings. In each settings and data set, we show results for different class distributions.

Table 5.3 and 5.4 show the two methods' performances when labelers exhibit correct within-group orderings, a scenario in which the baseline, SR, is optimal. The results indicate that MBA performs comparably well in this setting. When labelers conditionally misestimate the interaction of the sen-

Table 5.1: Spearman's rank-order ρ for MBA (ours) and the benchmark SR when labelers exhibit correct within-group ordering, ideal setting for SR. The ranks produced by **MBA** and **SR** both show significant correlation with true rank.

Dataset	**Setting**	**MBA (ours)**	**SR**
Adult	$P(Y=1\|A)=20\%$	0.947***	0.932**
Credit	$P(Y=1\|A)=20\%$	0.772*	0.936***
Employee	$P(Y=1\|A)=20\%$	0.895***	0.956***
Readmission	$P(Y=1\|A)=20\%$	0.934***	0.979***
Adult	$P(Y=1\|A)=30\%$	0.970***	0.966***
Credit	$P(Y=1\|A)=30\%$	0.860**	0.973***
Employee	$P(Y=1\|A)=30\%$	0.918***	0.983***
Readmission	$P(Y=1\|A)=30\%$	0.979***	0.989***

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: $p < 0.01$, and ***: $p < 0.001$.

sitive attribute with a feature (e.g., *sex* $\times$ *age* for the Adult dataset), while appearing to have the same selection rates, Table 5.3 and 5.4 show that the SR benchmark exhibits significantly poor performance and thus cannot be relied on in practice. By contrast, MBA produces an accurate rank of labelers' bias ($GAP_{\hat{Y}|Y,A}$) that is significantly correlated to the true rank ($GAP_{Y'|Y,A}$).

Figures 5.3, 5.4, 5.5, 5.6, show predicted bias, $GAP_{\hat{Y}|Y,A}$, produced by MBA and the SR benchmark, as well as labelers' true bias, $GAP_{Y'|Y,A}$, with 90% confidence bars. Recall that our goal is to recover the correct ranking of labelers' biases; hence, in these plots, we examine whether (and the degree to which) a labeler's bias was correctly positioned relative to others, as shown for the true biases. Figures 5.3, 5.4 show the ranking produced by the two methods

Table 5.2: Pearson correlation coefficients r for MBA (ours) and the benchmark SR when labelers exhibit correct within-group ordering, ideal setting for SR. The ranks produced by **MBA** and **SR** both show significant correlation with true rank.

Dataset	**Setting**	**MBA (ours)**	**SR**
Adult	$P(Y=1\|A)=20\%$	0.942***	0.943***
Credit	$P(Y=1\|A)=20\%$	0.775*	0.922***
Employee	$P(Y=1\|A)=20\%$	0.901***	0.961***
Readmission	$P(Y=1\|A)=20\%$	0.928***	0.981***
Adult	$P(Y=1\|A)=30\%$	0.964***	0.973***
Credit	$P(Y=1\|A)=30\%$	0.856**	0.976***
Employee	$P(Y=1\|A)=30\%$	0.931***	0.982***
Readmission	$P(Y=1\|A)=30\%$	0.975***	0.992***

*: p-value $<$ 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p $<$ 0.01, and ***: p $<$ 0.001.

Table 5.3: Spearman's rank-order ρ for MBA (ours) and benchmark SR when labelers exhibit incorrect within-group ordering. The ranks produced by MBA (ours) shows significant correlation with true rank, while the benchmark SR yielded all labelers having the same bias.

Dataset	**Setting**	**MBA (ours)**	**SR**
Adult	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.928***	-0.128
Credit	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.905**	0.079
Employee	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.841*	0.263
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=20\%$	0.975***	-0.337
Adult	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.942***	-0.058
Credit	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.942***	0.038
Employee	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.918***	0.477
Readmission	$P(Y=1\|A)=P(Y'=1\|A)=30\%$	0.977***	0.287

*: p-value $<$ 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: p $<$ 0.01, and ***: p $<$ 0.001.

Table 5.4: Pearson correlation coefficients r for MBA (ours) and benchmark SR when labelers exhibit incorrect within-group ordering. The ranks produced by MBA (ours) shows significant correlation with true rank, while the benchmark SR yielded all labelers having the same bias.

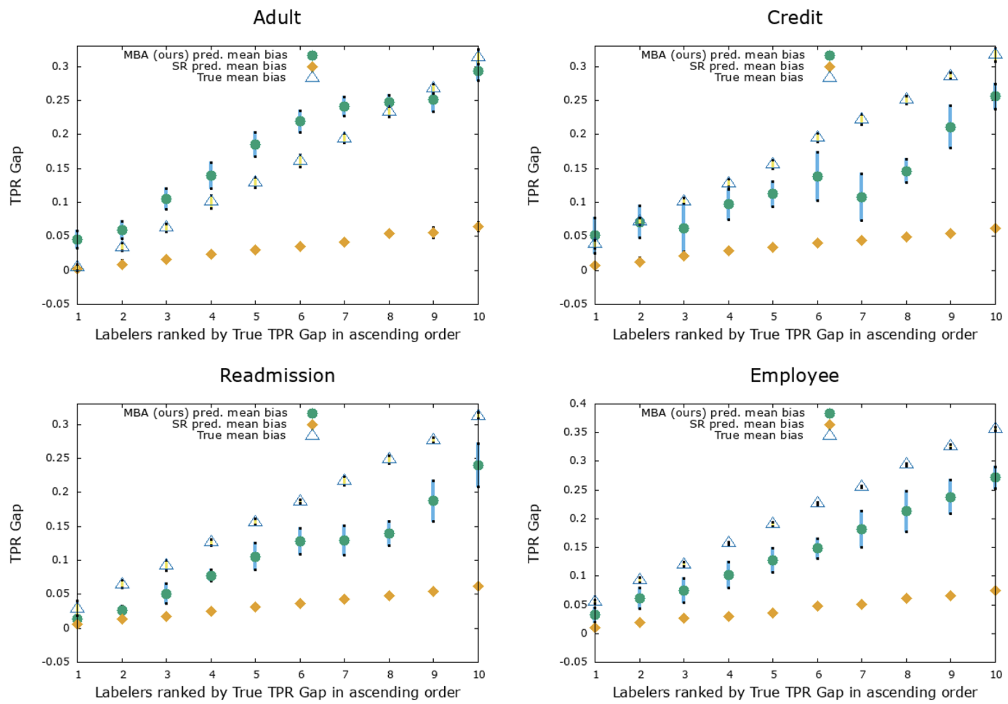
Dataset	**Setting**	**MBA (ours)**	**SR**
Adult	$P(Y=1\|A) = P(Y'=1\|A) = 20\%$	0.927***	-0.084
Credit	$P(Y=1\|A) = P(Y'=1\|A) = 20\%$	0.922***	0.080
Employee	$P(Y=1\|A) = P(Y'=1\|A) = 20\%$	0.881**	0.322
Readmission	$P(Y=1\|A) = P(Y'=1\|A) = 20\%$	0.964***	-0.329
Adult	$P(Y=1\|A) = P(Y'=1\|A) = 30\%$	0.934***	-0.031
Credit	$P(Y=1\|A) = P(Y'=1\|A) = 30\%$	0.938***	0.047
Employee	$P(Y=1\|A) = P(Y'=1\|A) = 30\%$	0.922***	0.637
Readmission	$P(Y=1\|A) = P(Y'=1\|A) = 30\%$	0.972***	0.543

*: p-value < 0.05, indicating that the correlation coefficient is different from zero and that a linear relationship exists, **: $p < 0.01$, and ***: $p < 0.001$.

for settings where labelers exhibit correct within-group ordering. Interestingly, even though both methods show high correlation with labelers' true rank in Table 5.1 and 5.2, the figures reveal how MBA approximates well both the relative ranking as well as the magnitude of the biases in all settings.

Figures 5.5 and 5.6 evaluate settings when labelers exhibit incorrect within-group orderings. This assessment visualizes the failure of the benchmark in this setting, which incorrectly yields all labelers as having the same (null) bias. Meanwhile, while MBA tends to underestimate the magnitude of the biases, it effectively recovers the correct rank of labelers' relative biases.

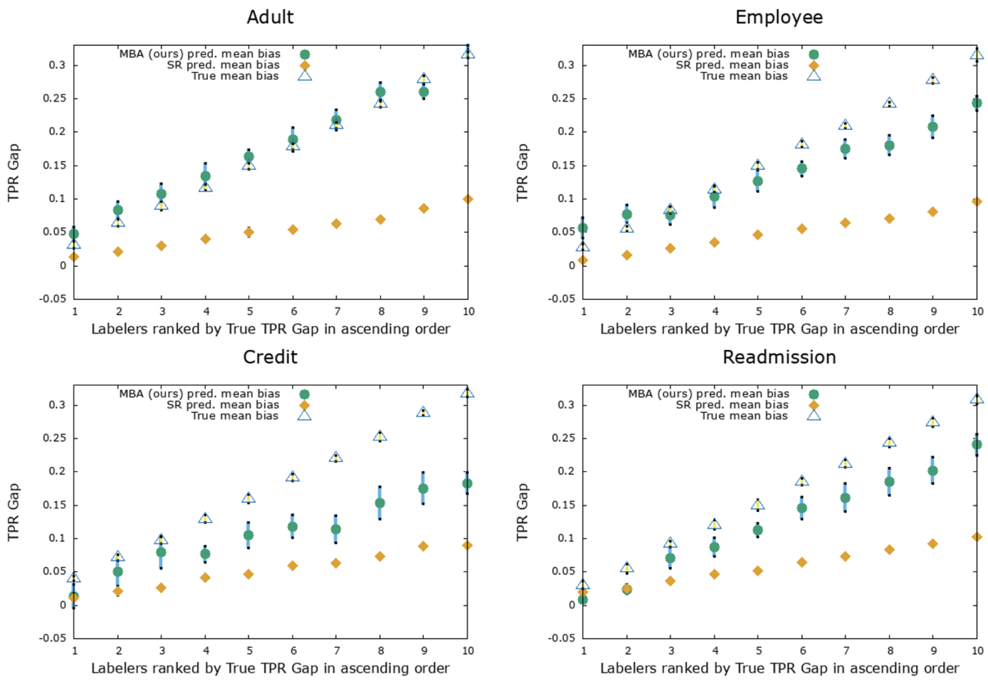
Figure 5.3: Predicted $GAP_{\hat{Y}|Y,A}$ by MBA (ours) and SR, and true $GAP_{Y'|Y,A}$ when labelers exhibit correct within-group ordering, and for 20% positive rate. Both MBA's and SR's ranking have significant correlation with true rank.



5.7 Discussion and Future Work

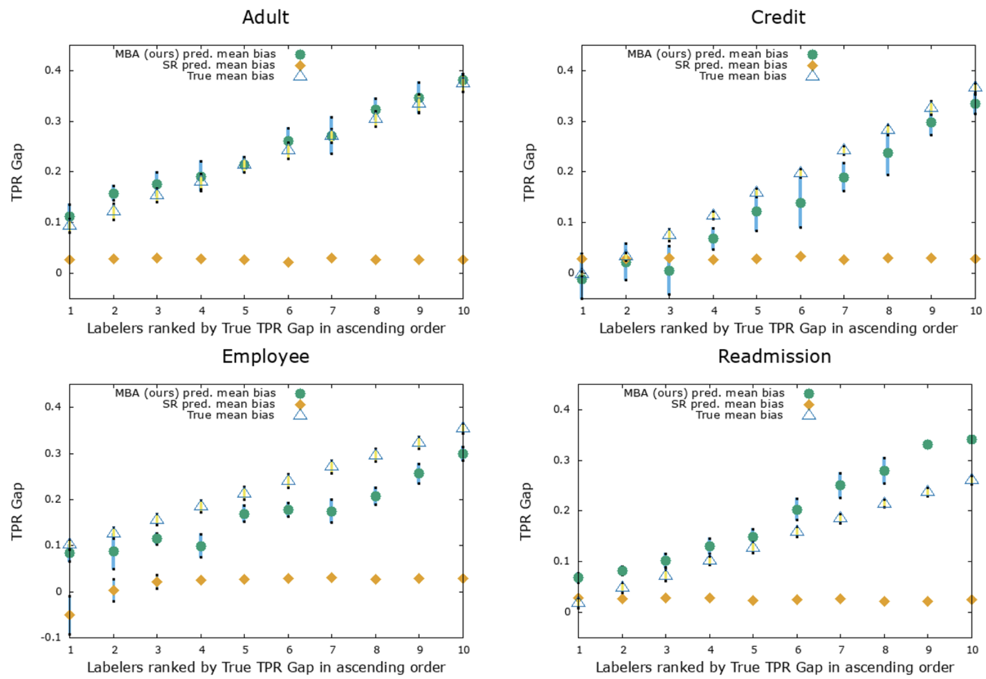
In this paper, we tackle the problem of assessing biases encoded in labelers' decisions. We propose an algorithm that returns an assessment of labelers' relative biases for a set of labelers, without requiring ground truth labels to be available for the instances assessed by the labelers, nor any overlap in the instances assessed across labelers'. The proposed approach estimates biases in terms of gaps in true positive rates, and we illustrate its performance by comparing it to the typically used alternative, selection rates (SR), which has

Figure 5.4: Predicted $GAP_{\hat{Y}|Y,A}$ by MBA (ours) and SR , and the true $GAP_{Y'|Y,A}$ when labelers exhibit correct within-group ordering, and 30% positive rate. MBA estimates follow the true rank better than SR.



the advantage of not requiring any ground-truth, but, as a result, also cannot account for the correctness of labelers' decisions. After providing theoretical guarantees for the proposed approach, we conduct an empirical assessment in which we consider different scenarios, both favorable and unfavorable for the baseline, SR. We show that our method performs well in what constitutes a best-case-scenario for SR, and then study a scenario in which SR can be misleading, revealing the advantages of the proposed approach in providing consistently good performance in both settings. While assessments of decisions and labeling biases based on selection rates are widespread, our results

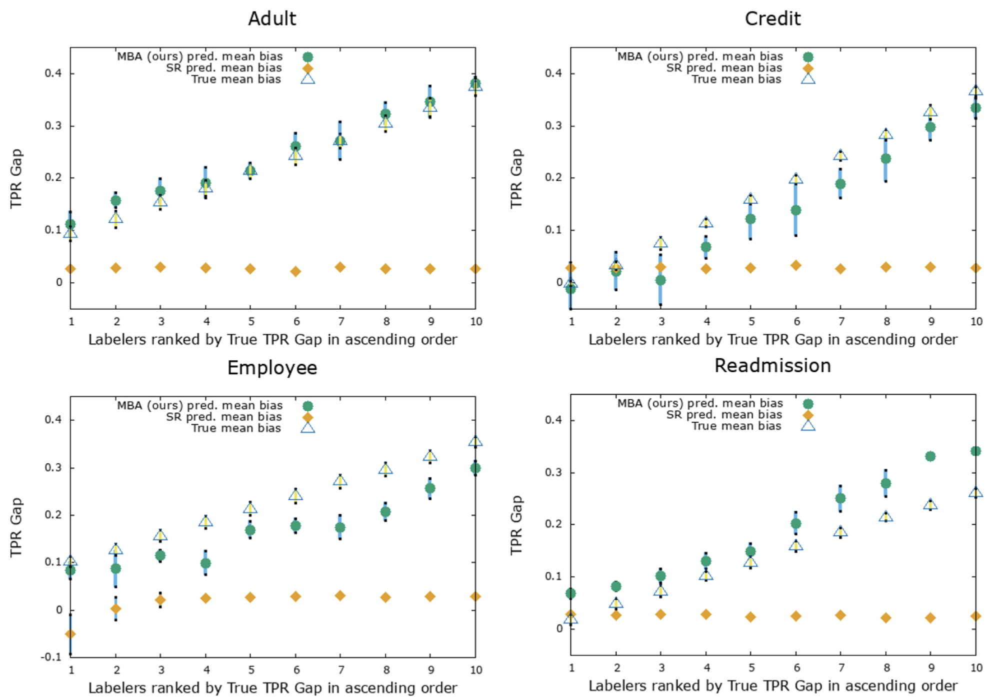
Figure 5.5: Predicted $GAP_{\hat{Y}|Y,A}$ by (ours) and SR, and the true $GAP_{Y'|Y,A}$ when labelers predict incorrect within-group ordering, and 20% positive rate. MBA yields correct ranking of labelers' biases while SR misestimates the biases to be apprixmately the equivalent.



show how SR may fail to differentiate between labelers exhibiting very different degrees of biases and are prone to being gamed by adversaries. The proposed approach addresses this problem and lays the groundwork towards reliable bias assessment in labeling. In future work we plan on conducting empirical studies using human-generated labels on a variety of tasks, to characterize both when the method succeeds and when the method fails in practice.

Increasing transparency in labelers' biases may have a variety of benefits. We are interested in identifying productive ways to bring the relative

Figure 5.6: Predicted $GAP_{\hat{Y}|Y,A}$ by MBA, SR, and the true $GAP_{Y'|Y,A}$ when labelers predict incorrect within-group ordering, and 30% positive rate. MBA infers the correct ranking of lablers biases, while SR failes to do so.



bias assessment to bear on related research questions and downstream tasks, including utilizing the output of our method when training an algorithm on human-generated labels. We are also interested in human-centered interventions that provide this piece of information to labelers as part of strategies meant to counter cognitive biases during labeling or decision-making. Finally, we intend to deepen our study of adversarial settings and modes of failure to better understand how and when different quantitative measures of quality and bias may be misleading and gameable, in order to better characterize its limitations and caution against its misuse as mechanisms for automated assessments.

Appendices

Appendix A

Algorithm Blocks

A.1 Algorithm: MDE

Algorithm 1: MDE

Algorithm MDE:
  Input: $\{S_{\mathrm{W}_k}\}_1^K$, GT
  // Creating C sets of N synthetic workers:
  for $c = 1 \ldots C$ **do** /* used $C = 10$, $N = 101$ */
    for $n = 1 \ldots N$ **do**
      $q_n^c = (1 - (n-1) * intv)$ /* used $intv = 0.005$, $\{q_n^c\}_1^N$ is in range $[0.5, 1]$ */
      $S_{sw_n}^c$ ←GT
      $S_{sw_n}^c$ ← Randomly draw a proportion of $(1 - q_n^c)$ instances from $\mathcal{S}_{sw_n}^c$ and invert their Y labels
  /* $\{\{S_{sw_n}^c\}_1^N\}_1^C$ synthetic workers created with accuracies $\{\{q_n^c\}_1^N\}_1^C$ */
  // Training base models on copies of real workers' data $\{S_{\mathrm{W}_k}\}_1^K$:
  $\{S_{\mathrm{W}_k}^{copy}\}_1^K \leftarrow \{S_{\mathrm{W}_k}\}_1^K$
  foreach $S_{\mathrm{W}_k^{copy}} \in \{S_{\mathrm{W}_k}^{copy}\}_1^K$ **do**
    foreach $\{X_i^k, \hat{Y}_i^k\} \in \mathcal{S}_{\mathrm{W}_k}^{copy}$ **do**: **if** $\{X_i^k, Y_i^k\} \in GT$ **then** replace $\hat{Y}_i^k$ with Y_i^k
    Train base model $B(S_{\mathrm{W}_k}^{copy})$ on $S_{\mathrm{W}_k}^{copy}$
  /* K base models $\{B_j\}_1^K$ created */
  // Using C sets of synthetic workers to produce C different mappings:
  for $c = 1 \ldots C$ **do** /* Each mapping c produced from N synthetic workers */
    $\{DQ_{sw_n}^c\}_1^N \leftarrow$ **for** $n = 1 \ldots N$ **do**: Produce DQ Scores($S_{sw_n}^c$, $\{B_j\}_1^K$)
    Induce a mapping: $f_c : DQ \rightarrow q$ from the set $\{DQ_{sw_n}^c, q_n^c\}_{n=1}^N$
  /* C mapping functions $\{f_c\}_1^C$ created */
  // Producing DQ scores and assessments for real workers $W = \{W_1, \ldots, W_k\}$:
  $\{DQ_{\mathrm{W}_k}\}_1^K \leftarrow$ **for** $k = 1 \ldots K$ **do**: Produce DQ Score (S_{Wk}, $\{B_j\}_1^K$)
  $\{\hat{q}_{W_k MDE}\}_1^K \leftarrow$ **for** $k = 1 \ldots K$ **do**: Produce Assessment(S_{W_k}, $\{f_c\}_1^C$, DQ_{W_k})
  return $\{\hat{q}_{W_k MDE}\}_1^K$, $\{B_j\}_1^K$, $\{f_c\}_1^C$
  /* $\{\hat{q}_{W_k MDE}\}_1^K$ are Alg.1's assessments of workers' accuracies. Alg.1 is applied in Alg.2, so $\{B_j\}_1^K$, $\{f_c\}_1^C$ are returned for reuse in Alg.2. */

Procedure Produce DQ Score(S_k, $\{B_j\}_1^K$):
  // Evaluating S_k's DQ (real or synthetic decisions S_{w_k} or $S_{\mathtt{sw}_k}$):
  $s_k^+ = \{\}$, $s_k^- = \{\}$
  foreach $\{X_i^k, Y_i^k\} \in S_k$ **do**
    $M(X_i^k) = \arg\max_z (\sum_{j=1, X_i^k \notin S_j}^K B_j(X_i^k)_z)$
    $Conf_{X_i^k} = \sum_{j=1, X_i^k \notin S_j}^K B_j(X_i^k)_{M(X_i^k)}$ /* $B_j(X_i^k)_{M(X_i^k)}$ denotes B_j's probability estimate that X_i^k maps to the class inferred by the ensemble M */
    if $\hat{Y}_i^k == M(X_i^k)$ **then** $s_j^+ = s_j^+ \bigcup \{X_i^k, \hat{Y}_i^k\}$
    else $s_j^- = s_j^- \bigcup \{X_i^k, \hat{Y}_i^k\}$
  return $DQ_k = \frac{\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^+} Conf_{X_i^k}}{(\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^+} Conf_{X_i^k}) + (\sum_{\{X_i^k, \hat{Y}_i^k\} \in S_k^-} Conf_{X_i^k})}$

Procedure Produce Assessment($\{f_c\}_1^C$, DQ_k):
  for $c = 1 \ldots C$ **do** /* Apply C mappings to predict a worker's accuracy */
    $\hat{q}_c = f_c(DQ_k)$) and truncate $\hat{q}_c$ into range $[0.5, 1]$ if necessary
  $\hat{q} = average(\{\hat{q}_c\}_1^C)$ /* Final estimate is the average of C assessments */
  return $\hat{q}$

A.2 Algorithm: MDE-HYB

Algorithm 2: MDE-HYB

Algorithm MDE-HYB:

Input: $\{S_{W_j}\}_1^K$, $GT = \{GT_k\}_1^K$

// Infer workers' accuracies, create base models, and induce mapping functions using MDE:

$\{\hat{q}_k^{\text{MDE}}\}_1^K$, $\{B_j\}_1^K$, $\{f_c\}_1^C \leftarrow$ Algorithm 1: MDE($\{S_{W_j}\}_1^K$, GT)

// Infer workers' accuracies using EAR (based on EQ 2):

$\{\hat{q}_k^{\text{EAR}}\}_1^K \leftarrow$ **for** $k = 1...K$ **do** : $\hat{q}_k^{\text{EAR}} = (\sum_{i=1}^{|GT_k|} I[Y_i == \hat{Y}_i])/|GT_k|$

// Generate distributions of MDE and EAR's errors

err_{MDE}, $err_{\text{EAR}} \leftarrow$ Generate Error Dist($\{\hat{q}_k^{\text{EAR}}\}_1^K$, $\{\hat{q}_k^{\text{MDE}}\}_1^K$, GT, $\{f_c\}_1^C$, $\{B_j\}_1^K$))

$p_{EAR} \leftarrow$ Compute p value: H_0^1:$(\mu_{\text{MDE}} - \mu_{\text{EAR}}) \leq d$ H_a^1:$(\mu_{\text{MDE}} - \mu_{\text{EAR}}) > d$ from $err_{\text{EAR}}, err_{\text{MDE}}$

$p_{MDE} \leftarrow$ Compute p value: H_0^2:$(\mu_{\text{EAR}} - \mu_{\text{MDE}}) \leq d$ H_a^2:$(\mu_{\text{EAR}} - \mu_{\text{MDE}}) > d$ from $err_{\text{EAR}}, err_{\text{MDE}}$

// μ_{EAR} and μ_{MDE} are the distribution means of err_{EAR} and err_{MDE} respectively

for $k = 1 \dots K$ **do** /* Estimate quality for each worker W_j */

  if $p_{MDE} \geq \alpha$ *and* $p_{EAR} \geq \alpha$ **then** /* Cannot reject either hypothesis */

    $\hat{q}_k = (\frac{p_{MDE}}{p_{MDE}+p_{EAR}} * \hat{q}_k^{\text{MDE}}) + (\frac{p_{EAR}}{p_{MDE}+p_{EAR}} * \hat{q}_k^{\text{EAR}})$

  else

    if $p_{MDE} < \alpha$ **then** $\hat{q}_k = \hat{q}_k^{\text{MDE}}$ **else** $\hat{q}_k = \hat{q}_k^{\text{EAR}}$

return $\{\hat{q}_k\}_1^K$

Procedure Generate Error Dist($\{\hat{q}_k^{\text{EAR}}\}_1^K$, $\{\hat{q}_k^{\text{MDE}}\}_1^K$, GT, $\{f_c\}_1^C$, $\{B_j\}_1^K$)):

$\{\hat{q}_{k_{\text{AVG}}}\}_1^K = \frac{1}{2}\{\hat{q}_k^{\text{EAR}} + \hat{q}_k^{\text{MDE}}\}_1^K$

$[\hat{q}_{lower}, \hat{q}_{upper}] \leftarrow \hat{\mu} \pm t_{\alpha=0.01,K-1} \frac{\hat{\theta}}{\sqrt{n}}$ /* $\hat{\mu}$ is mean and $\hat{\theta}$ is std of $\{\hat{q}_{k_{\text{AVG}}}\}_1^K$ */

$err_{\text{EAR}} = \{\}$; $err_{\text{MDE}} = \{\}$

for $r = 1 \dots R$ **do** /* Draw R different subsets of GT */

  $S_{sw_r} \leftarrow$ Randomly draw t instances from GT

  for $p = 1 \dots P$ **do** /* Repeat for P synthetic workers */

    $q \leftarrow$ Uniformly draw from $[\hat{q}_{lower}, \hat{q}_{upper}]$

    // Create synthetic workers' data, apply MDE, and estimate its errors:

    $S_{sw_{\text{MDE}}} \leftarrow S_{sw_r}$

    Flip each label in $S_{sw_{\text{MDE}}}$ with probability q

    $DQ \leftarrow$Produce DQ Score $(S_{sw_{\text{MDE}}}, \{B_j\}_1^K)$ /* Alg.1 Procedure */

    $\hat{q}_{\text{MDE}}$ =Produce Assessment $(S_{sw_{\text{MDE}}}, \{f_c\}_1^C, DQ)$ /* Alg.1 Procedure */

    $err_{\text{MDE}} \leftarrow err_{\text{MDE}} \bigcup \{|q - \hat{q}_{\text{MDE}}|\}$

    // Simulate decision errors, apply EAR, and estimate its errors:

    $GT_{PE} = \frac{|GT|}{|W|}$ /* average number of ground truth per expert */

    $S_{sw_{EAR}} \leftarrow$ Randomly draw GT_{PE} instances from S_{sw_r} and flip each of their labels with probability q

    $\hat{q}_{\text{EAR}} \leftarrow$ Apply EAR to $S_{sw_{EAR}}$

    $err_{\text{EAR}} \leftarrow err_{\text{EAR}} \bigcup \{|q - \hat{q}_{\text{EAR}}|\}$

return err_{MDE}, err_{EAR}

A.3 Algorithm: MBA

Algorithm 3: MBA

Algorithm MBA($\{S_{L^k}\}_{k=1}^{K}$, GS):

  foreach $S_{L^k} \in \{S_{L^k}\}_{k=1}^{K}$ **do** train base model $B(S_{L^k})$ on S_{L^k}

  // Step 1

  foreach $B^k \in \{B^k\}_{k=1}^{K}$ **do**

    $\{\hat{Y}^k\} \leftarrow$ use B^k to classify $\forall\, X_i^k \in S_{L^k}$

    Calculate $\{c'^k_{A=a}, c'^k_{A=\sim a}\}$ based on Eq.5.2

  $c_{opt.} \leftarrow$ Algorithm 2: Find Optimal C($\{c'^k_{A=a}, c'^k_{A=\sim a}\}_{k=1}^{K}$)

  $\{\pi^k_{A=a}, \pi^k_{A=\sim a}\}_{k=1}^{K} \leftarrow$ compute thresholds of $\{B^k\}_{k=1}^{K}$ with $c_{opt.}$

  // Step 2 ends

  foreach $B^k \in \{B^k\}_{k=1}^{K}$ **do**

    $\{\hat{Y}\}_{l=1}^{m} \leftarrow$ use B^k with $[\pi^k_{A=a}, \pi^k_{A=\sim a}]$ classify $GS = \{X_l, Y_l\}_{l=1}^{m}$

    $GAP^{\,k}_{\hat{Y}|Y,\,A} = TPR^{\,k}_{\hat{Y}|Y,\,a} - TPR^{\,k}_{\hat{Y}|Y,\,\sim a}$

  return $\{GAP_{\,\hat{Y}|Y,\,A}\}_{k=1}^{K}$ // Step 3 and 4 end

A.4 Algorithm: Find Optimal C

Algorithm 4: Find Optimal C

Algorithm `Find Optimal C`($\{c'_{j,\,A=a}, c'_{j,\,A=\sim a}\}_{j=1}^{K}$):

 $[c_{min}, c_{max}] \leftarrow$ minimum and maximum of $\{c'_{j,\,A=a}, c'_{j,\,A=\sim a}\}_{j=1}^{K}$

 $c_{step} \leftarrow c_{min}$

 do

  foreach $S_{L_j} = \{X_i^j, Y_i'^j\}_{i=1}^{n_j} \in \{S_{L_j}\}_{j=1}^{K}$ **do** `/* T-fold cross-validation */`

   Generate T stratified by $A = a, \sim a$ splits: $\{X^j, Y'^j\}_{t=1}^{T}$

   Train a model on the train splits and find corresponding thresholds based on c_{step} and the test split

   $\pi^p_{j,\,A=a}, \pi^p_{j,\,A=\sim a} \leftarrow average(\{\pi^t_{j,\,A=a}, \pi^t_{j,\,A=\sim a}\}_{t=1}^{T})$

  $c_{step} \leftarrow c_{step}$+ step_p

 while $c_{step} \leq c_{max}$

 $steps = \frac{c_{max} - c_{min}}{step\_p}$

 $c_{opt.} \leftarrow c_{step}$ which yields the $\min(\{std(\{TPR_{j,A=\sim a}\}_{j=1}^{K})_p\}_{p=1}^{steps})$

 return $c_{opt.}$

Appendix B

Theorem Proof

Lemma B.0.1. *If the correct functional form specification of each labeler' model B, a mapping $f : X \mapsto Y'$ is known, then $\hat{Y} \perp\!\!\!\perp Y|Y'$ and also $Y' \perp\!\!\!\perp Y|\hat{Y}$.*

Proof. Given the correct functional form for the labelers models ($f : X \to Y'$), then there exits a ratio $\frac{TPR^{l}_{\hat{Y}|Y',\, A}}{PPV^{l}_{\hat{Y}|Y',\, A}} = \frac{TPR^{k}_{\hat{Y}|Y',\, A}}{PPV^{k}_{\hat{Y}|Y',\, A}} = c$, such that if the biases exhibited in labelers l and k' models are folllowing $GAP^{\,l}_{\hat{Y}|Y,\, A} > GAP^{\,k}_{\hat{Y}|Y,\, A}$, then the decision biases of this pair of labelers are also following $GAP^{\,l}_{Y'|Y,\, A} > GAP^{\,k}_{Y'|Y,\, A}$, where $GAP^{\,i}_{\hat{Y}|Y,\, A} = TPR^{\,i}_{\hat{Y}|Y,\, a} - TPR^{\,i}_{\hat{Y}|Y,\, \sim a}$ and $GAP^{\,i}_{Y'|Y,\, A} = TPR^{\,i}_{Y'|Y,\, a} - TPR^{\,i}_{Y'|Y,\, \sim a}$. □

Proof. Given $GAP^{\,l}_{\hat{Y}|Y,\,A} > GAP^{\,k}_{\hat{Y}|Y,\,A}$, this can be rewritten as:

$$\begin{aligned} P(\hat{Y}_l = 1|A = 0, Y = 1) - P(\hat{Y}_l = 1|A = 1, Y = 1) > \\ P(\hat{Y}_k = 1|A = 0, Y = 1) - P(\hat{Y}_k = 1|A = 1, Y = 1) \end{aligned} \tag{B.1}$$

then

$$\begin{aligned} P(Y'_l = 1|A = 0, Y = 1) - P(Y'_l = 1|A = 1, Y = 1) > \\ P(Y'_k = 1|A = 0, Y = 1) - P(Y'_k = 1|A = 1, Y = 1) \end{aligned} \tag{B.2}$$

It is also true that,

$$\begin{aligned} P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y'_i = 1|A = a, Y = 1, \hat{Y}_i = 1) = \\ \tfrac{P(\hat{Y}_i=1, A=a, Y=1)}{P(A=a, Y=1)} * \tfrac{P(Y'_i=1, A=a, Y=1, \hat{Y}_i=1)}{P(A=a, Y=1, \hat{Y}_i=1)} = \tfrac{P(Y'_i=1, A=a, Y=1, \hat{Y}_i=1)}{P(A=a, Y=1)} = \\ P(Y'_i = 1, \hat{Y}_i = 1|A = a, Y = 1) \end{aligned} \tag{B.3}$$

By rearranging eq.B.3, we have

$$\begin{aligned} P(Y'_i = 1, \hat{Y}_i = 1|A = a, Y = 1) = \\ P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y'_i = 1|A = a, Y = 1, \hat{Y}_i = 1) \end{aligned} \tag{B.4}$$

It is also true that,

$$\frac{P(Y_i'=1,\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(Y_i'=1,\hat{Y}_i=1,A=a,Y=1)}{P(A=a,Y=1)} * \frac{P(A=a,Y=1)}{P(Y_i'=1,A=a,Y=1)} = P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{B.5}$$

By rearranging eq.B.5,

$$P(Y_i' = 1, \hat{Y}_i = 1|A = a, Y = 1) = P(Y_i' = 1|A = a, Y = 1) * P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{B.6}$$

From eq.B.4 and eq.B.6,

$$P(\hat{Y}_i = 1|A = a, Y = 1) * P(Y_i' = 1|A = a, Y = 1, \hat{Y}_i = 1) = P(Y_i' = 1|A = a, Y = 1) * P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{B.7}$$

By rearranging eq.B.7,

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(\hat{Y}_i=1|Y_i'=1,A=a,Y=1)}{P(Y_i'=1|\hat{Y}_i=1,A=a,Y=1)} \tag{B.8}$$

From Lemma B.0.1, it is true that $\hat{Y}Y|Y'$, and $\hat{Y}Y|A, Y'$; therefore,

$$P(\hat{Y}_i = 1|A = a, Y_i' = 1) = P(\hat{Y}_i = 1|Y_i' = 1, A = a, Y = 1) \tag{B.9}$$

and

$$P(Y_i' = 1|A = a, \hat{Y}_i = 1) = P(Y_i' = 1|\hat{Y}_i = 1, A = a, Y = 1) \tag{B.10}$$

From eq.B.8, eq.B.9, and eq.B.10, we have

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{P(Y_i'=1|A=a,Y=1)} = \frac{P(\hat{Y}_i=1|Y_i'=1,A=a)}{P(Y_i'=1|\hat{Y}_i=1,A=a)} \tag{B.11}$$

Note that right hand side of eq.B.11 is the "recall (TPR $_{\hat{Y}|Y',\, A}$) versus precision (PPV $_{\hat{Y}|Y',\, A}$) ratio" and we let the ratio equal to a constant c, so

$$\frac{P(\hat{Y}_i=1|Y_i'=1,A=a)}{P(Y_i'=1|\hat{Y}_i=1,A=a)} = \frac{TPR^{\,i}_{\hat{Y}|Y',\, A}}{PPV^{\,i}_{\hat{Y}|Y',\, A}} = c \tag{B.12}$$

From eq.B.12, it is true that

$$\frac{P(\hat{Y}_i=1|A=a,Y=1)}{c} = P(Y_i' = 1|A = a, Y = 1) \tag{B.13}$$

Given eq.B.1 above:

$$\begin{aligned} P(\hat{Y}_l = 1|A = 0, Y = 1) - P(\hat{Y}_l = 1|A = 1, Y = 1) > \\ P(\hat{Y}_k = 1|A = 0, Y = 1) - P(\hat{Y}_k = 1|A = 1, Y = 1) \end{aligned} \tag{B.1}$$

dividing both sides by c, we have:

$$\frac{P(\hat{Y}_l = 1|A = 0, Y = 1)}{c} - \frac{P(\hat{Y}_l = 1|A = 1, Y = 1)}{c} > \frac{P(\hat{Y}_k = 1|A = 0, Y = 1)}{c} - \frac{P(\hat{Y}_k = 1|A = 1, Y = 1)}{c} \tag{B.14}$$

which is equivalent to

$$P(Y_l' = 1|A = 0, Y = 1) - P(Y_l' = 1|A = 1, Y = 1) > P(Y_k' = 1|A = 0, Y = 1) - P(Y_k' = 1|A = 1, Y = 1) \tag{B.2}$$

□

Bibliography

[1] Balazs Aczel, Bence Bago, Aba Szollosi, Andrei Foldes, and Bence Lukacs. Measuring individual differences in decision biases: methodological considerations. *Frontiers in psychology*, 6:1770, 2015.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[3] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

[4] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi. org/10.31234/osf. io/9qdza*, 2020.

[5] Hossein Amirkhani and Mohammad Rahmati. Agreement/disagreement based crowd labeling. *Applied intelligence*, 41(1):212–222, 2014.

[6] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[7] Shlomo Argamon-Engelson and Ido Dagan. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11:335–360, 1999.

[8] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[9] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, pages 1770–1780. PMLR, 2020.

[10] Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M. Redmiles, Meeyoung Cha, and Krishna P. Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, pages 1–12, 2021.

[11] Kevin Bache and Moshe Lichman. Uci machine learning repository, 2013.

[12] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23, 1999.

[13] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the eth-

ical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76. PMLR, 2018.

[14] Fabian Beigang. Shortcomings of counterfactual fairness and a proposed modification. *arXiv preprint arXiv:2011.07312*, 2020.

[15] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

[16] Mustafa Bilgic and Lise Getoor. Link-based active learning. In *NIPS Workshop on Analyzing Networks and Learning with Graphs*, volume 4, page 9, 2009.

[17] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.

[18] Andriy Bodnaruk and Andrei Simonov. Do financial experts make better investment decisions? *Journal of Financial Intermediation*, 24(4):514–536, 2015.

[19] Sarah E Bonner. A model of the effects of audit task complexity. *Accounting, organizations and society*, 19(3):213–234, 1994.

[20] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7474–7483, 2017.

[21] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[22] Erik Brynjolfsson and Andrew McAfee. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy.* Brynjolfsson and McAfee, 2011.

[23] Ethan R Burris. The risks and rewards of speaking up: Managerial responses to employee voice. *Academy of management journal*, 55(4):851–875, 2012.

[24] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[25] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[26] Gretchen B Chapman and Arthur S Elstein. Cognitive processes and biases in medical decision making. 2000.

[27] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Data programming using continuous and quality-guided labeling functions. *arXiv preprint arXiv:1911.09860*, 2019.

[28] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[29] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[30] Yiling Chen, Ian Kash, Mike Ruberry, and Victor Shnayder. Decision markets with good incentives. In *Internet and Network Economics: 7th International Workshop, WINE 2011, Singapore, December 11-14, 2011. Proceedings 7*, pages 72–83. Springer, 2011.

[31] Zhijun Chen, Huimin Wang, Hailong Sun, Pengpeng Chen, Tao Han, Xudong Liu, and Jie Yang. Structured probabilistic end-to-end learning from crowds. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1512–1518. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[32] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[33] Maria Christoforaki and Panagiotis G. Ipeirotis. A system for scalable and reliable technical-skill testing in online labor markets. *Computer Networks*, 90:110–120, 2015. Crowdsourcing.

[34] Marvin S Cohen. Three paradigms for viewing decision biases. *Decision making in action: Models and methods*, 1:36–50, 1993.

[35] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[36] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. In *Advances in neural information processing systems*, pages 705–712, 1995.

[37] Brian W Collins. Tackling unconscious bias in hiring practices: The plight of the rooney rule. *NYUL Rev.*, 82:870, 2007.

[38] Pat Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780, 2003.

[39] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

[40] Peng Dai, Christopher H Lin, Daniel S Weld, et al. Pomdp-based control of workflows for crowdsourcing. *Artificial Intelligence*, 202:52–85, 2013.

[41] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294, 2013.

[42] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

[43] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications, 2018.

[44] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*, 2021.

[45] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[46] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[47] Kurt De Grave, Jan Ramon, and Luc De Raedt. Active learning for primary drug screening. In *Benelearn 08, The Annual Belgian-Dutch Machine Learning Conference*, volume 2008, pages 55–56, 2008.

[48] Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787):684–687, 2006.

[49] Ofer Dekel and Ohad Shamir. Vox populi: Collecting high-quality labels from a crowd. In *COLT*, 2009.

[50] Meghana Deodhar, Joydeep Ghosh, Maytal Saar-Tsechansky, and Vineet Keshari. Active learning with multiple localized regression models. *INFORMS Journal on Computing*, 29(3):503–522, 2017.

[51] Bimbisar Desai, Karen Dixon, Elizabeth Farrant, Qixing Feng, Karl R Gibson, Willem P van Hoorn, James Mills, Trevor Morgan, David M Parry, Manoj K Ramjee, et al. Rapid discovery of a novel series of abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. *Journal of medicinal chemistry*, 56(7):3033–3047, 2013.

[52] R Key Dismukes, Benjamin A Berman, and Loukia Loukopoulos. *The limits of expertise: Rethinking pilot error and the causes of airline accidents*. Routledge, 2017.

[53] Wanxue Dong, Maytal Saar-Tsechansky, and Tomer Geva. A machine learning framework towards transparency in experts' decision quality. *arXiv preprint arXiv:2110.11425*, 2021.

[54] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[55] Itiel E Dror and David Charlton. Why experts make errors. *Journal of Forensic Identification*, 56(4):600, 2006.

[56] Itiel E Dror, David Charlton, and Ailsa E Péron. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic science international*, 156(1):74–78, 2006.

[57] Itiel E Dror, Alvaro Pascual-Leone, Vilayanur Ramachandran, et al. The paradox of human expertise: why experts get it wrong. *The paradoxical brain*, 177, 2011.

[58] Itiel E Dror, Ailsa E Peron, Sara-Lynn Hind, and David Charlton. When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(6):799–809, 2005.

[59] Cynthia DuBois. The impact of "soft" affirmative action policies on minority hiring in executive leadership: The case of the nfl's rooney rule. *American Law and Economics Review*, 18(1):208–233, 2015.

[60] Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing ground truth for medical relation extraction. *arXiv preprint arXiv:1701.02185*, 2017.

[61] Michael W Dunham, Alison Malcolm, and J Kim Welford. Improved well log classification using semisupervised gaussian mixture models and a new hyper-parameter selection strategy. *Computers & Geosciences*, 140:104501, 2020.

[62] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[63] M. J. Bertin et al. *Pisot and Salem Numbers.* user Verlag, Berlin, 1992.

[64] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657, 2016.

[65] Sina Fazelpour and Maria De-Arteaga. Diversity in sociotechnical machine learning systems. *arXiv preprint arXiv:2107.09163*, 2021.

[66] George Fein, L Klein, and P Finn. Impairment on a simulated gambling task in long-term abstinent alcoholics. *Alcoholism: Clinical and Experimental Research*, 28(10):1487–1491, 2004.

[67] Michael Feldman. *Computational fairness: Preventing machine-learned discrimination.* PhD thesis, 2015.

[68] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[69] Mark Fenton-O'Creevy, Emma Soane, Nigel Nicholson, and Paul Willman. Thinking, feeling and deciding: The influence of emotions on the decision making and performance of traders. *Journal of Organizational Behavior*, 32(8):1044–1061, 2011.

[70] Gerald R Ferris, Timothy P Munyon, Kevin Basik, and M Ronald Buckley. The performance evaluation context: Social, emotional, cognitive, political, and relationship components. *Human Resource Management Review*, 18(3):146–163, 2008.

[71] David N Figlio. Teacher salaries and teacher quality. *Economics Letters*, 55(2):267–271, 1997.

[72] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. Fairness evaluation in presence of biased noisy labels. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2325–2336. PMLR, 26–28 Aug 2020.

[73] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and

Alexandra Chouldechova. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. *arXiv preprint arXiv:2105.04953*, 2021.

[74] Antonio Foncubierta Rodríguez and Henning Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 9–14, 2012.

[75] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[76] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 562–577. Springer, 2014.

[77] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337 – 407, 2000.

[78] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM Conference on*

Hypertext and Social Media, HT '14, page 218–223, New York, NY, USA, 2014. Association for Computing Machinery.

[79] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[80] Ruijiang Gao and Maytal Saar-Tsechansky. Cost-accuracy aware adaptive labeling for active learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2569–2576, Apr. 2020.

[81] Ruijiang Gao and Maytal Saar-Tsechansky. Cost-accuracy aware adaptive labeling for active learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2569–2576, 2020.

[82] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[83] Tomer Geva and Maytal Saar-Tsechansky. Who's a good decision maker? data-driven expert worker ranking under unobservable quality. In *Proceedings of the Thirty Seventh International Conference on Information Systems*, 2016.

[84] Tomer Geva and Maytal Saar-Tsechansky. Who is a better decision

maker? data-driven expert ranking under unobserved quality. *Production and Operations Management*, 30(1):127–144, 2021.

[85] Tomer Geva, Maytal Saar-Tsechansky, and Harel Lustiger. More for less: adaptive labeling payments in online labor markets. *Data Mining and Knowledge Discovery*, 33(6):1625–1673, 2019.

[86] Arpita Ghosh and Preston McAfee. Crowdsourcing with endogenous entry. In *Proceedings of the 21st international conference on World Wide Web*, pages 999–1008, 2012.

[87] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[88] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019.

[89] Mark Graber, Ruthanna Gordon, and Nancy Franklin. Reducing diagnostic errors in medicine: what's the goal? *Academic Medicine*, 77(10):981–992, 2002.

[90] Richard C Grote. *Forced ranking: Making performance management work*. Harvard Business School Press Boston, MA, 2005.

[91] Yuhong Guo. Active instance sampling via matrix partition. *Advances in Neural Information Processing Systems*, 23, 2010.

[92] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[93] Michael M Harris and John Schaubroeck. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *personnel psychology*, 41(1):43–62, 1988.

[94] Rodney A Hayward. Counting deaths due to medical errors. *JAMA*, 288(19):2404–2404, 2002.

[95] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[96] Shen-Shyang Ho and Harry Wechsler. Query by transduction. *IEEE transactions on pattern analysis and machine intelligence*, 30(9):1557–1571, 2008.

[97] Mokter Hossain and Ilkka Kauranen. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 2015.

[98] Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. Cost-effective active learning from diverse labelers. In *IJCAI*, pages 1879–1885, 2017.

[99] Robert JB Hutton and Gary Klein. Expert decision making. *Systems Engineering: The Journal of The International Council on Systems Engineering*, 2(1):32–45, 1999.

[100] Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, 2014.

[101] Stefanie K Johnson. What amazon's board was getting wrong about diversity and hiring'. *Harvard Business Review*, 2018.

[102] Stefanie K Johnson, David R Hekman, and Elsa T Chan. If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, 26(04), 2016.

[103] Paul E Jones and Peter HMP Roelofsma. The potential for social contextual and group biases in team decision-making: Biases, conditions and psychological mechanisms. *Ergonomics*, 43(8):1129–1152, 2000.

[104] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009.

[105] Daniel Kahneman. Article commentary: Judgment and decision making: A personal view. *Psychological science*, 2(3):142–145, 1991.

[106] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases.* Cambridge university press, 1982.

[107] Ece Kamar and Eric Horvitz. Incentives for truthful reporting in crowdsourcing. In *AAMAS*, volume 12, pages 1329–1330, 2012.

[108] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

[109] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[110] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. *Advances in Neural Information Processing Systems*, 24:1953–1961, 2011.

[111] David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[112] Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In *European Conference on Information Retrieval*, pages 165–176. Springer, 2011.

[113] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944, 2011.

[114] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval*, 16(2):138–178, 2013.

[115] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

[116] Michael P Kim, Omer Reingold, and Guy N Rothblum. Fairness through computationally-bounded awareness. *arXiv preprint arXiv:1803.03239*, 2018.

[117] M Kiruthiga and P Sangeetha. Improving labeling quality using positive label frequency threshold algorithm. *International Journal of Computer Science and Engineering Communications*, 4(6):1467–1473, 2016.

[118] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

[119] Gary A Klein, Judith Orasanu, Roberta Calderwood, Caroline E Zsam-

bok, et al. *Decision making in action: Models and methods*, volume 3. Ablex Norwood, NJ, 1993.

[120] Richard J Klimoski and Manuel London. Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59(4):445, 1974.

[121] Donald K. Knuth. *The TEXbook*. Addison-Wesley, 1984.

[122] Akhil Alfons Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *Researchgate Preprint*, pages 1–19, 2019.

[123] Marios Kokkodis. Dynamic recommendations for sequential hiring decisions in online labor markets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 453–461, 2018.

[124] Marios Kokkodis. Dynamic, multidimensional, and skillset-specific reputation systems for online work. *Information Systems Research*, 32(3):688–712, 2021.

[125] Marios Kokkodis and Panagiotis G Ipeirotis. Reputation transferability in online labor markets. *Management Science*, 62(6):1687–1706, 2016.

[126] Miriam Komaromy, Kevin Grumbach, Michael Drake, Karen Vranizan, Nicole Lurie, Dennis Keane, and Andrew B Bindman. The role of black and hispanic physicians in providing health care for underserved populations. *New England Journal of Medicine*, 334(20):1305–1310, 1996.

[127] Daniel Kondermann. Ground truth design principles: an overview. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, pages 1–4, 2013.

[128] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606, 2015.

[129] Gloria J Kuhn. Diagnostic errors. *Academic Emergency Medicine*, 9(7):740–750, 2002.

[130] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[131] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189, 2020.

[132] Leslie Lamport. *LaTeX: A document preparation system.* Addison-Wesley, 2nd edition, 1994.

[133] Alexandre Louis Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. Noise-tolerant fair classification. *arXiv preprint arXiv:1901.10837*, 2019.

[134] Frank J Landy and James L Farr. Performance rating. *Psychological bulletin*, 87(1):72, 1980.

[135] Lucian L Leape, Troyen A Brennan, Nan Laird, Ann G Lawthers, A Russell Localio, Benjamin A Barnes, Liesi Hebert, Joseph P Newhouse, Paul C Weiler, and Howard Hiatt. The nature of adverse events in hospitalized patients: results of the harvard medical practice study ii. *New England journal of medicine*, 324(6):377–384, 1991.

[136] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[137] Wil Leitner. Nfl disgraces rooney rule by giving 'token interviews' to black candidates, Jan 2020.

[138] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[139] Alexander Hanbo Li and Jelena Bradic. Boosting in the presence of outliers: adaptive classification with nonconvex loss functions. *Journal of the American Statistical Association*, 113(522):660–674, 2018.

[140] Jiyi Li, Yasushi Kawase, Yukino Baba, and Hisashi Kashima. Performance as a constraint: An improved wisdom of crowds using performance regularization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1534–1541, 2020.

[141] Moshe Lichman et al. Uci machine learning repository, 2013, 2013.

[142] Christopher H Lin, M Mausam, and Daniel S Weld. To re (label), or not to re (label). In *HCOMP*, 2014.

[143] Jeffrey A Linder, Jason N Doctor, Mark W Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker, and Craig R Fox. Time of day and the decision to prescribe antibiotics. *JAMA internal medicine*, 174(12):2029–2031, 2014.

[144] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[145] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

[146] Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

[147] Chen Change Loy, Tao Xiang, and Shaogang Gong. Stream-based active unusual event detection. In *Asian Conference on Computer Vision*, pages 161–175. Springer, 2010.

[148] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[149] F Mittelbach M Goosens and A Samarin. *The LaTeXCompanion*. Addison-Wesley, 1994.

[150] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[151] Winter Mason and Duncan J Watts. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 77–85, 2009.

[152] Carlos A Mazas, Peter R Finn, and Joseph E Steinmetz. Decision-making biases, antisocial personality, and early-onset alcoholism. *Alcoholism: Clinical and Experimental Research*, 24(7):1036–1040, 2000.

[153] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[154] Richard J McMurray, Oscar W Clarke, John A Barrasso, Dexanne B Clohan, Charles H Epps, John Glasson, Robert McQuillan, Charles W

Plows, Michael A Puzak, David Orentlicher, et al. Gender disparities in clinical decision making. *JAMA*, 266(4):559–562, 1991.

[155] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[156] Anay Mehrotra and L Elisa Celis. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 237–248, 2021.

[157] Prem Melville, Stewart M Yang, Maytal Saar-Tsechansky, and Raymond Mooney. Active learning for probability estimation using jensen-shannon divergence. In *European conference on machine learning*, pages 268–279. Springer, 2005.

[158] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134. PMLR, 2015.

[159] Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent corruption. *arXiv preprint arXiv:1605.00751*, 2016.

[160] G.T. Milkovich, J.M. Newman, and B.A. Gerhart. *Compensation*. Compensation. McGraw-Hill Irwin, 2011.

[161] Robert Miranda Jr, James MacKillop, Lori A Meyerson, Alicia Justus, and William R Lovallo. Influence of antisocial and psychopathic traits on decision-making biases in alcoholics. *Alcoholism: Clinical and Experimental Research*, 33(5):817–825, 2009.

[162] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[163] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, volume 26, pages 1196–1204, 2013.

[164] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *J. Mach. Learn. Res.*, 18(1):5666–5698, 2017.

[165] David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "big three". *Diagnosis*, 8(1):67–84, 2021.

[166] Minh-Quoc Nghiem and Sophia Ananiadou. Aplenty: annotation tool for creating high-quality datasets using active and proactive learning.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 108–113, 2018.

[167] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

[168] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[169] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134, 2000.

[170] Geoffrey R. Norman, Donald Rosenthal, Lee R. Brooks, Scott W. Allen, and Linda J. Muzzin. The Development of Expertise in Dermatology. *Archives of Dermatology*, 125(8):1063–1068, 08 1989.

[171] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM, 2010.

[172] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

[173] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

[174] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3934–3938. IEEE, 2017.

[175] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575, 2016.

[176] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(4), 2010.

[177] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[178] Daniel Reker. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies*, 32:73–79, 2019.

[179] Daniel Reker, Petra Schneider, and Gisbert Schneider. Multi-objective active machine learning rapidly improves structure–activity models and

reveals new protein–protein interaction inhibitors. *Chemical science*, 7(6):3919–3927, 2016.

[180] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *arXiv preprint arXiv:1709.01779*, 2017.

[181] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436, 2013.

[182] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

[183] Maytal Saar-Tsechansky and Foster Provost. Active sampling for class probability estimation and ranking. *Machine learning*, 54(2):153–178, 2004.

[184] Ali S Saber Tehrani, HeeWon Lee, Simon C Mathews, Andrew Shore, Martin A Makary, Peter J Pronovost, and David E Newman-Toker. 25-year summary of us malpractice claims for diagnostic errors 1986–2010: an analysis from the national practitioner data bank. *BMJ Quality & Safety*, 22(8):672–680, 2013.

[185] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Pro-*

ceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pages 253–260, 2002.

[186] Craig Eric Schneier and Richard W Beatty. The influence of role prescriptions on the performance appraisal process. *Academy of Management Journal*, 21(1):129–135, 1978.

[187] Burr Settles. Active learning literature survey. *Technical Report*, 2009.

[188] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

[189] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[190] James Shanteau. Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2):252–266, 1992.

[191] James Shanteau, David J Weiss, Rickey P Thomas, and Julia C Pounds. Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, 136(2):253–263, 2002.

[192] Victor Sheng, Jing Zhang, Bin Gu, and Xindong Wu. Majority voting and pairing with multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[193] Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, pages 1308–1318. PMLR, 2020.

[194] Hardeep Singh, Ashley N D Meyer, and Eric J Thomas. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations. *BMJ Quality & Safety*, 23(9):727–731, 2014.

[195] Rebecca Smith-Coggins, Mark R Rosekind, Stacy Hurd, and Kenneth R Buccino. Relationship of day versus night sleep to physician performance and mood. *Annals of emergency medicine*, 24(5):928–934, 1994.

[196] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[197] Michael Spivak. *The joy of TEX*. American Mathematical Society, Providence, R.I., 2nd edition, 1990.

[198] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.

[199] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019.

[200] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[201] Elliot J Sussman, William G Tsiaras, and Keith A Soper. Diagnosis of diabetic eye disease. *Jama*, 247(23):3231–3234, 1982.

[202] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[203] Wei Tang and Matthew Lease. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*, pages 1–6, 2011.

[204] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

[205] Philip E Tetlock. *Expert political judgment: How good is it? How can we know?-New edition*. Princeton University Press, 2017.

[206] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[207] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310, 2019.

[208] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[209] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[210] Alf J. van der Poorten. Some problems of recurrent interest. Technical Report 81-0037, School of Mathematics and Physics, Macquarie University, North Ryde, Australia 2113, August 1981.

[211] Monica Van Such, Robert Lohr, Thomas Beckman, and James M Naessens. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice*, 23(4):870–874, 2017.

[212] Jasmin Vassileva, Raul Gonzalez, Antoine Bechara, and Eileen M Martin. Are all drug addicts impulsive? effects of antisociality and extent of multidrug use on cognitive and motor impulsivity. *Addictive Behaviors*, 32(12):3071–3076, 2007.

[213] Vincent Violago and Nikko Quevada. Ai: The issue of bias. *Managing Intell. Prop.*, 277:32, 2018.

[214] Christopher JD Wallis, Angela Jerath, Natalie Coburn, Zachary Klaassen, Amy N Luckenbaugh, Diana E Magee, Amanda E Hird, Kathleen Armstrong, Bheeshma Ravi, Nestor F Esnaola, et al. Association of surgeon-patient sex concordance with postoperative outcomes. *JAMA surgery*, 2021.

[215] Guihua Wang, Jun Li, Wallace J Hopp, Franco L Fazzalari, and Steven F Bolling. Using patient-specific quality information to unlock hidden healthcare capabilities. *Manufacturing & Service Operations Management*, 21(3):582–601, 2019.

[216] Jing Wang, Panagiotis G Ipeirotis, and Foster Provost. Cost-effective quality assurance in crowd labeling. *Information Systems Research*, 28(1):137–158, 2017.

[217] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[218] Fabian L Wauthier and Michael Jordan. Bayesian bias mitigation for crowdsourcing. *Advances in neural information processing systems*, 24:1800–1808, 2011.

[219] Jeff A Weekley and Joseph A Gier. Ceilings in the reliability and validity of performance ratings: The case of expert raters. *Academy of Management Journal*, 32(1):213–222, 1989.

[220] Nicola White, Fiona Reid, Adam Harris, Priscilla Harries, and Patrick Stone. A systematic review of predictions of survival in palliative care: how accurate are clinicians and who are the experts? *PloS one*, 11(8):e0161407, 2016.

[221] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

[222] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting racial bias in hate speech detection. *arXiv preprint arXiv:2005.12246*, 2020.

[223] Xianglei Xing, Yao Yu, Hua Jiang, and Sidan Du. A multi-manifold semi-supervised gaussian mixture model for pattern classification. *Pattern Recognition Letters*, 34(16):2118–2125, 2013.

[224] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine learning*, 95(3):291–327, 2014.

[225] Changchang Yin, Buyue Qian, Shilei Cao, Xiaoyu Li, Jishang Wei, Qinghua Zheng, and Ian Davidson. Deep similarity-based batch mode

active learning with exploration-exploitation. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 575–584. IEEE, 2017.

[226] Jenny Yourstone, Torun Lindholm, Martin Grann, and Ola Svenson. Evidence of gender bias in legal insanity evaluations: A case vignette study of clinicians, judges and students. *Nordic journal of psychiatry*, 62(4):273–278, 2008.

[227] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[228] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[229] Bo Zhang, Zhen Chen, and Paul S Albert. Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. *Biometrics*, 68(4):1294–1302, 2012.

[230] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[231] Huanyu Zhang, Ying Xiao, Xinyue Zhao, Zhuang Tian, Shu-yang Zhang, and Dong Dong. Physicians' knowledge on specific rare diseases and its associated factors: a national cross-sectional study from china. *Orphanet Journal of Rare Diseases*, 17(1):1–13, 2022.

[232] Jing Zhang, Xindong Wu, and Victor S Shengs. Active learning with imbalanced multiple noisy labeling. *IEEE transactions on cybernetics*, 45(5):1095–1107, 2014.

[233] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*, 2018.

[234] Denny Zhou, John C Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *In M. I. Jordan, Y. LeCun, S. A. Solla (eds.), Advances in Neural Information Processing Systems (2204–2212). MIT Press, Cambridge, MA.*, 2012.

[235] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[236] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*, 2002.

[237] Laura Zwaan and Hardeep Singh. The challenges in defining and measuring diagnostic error. *Diagnosis*, 2(2):97–103, 2015.

Index

Vita

Craig William McCluskey was born in Minneapolis, Minnesota on 20 May 1950, the son of Dr. William R. McCluskey and Lucilla W. McCluskey. He received the Bachelor of Science degree in Engineering from the California Institute of Technology and was commissioned an Officer in the United States Air Force in 1971. He entered active duty in October, 1971, and was stationed in Denver, Colorado, Colorado Springs, Colorado, Panama City, Florida, and Sacramento, California. He separated from the USAF in 1975 and worked as an engineer for several small electronics companies in California before moving to Colorado Springs, Colorado to work for Hewlett-Packard in 1979. He left Hewlett-Packard in 1989 and joined a small company based in Herndon, Virginia, working out of his house as a "remote" engineer designing parts of the Alexis satellite for Los Alamos National Laboratories. Laid off when his portion of the satellite was completed, he applied to the University of Texas at Austin for enrollment in their physics program. He was accepted and started graduate studies in August, 1991.

Permanent address: 521 White Tail Ter.
Waxhaw, NC 78713

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.