Enhancing Image Classification Capabilities of Crowdsourcing-Based Methods through Expanded Input Elicitation

Romena Yasmin, [1] **Joshua T. Grassel,** [1] **Md Mahmudulla Hassan,** [2] **Olac Fuentes,** [2] **Adolfo R. Escobedo** [1]

[1] School of Computing and Augmented Intelligence, Arizona State University
[2] Department of Computer Science, University of Texas at El Paso
ryasmin@asu.edu, jgrassel@asu.edu, mhassan@miners.utep.edu, ofuentes@utep.edu, adres@asu.edu

Abstract

This study investigates how different forms of input elicitation obtained from crowdsourcing can be utilized to improve the quality of inferred labels for image classification tasks, where an image must be labeled as either positive or negative depending on the presence/absence of a specified object. Three types of input elicitation methods are tested: binary classification (positive or negative); level of confidence in binary response (on a scale from 0-100%); and what participants believe the majority of the other participants' binary classification is. We design a crowdsourcing experiment to test the performance of the proposed input elicitation methods and use data from over 200 participants. Various existing voting and machine learning (ML) methods are applied and others developed to make the best use of these inputs. In an effort to assess their performance on classification tasks of varying difficulty, a systematic synthetic image generation process is developed. Each generated image combines items from the *MPEG-7 Core Experiment CE-Shape-1 Test Set* into a single image using multiple parameters (e.g., density, transparency, etc.) and may or may not contain a target object. The difficulty of these images is validated by the performance of an automated image classification method. Experimental results suggest that more accurate classifications can be achieved when using the average of the self-reported confidence values as an additional attribute for ML algorithms relative to what is achieved with more traditional approaches. Additionally, they demonstrate that other performance metrics of interest, namely reduced false-negative rates, can be prioritized through special modifications of the proposed aggregation methods that leverage the variety of elicited inputs.

Introduction

Recent computer vision research has demonstrated significant performance improvement of machine learning (ML) approaches, in particular those based on deep convolutional neural networks, over conventional approaches for image classification and annotation tasks (Krizhevsky, Sutskever, and Hinton 2012; Tan and Le 2019; Zhai et al. 2021). However, these algorithms typically require a large and diverse set of annotated data to generate accurate classifications. Large amounts of annotated data are not always available, especially for tasks where producing high-quality meta-data is costly, such as-image based medical diagnosis (Cheplygina, de Bruijne, and Pluim 2019), pattern recognition in geospatial remote sensing data (Rasp et al. 2020; Stevens et al. 2020), etc. In addition, ML algorithms are often sensitive to perturbations in the data for complex visual tasks, that to some extent are even difficult for humans, such as object detection in cluttered backgrounds and detection of adversarial examples (Papernot et al. 2016; McDaniel, Papernot, and Celik 2016), due to the high dimensionality and variability of the feature space of the images.

A complementary approach for image classification that has received significant attention in various domain-specific applications is crowdsourcing. Its growth has been accompanied and propelled by the emergence of online crowdsourcing platforms (e.g., MTurk, Prolific), which are widely employed to recruit and compensate human participants to annotate and classify data that are difficult for machine-only approaches. In general, crowdsourcing works by utilizing the concept of the "wisdom of the crowd" (Surowiecki 2005), where the judgments or predictions of multiple participants are used to sift out the noise in individual predictions and better approximate a ground truth (Yi et al. 2012). Numerous studies over the last decade have established that, under the right circumstances and with the proper aggregation methods, the collective predictions of multiple non-experts are uncontroversially more accurate than those from almost any individual including well-informed experts. This concept of using groups to make collective decisions has been successfully applied to a number of visual tasks ranging from simple classification and annotation (Russakovsky et al. 2015) to complex real-world applications, including assessment of damages caused by natural disasters (Barrington et al. 2012) and segmentation of biomedical images for diagnostic purposes (Gurari et al. 2015).

One of the main challenges in crowdsourcing is judgment/estimation aggregation, that is, the combining or fusing of multiple sources of potentially conflicting information into a single representative judgment. Since the quality of the predictions is highly dependent on the method employed to consolidate the crowdsourced inputs (Mao, Procaccia, and Chen 2013), a vast number of works have focused on developing effective algorithms to tackle this task. Computational social choice is a field dedicated to the rigorous analysis and design of such data aggregation mechanisms (Brandt et al.

2016). Researchers in this field have studied extensively the properties of various voting rules (e.g., Majority Voting) to develop better classification algorithms. Traditional voting methods elicit a single form of input from each individual. A major drawback in practice is that popular voting methods are susceptible to outliers and to input errors when only a small number of inputs are involved (Yoo, Escobedo, and Skolfield 2020). More generally, relying on a single form of input elicitation may decrease the quality of the collective decision due to cognitive biases such as anchoring, bandwagon effect, decoy effect, etc. (Eickhoff 2018). Studies have also found that the choice of input modality (e.g., using rankings or ratings as a subjective judgment response) can play a significant role in the accuracy of group decisions (Escobedo, Moreno-Centeno, and Yasmin 2021) and predictions (Rankin and Grube 1980). These difficulties in data collection and aggregation mechanisms of crowdsourcing become even more prominent when the task at hand is complex. Researchers have suggested many potential ways of mitigating these limitations. One promising direction is the collection of richer data, i.e., using multiple forms of input elicitation. As a parallel line of inquiry, previous work suggests that specialized aggregation methods for integrating this data should be considered for making good use of these different pieces of information (Kemmer et al. 2020).

Although ML methods have been shown to perform exceedingly well in various classification tasks, these outcomes typically depend on relatively large data sets (Hsing et al. 2018). However, high amounts of richly annotated data are inaccessible in various situations and/or obtaining them is prohibitively costly. Yet in such situations where less data is available, ML methods provide a natural mechanism for incorporating multiple forms of crowdsourced inputs, since they are tailor-made for classification based on input features. This study investigates how the performance of crowdsourcing-based voting and ML methods can be improved for image classification tasks involving smaller data sets. Its aim is to answer three main questions:

- How can different forms of input elicitation be utilized to improve accuracy in image classification tasks?
- Does the incorporation of multiple forms of participant inputs within ML approaches produce more accurate predictions from small data sets?
- How can traditional voting methods be modified to improve different classification performance metrics such as reduced false negative rates?

To address these questions, we designed a crowdsourced experiment that elicits three inputs from participants: binary classification (1=positive or 0=negative); level of confidence in binary response (on a scale from 0-100%); and what participants believe the majority of the other participants' binary classification is. We develop simple rules inspired by voting methods to aggregate these inputs into a collectively determined classifier. In an attempt to harness the benefits of both collective human intelligence and machine intelligence, we also incorporate these elicited inputs into ML algorithms to develop more reliable and accurate visual screening classifiers. The results indicate that integrating these diverse forms of input elicitation, including self-reported confidence values, can improve accuracy and efficiency of crowdsourced computation.

Literature Review

In recent years, crowdsourcing has been widely applied to complete a variety of image labeling/classification tasks, from those requiring simple visual identification abilities to those that rely on specific domain expertise. Many studies have leveraged crowdsourcing to annotate large scale data sets often requiring subjective analysis such as conceptualized images (Nowak and Rüger 2010), scene-centric images (Zhou et al. 2014), and general purpose images from publicly available data sets (Deng et al. 2009; Everingham et al. 2010). Crowdsourcing techniques have also been successfully tailored to many other complex visual labeling/classification contexts that require profound domain knowledge including identifying fish and plants (He, van Ossenbruggen, and de Vries 2013; Oosterman et al. 2014), endangered species through camera trap images (Swanson et al. 2015), locations of targets (Salek, Bachrach, and Key 2013), land covers (Foody et al. 2018), and sidewalk accessibility (Hara, Le, and Froehlich 2012). Due to its low cost and rapid processing capabilities, another prominent use of crowdsourcing is classification of CT images in medical applications. Such tasks have included identifying malaria-infected red blood cells (Mavandadi et al. 2012), detecting clinical features of glaucomatous optic neuropathy (Mitry et al. 2016), categorizing dermatological features (Cheplygina and Pluim 2018), labeling protein expression (Irshad et al. 2017), and various other tasks (Nguyen et al. 2012; Mitry et al. 2013).

Despite its effectiveness at processing high work volumes, numerous technical challenges need to be addressed to maximize the benefits from this paradigm. One such challenge is obtaining good quality data (Ipeirotis, Provost, and Wang 2010a), which to a large extent depends on workers' inherent characteristics (Hube, Fetahu, and Gadiraju 2019; Eickhoff 2018) as well as monetary incentives (Mason and Watts 2009). In addition, tasks are often hindered by insincere and even malicious workers who aim to complete them as quickly as possible to receive payment (Downs et al. 2010). Several studies have performed comprehensive experiments to analyze the impacts of participant behavior on the quality of data and have suggested methods for better quality control. Sheng, Provost, and Ipeirotis (2008) suggested the use of repeated labeling with a cost to improve quality, especially in the presence of noisy data. They also showed that when the cost of processing is not free, repeated labeling is more effective and robust in producing data of good quality. For complex tasks where repeated labeling is too expensive, Su, Deng, and Fei-Fei (2012) proposed a multi-step rechecking approach to collect tight bounding boxes, where an initial worker is used for the main task and a second worker is used to check the validity of the result. Other quality verification approaches include using gold standard questions (Gadiraju et al. 2015), inferring worker quality probabilistically (Welinder et al. 2011), and using crowdsourc-

ing platforms to validate automatically labeled data (Foncubierta Rodríguez and Müller 2012; de Herrera et al. 2014). It is worth adding that, while automated machine learning methods can be used to detect and mitigate human label noise (Patrini et al. 2017; Lee et al. 2018), they come with a major caveat. Namely, these techniques heavily rely on large-scale data sets that are often unavailable for specialized tasks.

Another technical challenge in crowdsourcing is deploying effective aggregation mechanisms. The most commonly used method across various types of tasks is Majority Voting (MV) (Hastie and Kameda 2005). MV attains high accuracy on simple idealized tasks, but its performance tends to degrade on tasks that require more expertise. A suggested MV enhancement for the harder tasks is to elicit the participant's level of confidence (as a proxy of expertise) and to integrate these inputs within the aggregation mechanism. Grofman, Owen, and Feld (1983) suggested weighing each individual's inputs based on self-reported confidence of their respective responses, in accordance with the belief that individuals can estimate reliably the accuracy of their own judgments (Griffin and Tversky 1992). Another prominent method is the Slating algorithm developed by Koriat (2012a), which determines the response according to the most confident participant. For additional uses of confidence values to make decisions, see Mannes, Soll, and Larrick (2014) and Litvinova et al. (2020). Although subjective confidence values can in some cases be a valid predictor of accuracy (Görzen, Laux et al. 2019; Matoulkova 2017), in many others they may degrade performance owing to cognitive biases that prevent a realistic assessment of one's abilities (Saab et al. 2019). Another approach is to weigh responses based on some form of worker reliability. Khattak and Salleb-Aouissi (2011) used trapping questions with expert annotated labels to estimate the expertise level of workers. For domain-specific tasks where the majority can be systematically biased, Prelec, Seung, and McCoy (2017) introduced the Surprisingly Popular Voting method, which elicits two responses from participants: their own answer and what they think the majority of other participants' answer is. It then selects the answer that is "more popular than people predict".

Experimental Design

Prior to introducing the components of the experimental design, we describe the *MPEG-7 Core Experiment CE-Shape-1 Test Set* (Jeannin and Bober 1999), which is the source data from which the featured crowdsourcing activities are constructed. The data set is composed of black and white images of a diverse set of shapes and objects including animals, geometric shapes, common household objects, etc. In total, the data set consists of 1200 objects/shapes (referred to here as *templates*) divided into 60 object/shape classes, with each class containing 20 members. Figure 1 provides representative templates from 12 of these classes.

Instantiation of Images

The images used in the crowdsourcing experiment are constructed by instantiating and placing multiple *MPEG-7 Core Experiment CE-Shape-1 Test Set* templates onto a single image frame. The instantiation of the image template is specified with six adjustable parameters: density, scale, color, transparency, rotation, and target object. Each of these parameters is described in the following paragraphs.

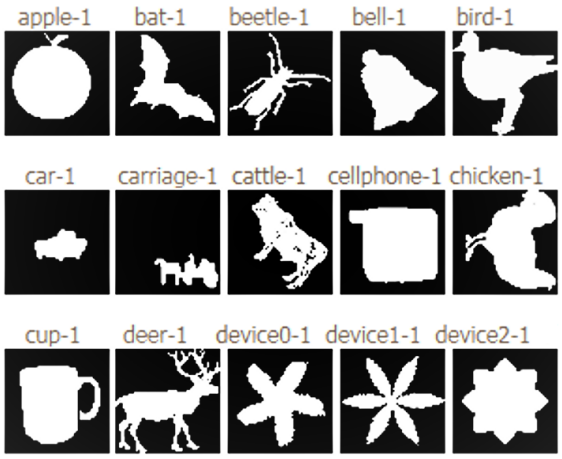
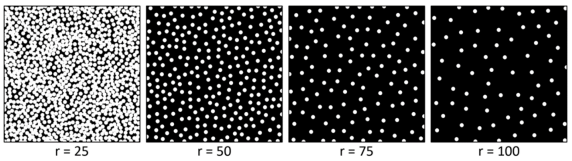
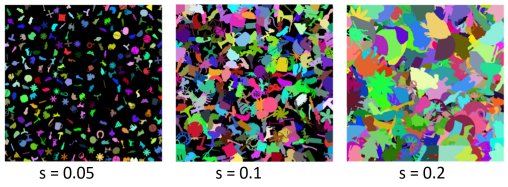


Figure 1: Object/shape templates from the *MPEG-7 Core Experiment CE-Shape-1 Test Set.*

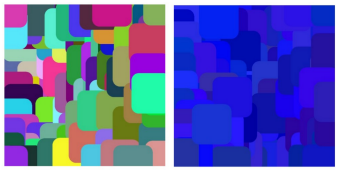
- **Density:** The center points of instantiated templates are set using a Poisson-disk distribution. This distribution produces closely packed individual points while maintaining a specified pixel distance, r. An example of the placement of dots over four settings of r (i.e., decreasing density) on a 1000×1000 pixel background is shown in Figure 2a. Once the dots are placed, the selected templates are inserted into the image to substitute them.
- **Scale:** The scale parameter, S, is defined on the interval $[0, 1]$. It specifies the relative size of the templates within an image, i.e., the ratio of the height of the object/shape template to the background. Figure 2b displays three images generated by increasing the value of S; the larger S is, the more cluttered the generated image becomes. S is randomly drawn from a triangular distribution $T(a, b, c)$.
- **Color:** The colors of the instantiated objects/shapes are set using RGB channels (red, green, and blue) where each channel is encoded as a number between 0 and 255; a specific combination of these values generates a unique color. Each instantiated object is assigned a single color. The RGB channel values are uniformly sampled from a restricted range, as specified by the distributions $U(a^R, b^R)$, $U(a^G, b^G)$, $U(a^B, b^B)$, where a and b are lower and upper bounds, respectively. Figure 2c displays an example of two color distributions. Alternatively, the color of an instantiated object can be randomly selected from a discrete set of RGB tuples.
- **Transparency:** The transparency value of an image, α, is drawn from a discrete uniform distribution $U(a, b)$. The parameter α is a number between 0 and 255, where smaller (larger, resp.) values are more transparent (opaque, resp.). Figure 2d displays four images generated



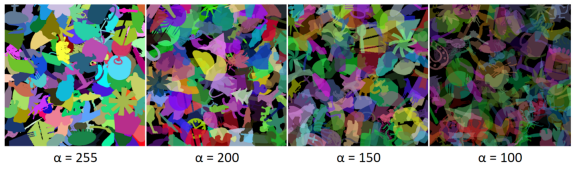
(a) Density



(b) Scale



(c) Color



(d) Transparency

Figure 2: Example images generated by varying parameters

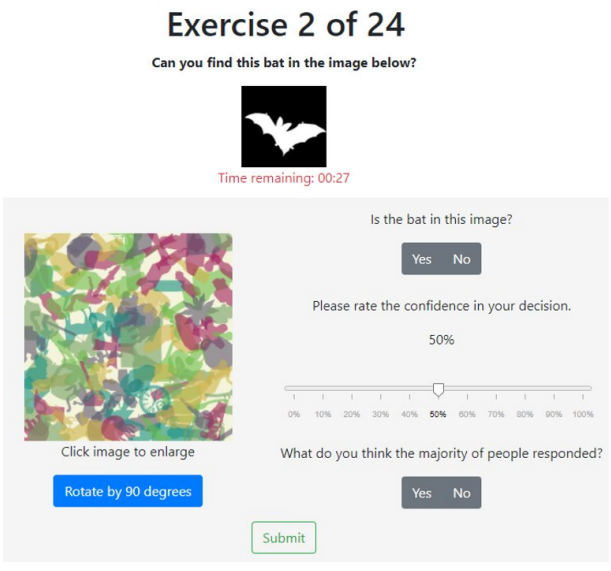


Figure 3: User interface of image classification task

with decreasing values of α.

- **Rotation:** The relative rotation, in degrees, of individual templates for an image is chosen by sampling from a uniform distribution $U(a, b)$.
- **Target Object:** When a target is specified, all templates of the same class are removed from the normal generation process. The target template is used in place of a single random template during the generation process to ensure it is present only once on the image frame.

In preliminary tests, we implemented the above image generation process to explore what image parameters contribute to the difficulty of an image classification task. With this knowledge, we designed experiments consisting of classification tasks of reasonable difficulty—that is, neither trivial nor impossible to complete.

Description of Activities

The crowdsourcing activities seek to address the research questions posed in the "Introduction" section. To that end, multiple forms of input were gathered from participants to complete a number of image classification tasks. To deploy the experiments, a web application was developed and then published on Amazon Mechanical Turk (MTurk). After participants were briefed about the study, they were prompted to complete a set of image classification tasks; images were shown to participants in randomized order. Figure 3 provides an example of the computer interface for one of these tasks along with the three questions posed to the participants (the associated inputs are discussed in the "Methodology" section). Participants had 60 seconds to provide their inputs before the image was hidden, at which point they had to provide their answers to proceed. The time limit was imposed to ensure the scalable implementation of a high number of tasks. In particular, the goal is to develop activities that can capture enough quality inputs from participants to improve collective estimates while mitigating potential cognitive fatigue. Seven experiments were completed and grouped into two sets: Experiment Set A and Experiment Set B. Each experiment used a balanced set of images, with half containing the target template; target objects were chosen so as to avoid confusion with other template classes. Images were generated using the above process with varying parameters on 1080×1080 beige background (RGB values $(245, 245, 220)$). The rotation of all templates follows the uniform distribution $U(0, 360)$. The remaining parameters are specific to each experiment and are summarized in Table 1. The image parameter ranges selected for Experiment Set A were designed to keep the difficulty of the classification tasks relatively uniform. On the other hand, a more complex set of parameters was selected for Experiment Set B, in part to explore the effects of the parameters on performance. The difference in difficulty is reflected in the individual performance achieved in these two experiment sets, measured by the average percentage of participants with the correct classification. For Experiment Set A, average individual performance was between 59% to 77% for each of the four experiments, whereas for Experiment Set B, it was between 53% to 80% for each of the three experiments.

Participants

A total of 251 participants across all experiments were recruited from Amazon MTurk. All participants were paid for their participation. Participants in Experiment Set A were paid $1.25; those in Experiment Set B were paid $2.00, commensurate with a higher number of questions and time spent. From these 251 participants, 45 were identified as *insincere participants* and their responses were subsequently removed from further analysis (see "Filtering of Insincere Participants" subsection for details). Of the remaining 206 participants, 174 completed the demographics survey. Their reported age ranged from 21 to 71 years old with a mean and median of 37 and 34, respectively. 114 participants reported their gender as male, 60 as female, and 0 as other. In terms of reported education level, 2 participants did not finish high school completion and 20 finished a high school/GED, 24 some college, 9 a two-year degree, 99 a four-year degree, 17 a master's degree, 1 a professional degree, and 2 a doctoral degree.

Methodology

Consider the image label aggregation problem where a set of images I are to be labeled by a set of participants P; without loss of generality, assume each image and participant has an unique identifier, that is, $I = \{i_1, i_2, ..., i_n\}$ and $P = \{p_1, p_2, ..., p_m\}$, where n and m represent respectively the total number of images and participants. For each image $i_k \in I$, the objective is to infer the binary ground truth label $y_k \in \{0, 1\}$, where $y_k = 1$ if the specified target object is present in the image and $y_k = 0$ otherwise. Since in this experiment each worker may label only a subset of the images, let $P_k \subseteq P$ be the set of participants who complete the labeling task of image $i_k \in I$. In contrast to most crowdsourced labeling tasks, where only a single label estimate is collected from participants, each participant $p_j \in P_k$ in the experiment was asked to provide three inputs. The first input is their binary choice label, $l_k^j \in \{0, 1\}$, regarding the presence/absence of the target object in image i_k. The second input is a numeric confidence value, $c_k^j \in [0, 100]$, indicating the degree of confidence in their binary choice label. The third input is another binary choice value, $g_k^j \in \{0, 1\}$, indicating what p_j thought the binary label assigned to i_k by the majority of the participants in P_k would be. This input value is referred to in this study as the Guess of Majority Elicitation (GME).

The ensuing subsection briefly discusses some of the traditional aggregation methods: Majority Voting, Confidence Weighted Majority Voting, and Surprisingly Popular Voting. Each of these methods uses a different input format and/or multiple inputs for the classification task. We also introduce a customized voting method that utilizes both the c_k^j and g_k^j parameters within a Majority Voting framework, denoted here as the Modified Majority Voting method. In addition to these voting methods, we describe how the collected input elicitations are incorporated as features in ML methods to generate predictions.

Majority Voting (MV)

Majority Voting is the most widely used aggregation method due to its computational simplicity. Recall that in the featured experiments each participant's binary choice answer can be either 0 or 1. Therefore in this study, for an image i_k, the binary choice value that receives the highest number of votes is selected as the final label when MV is used. This label can be written as:

$$y_k^* = \underset{l \in \{0,1\}}{\arg\max} \sum_{p_j \in P_k} \mathbb{1}(l_k^j = l),$$

where, $\mathbb{1}(.)$ is an indicator function, which equals to 1 whenever the given argument inside the bracket is true; otherwise it equals to 0.

Confidence Weighted Majority Voting (CWMV)

An implicit assumption made when using Majority Voting is that participants within a crowdsourcing platform are equally reliable and, therefore, their provided labels should have equal weights when they are aggregated. However, this is not always ideal, specially in the presence of noisy labelers. One simple approach to address this problem is to weigh the participants' labels according to their self reported confidence values. This is based on the idea that participants can accurately assess confidences in their independently formed decisions (Meyen et al. 2021). Therefore more weight is given to participants who are more confident regarding their answers than those who are more conflicted. In this method, for a given image i_k, the binary choice label for which the sum of the individual confidence values is the highest is selected as the final label; this can be written as:

$$y_k^* = \underset{l \in \{0,1\}}{\arg\max} \sum_{p_j \in P_k} c_k^j \mathbb{1}(l_k^j = l).$$

Surprisingly Popular Voting (SPV)

The Surprisingly Popular Voting method leverages the idea that for some domain–specific questions where the majority of the crowd is highly inaccurate, participants who are accurate but are in the minority may also know that their response is rare (Rutchick et al. 2020). Comparison of the number of participants who agree on a label to the number of participants who think the given label will be the choice provided by the majority often allows the minority choice to be selected as the final label. The predicted label of image i_k using this method can be written as:

$$y_k^* = \underset{l \in \{0,1\}}{\arg\max} \sum_{p_j \in P_k} \left[\mathbb{1}(l_k^j = l) - \mathbb{1}(g_k^j = l) \right].$$

Modified Majority Voting (MMV)

In general, self-reported confidence values represent the level of conviction in one's decision rather than its correctness. Therefore, for consensually wrong questions (i.e., questions where most participants are wrong), reliance on the most confident participant can result in a higher error rate (Koriat 2012b). In contrast to this belief, for specific use

Exp.		Images	Density	Scale	Color	Transparency	Target
Set A	#1 #2 #3 #4	16	{100, 120, 140, 160}	$\{T(0.08, 0.2, 0.32), .., T(0.53, 0.65, 0.77)\}$	Discrete: {4}	$U(100, 200)$	Bat Butterfly Apple Stingray
Set B	#5 #6 #7	24	{80} {80,100,120} {100, 150}	$\{T(0.15, 0.2, 0.25), T(0.25, 0.3, 0.35)\}$ $\{T(0.15, 0.2, 0.25), .., T(0.35, 0.4, 0.45)\}$ $\{T(0.15, 0.2, 0.25), T(0.25, 0.3, 0.35)\}$	Discrete: {1,...,6} $U(10, 255), U(10, 255), U(10, 255)$	$U(140, 170)$	Bat Turtle Various-7

Table 1: Summary of experiment image parameters

cases such as the featured image classification tasks, confidence values can provide additional useful information to binary responses. Participants able to locate the target object in a challenging image may intuitively express higher confidence in their response. Moreover, for these specific cases, it may be appropriate to assume that these participants will also be able to infer the difficulty of the question. The latter will be reflected through the difference between their two binary choice response parameters, l_k^j and g_k^j. In the customized voting method, these two concepts are leveraged for reclassifying, the images for which the MV inference label was 0, i.e., the specified object is absent. In addition, CWMV is used as a tie-breaking rule to generate more decisive labels as an intermediate step. The steps used for labeling an image i_k using MMV can be summarized as follows:

- **Step 1:** Predict the image class, y_k^* using MV
- **Step 2:** Use CWMV as a tie-breaking rule for instances where, $\sum_{p_j \in P_k} \mathbb{1}(l_k^j = 1) = \sum_{p_j \in P_k} \mathbb{1}(l_k^j = 0)$
- **Step 3:** If $y_k^* = 0$, check if there exists a participant $p_j \in P_k$ with confidence value $c_k^j = 100\%$ such that, $l_k^j = 1$ and $g_k^j = 0$.
- **Step 4:** If there exists a participant $p_j \in P_k$ whose responses matches the requirements stated in step 3, then set $y_k^* = 1$.

Crowdsourcing-based ML Methods

Four ML based binary classification approaches were selected: Logistic Regression (LR), Random Forest Classifier (RF), Support Vector Machines (SVM), and K-Nearest Neighbor (KNN). These were selected as reasonable representatives of available methods. The ML classifiers were trained and evaluated using built-in functions of the Python *scikit-learn* library (Pedregosa et al. 2011). The hyperparameters were optimized on a linear grid search with a nested 5-fold cross validation strategy. Four features were extracted from the three inputs elicitated from participants for use with the ML classifiers; these features are described in the ensuing paragraphs.

- **Binary Choice Elicitation (BCE)**: For each image $i_k \in I$ the binary choice elicitation values are divided into two sets: one containing the participants with response $l_k^j = 1$ and the other containing participants with response $l_k^j = 0$. The number of participants in each set can be used as an input feature within a ML classifier. However, in practical settings since the number of participants can vary from image to image, it is more prudent to use the relative size of the sets. Note that these relative sizes are complements of each other, i.e., the fraction of participants who chose $l_k^j = 1$ as their binary choice label can be determined by subtracting from 1.0 the fraction of participants who chose $l_k^j = 0$. Therefore, to remove redundancy and colinearity within the features, only one of these values is used as an input as follows:

$$x_k^0 = \frac{\sum_{p_j \in P_k} \mathbb{1}(l_k^j = 0)}{|P_k|},$$

  where, x_k^0 represents the fraction of participants who agree on the absence of the target object in image i_k.
- **Guess of Majority Elicitation (GME)**: Similar to BCE, GME is also converted into a single feature based on the number of participants whose g_k^j response value is 0:

$$x_k^{GME,\,0} = \frac{\sum_{p_j \in P_k} \mathbb{1}(g_k^j = 0)}{|P_k|}.$$

- **Confidence Elicitation (CE)**: Although previous research has explored using confidence scores to improve annotation quality of crowdsourced data (Ipeirotis, Provost, and Wang 2010b), very few works have incorporated this input within a machine learning model. Another issue with using confidence values is that, even though some participants might be accurate in judging their performance, others might be prone to metacognitive biases, i.e., overconfident or underconfident in judging their actual abilities (Oyama et al. 2013). Hence, self-reported confidence should not be taken at face value, and specific confidence values should not be assumed to convey the same meaning across different individuals. In an attempt to mitigate such biases, the confidence values, $c_{k \subseteq \{1,2,..,n\}}^j$ provided by participant $p_j \in P$ were rescaled linearly between 0 and 100, with the lowest confidence value expressed by p_j being mapped to 0 and the greatest to 100. Let $I^j \subseteq I$ be the set of images for which p_j provided a label, then the confidence value for participant p_j classifying image i_k is rescaled as follows:

$$c_k^{j*} = \frac{c_k^j - \min_{i_q \in I^j} c_q^j}{\max_{i_q \in I^j} c_q^j - \min_{i_q \in I^j} c_q^j} \times 100.$$

The confidence values are further divided into two sets based on the l_k^j values. The average confidence values of each set are then used as additional features for the ML classifier. For image i_k, these two inputs can be expressed as follows:

$$x_k^{conf,\,1} = \frac{\sum\limits_{p_j \in P_k} c_k^{j*} \mathbb{1}(l_k^j = 1)}{\sum\limits_{p_j \in P_k} \mathbb{1}(l_k^j = 1)};$$

$$x_k^{conf,\,0} = \frac{\sum\limits_{p_j \in P_k} c_k^{j*} \mathbb{1}(l_k^j = 0)}{\sum\limits_{p_j \in P_k} \mathbb{1}(l_k^j = 0)}.$$

To prevent overfitting and mitigate a false sense of confidence in the accuracy values, a 5-fold cross-validation strategy was used to train and evaluate the ML classifiers. Furthermore, each experiment was repeated 20 times and the average values were recorded.

Results

This section compares the performance of the four voting methods and four ML methods listed in the "Methodology" section. The results are divided according to two performance metrics of interest: accuracy and false negative rate (i.e. rate at which method incorrectly fails to identify the presence of the target object).

Before comparing the crowdsourcing-based methods, we evaluate the performance of an automated image classification method on images generated according to the process described in the "Instantiation of Images" subsection. The automated method is built on a deep neural network called ResNet-34 (He et al. 2015). The purpose of this supplementary analysis is to validate the complexity of the classification tasks by exhibiting that state-of-the-art neural networks do not yet achieve satisfactory results on these images rather than to make any direct comparisons with the crowdsourcing-based methods. Nonetheless, its results suggest that these tasks are not only nontrivial but that they require human inputs to achieve better results.

ResNet-34 consists of 34 layers. By default, the model is trained on the ImageNet (Russakovsky et al. 2014) data set, which consists of 1.28 million images from 1,000 different classes. In preliminary testing, pre-trained weights were not found to be useful (due to the different characteristics of ImageNet and the synthetic images generated in this study). Hence, we trained ResNet-34 from scratch (without using pre-trained weights) on a data set that contains approximately 60,000 images. Almost half of these images contain the target object, a bat, and the rest do not. The images are representative of those used in Experiment Sets A and B, specifically experiments #1 and #5, in terms of difficulty and parameter settings described in the "Instantiation of Images" subsection. We chose a representative object (a bat) as the target for the classification task because training the model for multiple objects would require more time and data; we assume the models' performance will be similar in such settings. An important change made to the original implementation of ResNet-34 for our study was to modify the last few fully-connected layers so that it is suitable for binary classification. To be specific, the number of neurons in the last three fully-connected layers were changed, where the output layers have only 2 neurons instead of the original 1,000. Its objective function, categorical cross entropy loss, was also replaced with binary cross entropy with logit loss, for this reason.

The ResNet-34 model was tested on a set of 3,000 images and achieved an accuracy of 61.8%. We attribute its somewhat poor performance to two reasons. First, deep neural networks require an ample amount of data to learn image features. Although the model used a few hundred times more images than the crowdsourcing-based ML methods to train, the size of this data set was not large enough for the model to learn all distinct image features and differentiate between positive and negative classes. Second, convolutional neural networks are known to strongly rely on texture-dependent features, which are absent in the generated images. This dependency hurts the training of the ResNet-34 model and leads to improper model optimization for the given task. Since most of the traditional convolutional neural networks have similar characteristics in terms of feature learning, we infer that they would exhibit a similar performance on this data set.

Filtering of Insincere Participants

Before proceeding, it is necessary to address the quality of the participants recruited via the MTurk platform and the quality of data they provide. Because of the endemic presence in most crowdsourcing platforms of *insincere participants* (i.e. annotators who do not demonstrate an earnest effort), some criteria should be defined to detect their low quality inputs. This study defined two criteria for characterizing (and filtering out) an annotator as insincere:

- **Criterion 1:** The participant answered more than 75% of the questions in less than or equal to 10 seconds.
- **Criterion 2:** The participant's responses to all binary input questions were exclusively 0 or exclusively 1 for the entire question set.

Using the two above defined criteria, a total of 45 out of 251 participants were removed from the two study sets. Among these 45 participants, 19 fell under criterion 1 and the rest under criterion 2. As expected, cleansing of the raw data based on these criteria yielded less noisy inputs for the crowdsourcing-based aggregation methods.

Performance of Aggregation Methods

The first set of results assess the performance of the aggregation methods described in the "Methodology" section. The overall results of this experiment are summarized in Table 2. The performance of the ML methods is quantified via three performance metrics: Accuracy (Acc.), False Negative Rate (FNR), and Area Under the ROC Curve (AUC). For the voting methods, only the first two of these metrics are reported. For each of the ML classifiers, the best Accuracy, FNR, and

Input	Voting Methods								Machine Learning Methods											
	MV		CWMV		SPV		MMV		RF			LR			KNN			Linear SVM		
Elicitations	Acc.	FNR	Acc.	FNR	Acc.	FNR	Acc.	FNR	Acc.	FNR	AUC	Acc.	FNR	AUC	Acc.	FNR	AUC	Acc.	FNR	AUC
Experiment Set A																				
BCE	0.73	0.53	–	–	–	–	–	–	**0.86**	**0.18**	0.86	**0.88**	**0.16**	**0.90**	**0.89**	**0.16**	0.87	**0.87**	0.19	**0.91**
BCE, CE	0.73	0.53	0.70	0.56	–	–	–	–	0.81	0.22	**0.91**	0.85	0.20	0.89	0.84	0.26	**0.88**	0.86	0.20	0.90
BCE, GME	–	–	–	–	0.45	0.94	–	–	0.83	0.20	0.84	0.87	0.17	0.89	0.85	0.20	0.85	**0.87**	**0.16**	0.90
BCE, CE, GME	–	–	–	–	–	–	0.69	0.53	0.82	0.22	0.88	0.83	0.21	0.89	0.80	0.27	0.83	0.84	0.21	0.90
Experiment Set B																				
BCE	0.69	0.56	–	–	–	–	–	–	0.72	0.33	0.77	0.76	0.32	0.84	0.75	0.28	0.80	0.75	0.34	0.84
BCE, CE	0.71	0.53	0.72	0.42	–	–	–	–	**0.82**	**0.21**	**0.87**	0.81	0.25	0.88	**0.81**	**0.24**	**0.87**	0.80	0.26	0.86
BCE, GME	–	–	–	–	0.53	0.78	–	–	0.68	0.30	0.74	0.77	0.30	0.84	0.73	0.29	0.79	0.73	0.37	0.84
BCE, CE, GME	–	–	–	–	–	–	0.79	0.28	**0.82**	0.24	0.85	**0.83**	**0.22**	**0.89**	**0.81**	0.25	**0.87**	**0.83**	**0.24**	**0.88**
Combined Experiment Set																				
BCE	0.71	0.54	–	–	–	–	–	–	**0.82**	**0.23**	0.82	**0.84**	**0.21**	**0.88**	**0.83**	**0.21**	0.83	0.83	**0.22**	**0.89**
BCE, CE	0.72	0.53	0.71	0.49	–	–	–	–	0.80	0.24	**0.86**	0.81	0.24	0.87	0.81	0.22	**0.86**	0.80	0.26	0.86
BCE, GME	–	–	–	–	0.49	0.85	–	–	0.78	**0.23**	0.81	**0.84**	**0.21**	**0.88**	0.80	0.25	0.83	**0.84**	**0.22**	0.88
BCE, CE, GME	–	–	–	–	–	–	0.74	0.40	0.79	0.24	0.85	0.81	0.23	0.87	0.80	0.22	0.85	0.80	0.27	0.87

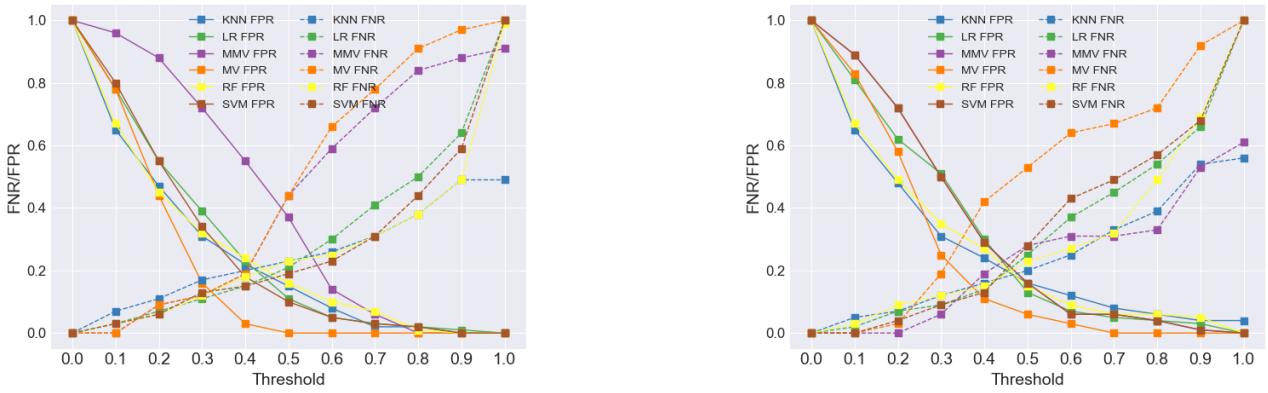
Table 2: Performance analysis of crowdsourcing based aggregation methods

AUC values among the different input elicitation combinations are marked in bold. Before proceeding, it is worthwhile to mention three additional points regarding the values presented in the table. First, the Combined Experiment Set is created by merging the data from Experiment Sets A and B. Second, when calculating the accuracy and FNR values of the voting methods, images with undecided outcomes (i.e. ties) are considered as a separate label. Third, the second row (BCE, CE) for each data set under column MV represents the performance of MV when CWMV is used as a tie-breaking rule.

Let us first discuss the performance of the aggregation models in terms of accuracy. For Experiment Sets A and B as well as the Combined Experiment Set, the average accuracy values of MV were stable in the range of $69\% - 73\%$. Both CWMV and the customized voting method MMV performed slightly worse than MV in Experiment Set A but had a higher accuracy in Experiment Set B. However, upon further inspection of the images in Experiment Set A, we observed that the silhouettes of two of the target images (the sting ray and the cup) were too similar and could not be differentiated properly by participants. This may have created a false sense of confidence in the participants' responses, thereby reducing the accuracy of the voting methods. SPV was the worst performer across the board with an average accuracy value of less than 50%. This low performance value can be largely attributed to the excessive number of tied labels generated compared to the other methods. In SPV, 21 out of the 136 instances were classified as tied (i.e., participants were undecided regarding the guess of the majority's estimate). By comparison, there were only 5 tied instances with MV and 1 with CWMV. The latter result motivated the use of CWMV as a tie-breaking rule within the MMV algorithm.

The results of the ML classifiers were relatively consistent in terms of both accuracy and AUC values for all four combinations of the input elicitations for the Combined Experiment Set (i.e., when a larger training data set is available). However, in Experiment Set B, the ML classifiers showed increase in accuracy and AUC values when the (BCE, CE) or (BCE, CE, GME) combination were used, as compared to the standalone BCE input. In Experiment Set A, the results followed a different pattern. In this case, the classifiers achieved the highest accuracy values when only the BCE input was used. The classifiers for this experiment set performed extremely well with accuracy values above 80% for all four combinations of input features. The fact that the images in Experiment Set A were generated using parameter ranges that were more consistent and less variable in difficulty may have been a contributing factor in this case. For this set, the highest accuracy was achieved by KNN for the BCE input. For RF, even though accuracy decreased from 86% to 81% when using (BCE,CE) combination instead of only BCE as inputs, AUC values were much larger (from 86% to 91%) for the (BCE, CE) combination, indicating that this input helps the model distinguish between classes more accurately. These results suggests that, even though the self-reported confidence values are not always particularly helpful when used within the traditional voting methods context (Saab et al. 2019; Li and Varshney 2017)—as can also be seen by the performance of the CWMV algorithm in this study—integrating CE into an ML classifier can help attain more accurate predictions when the sample size is small. Overall, ML classifiers outperformed the voting methods, with the exception of MMV which showed comparable performance in experiment set B. Note that a distinctive advantage of MMV over ML methods is that it does not require training data.

Another performance metric of interest is FNR, which denotes the fraction of images the methods label as 0 when their true label is 1. A high FNR is of much concern in a number of critical engineering and medical practices where a false negative may be more detrimental than a false positive, since the latter can be easily verified in subsequent steps. For example, FNR has significant importance on the detection of lung cancer from chest X-rays. If the model



(a) Set A

(b) Set B

Figure 4: Change in FNR/FPR of different aggregation methods under varying thresholds

falsely classifies an X-ray as negative, the patient may not receive needed medical care in a timely fashion. Returning to Table 2, the FNRs of the voting methods (MV, CWMV, SPV) are high across the board, with SPV again having the worst performance. The high FNRs of MV and CWMV can be attributed to the fact that people tend to provide the label "0" whenever they cannot find the object; these methods are unable to extract any additional useful information in these cases. MMV was specifically customized to address these limitations of the voting methods. As the results demonstrate, it successfully reduced FNRs in Experiment Set A, where MMV is able to reclassify some of the images as positive which the MV and CWMV method misclassified as negative. The performance increase of MMV compared to MV that can be attributed to the tie-breaking step (Step 2) is reported in the second rows of the MV column (BCE, CE). It is clear from the comparison between rows 1 and 2 of each set that, the tie-breaking rule slightly increases accuracy, but it does not help to reduce FNR. The decrease in FNR values is attributed primarily to the incorporation of the g_k^j and maximum c_k^j parameter within the algorithm.

The FNRs of the ML classifiers seem to be highly correlated with their accuracy values. More specifically, FNR values decrease as accuracy values increase. In fact, when the classifiers have an accuracy of over 80%, FNR values are always less than 30%. This outcome suggests that the ML classifiers are able to weigh the given information more accurately and provide a more balanced classification.

Changing the Threshold of Positive Classification

This subsection examines how voting methods can be modified to emphasize other important metrics of image classification. In particular, we seek to prioritize reduced false negative rates, which are relevant in various critical applications. We reduce FNRs of voting methods by lowering the threshold at which a positive classification is returned (i.e., changing the tipping point for returning a positive collective response). However, care must be exercised when lowering the threshold since this implicitly increases False Positive Rates (FPRs), which can also be problematic.

By default, the threshold at which voting methods return a positive response is fixed; for example, MV requires more than 50% of positive responses to return the positive class. Figure 4 illustrates the impacts of adjusting the thresholds for the voting methods as well as for the ML methods. The figure separates FNRs from FPRs for each method. Using MV as an example, note that decreasing the threshold from 0.5 to approximately 0.3 results in relatively small increases to the FPR but significantly decreases the FNR; further decreases cause a disproportionate increases to FPR. Hence, these inflection points can help guide how the thresholds can be set for each voting method to prioritize FNR. Note, that similar observations can be made about the FNRs of the ML methods.

Discussion

This section highlights key observations related to the research questions, along with the limitations of the study.

The experimental results demonstrate that supplementing binary choice elicitation with other forms of inputs can generate better classifiers. This is true for voting methods as long as they are used within a structured framework such as the MMV method. Because different combinations of input elicitation provide different information about an individual's ability to analyze an image (e.g., for positive images, confidence values coupled with selected labels can provide information about the reliability of the binary response), using them within a defined framework can provide improved classifications under the right circumstances. This is even more true of ML methods. By including diverse inputs as features in these methods, it is possible to obtain better classification at a relatively low cost (i.e., smaller training data sets compared to other ML approaches). Overall, when the training sample is small, the BCE and CE inputs implemented together within any of the ML classifiers generated more dependable results for data sets of varying levels of difficulty. These diverse inputs also help these ML classifiers improve other performance metrics such as AUC

values, which measure how much the model is capable of distinguishing between labels. Voting methods had a really poor performance with respect to FNRs. However, simple parametric modification (i.e., changing the threshold value) was shown to significantly reduce the FNR with comparatively small increases in the FPR.

Our study admittedly has some limitations. First of all, the approach used to filter "insincere participants" was relatively simple. To obtain a better quality data set, in future studies it will be helpful to deploy more sophisticated quality control techniques (e.g., Honeypot questions) to filter out unreliable or poor quality participants. Second, the designed data sets are balanced in terms of the number of positive and negative images present in each experiment. Future studies will assess the featured methods on more imbalanced data sets applicable to other critical contexts. Similarly, a third limitation is that the synthetic images generated for this work have certain characteristics that may not be generalizable to various real-world situations. Future work will look into expanding these methodologies.

Conclusion

Although crowdsourcing methods have been productive in image classification, they do not tap into the full potential of the wisdom of the crowd in one important respect. Tasks can be amplified to elicit and integrate multiple inputs from each participant; an easy-to-implement option, for example, is eliciting the level of confidence in one's binary response. This paper investigates how this information can be utilized with machine learning to enhance the capabilities of crowdsourcing. Its three main contributions are as follows. First it introduces a systematic synthetic image generation process that can be used to create image classification tasks of varying difficulty. Second, it demonstrates that, while reported confidence in one's response did not help much in improving the performance of voting methods, it enhanced the performance of machine learning methods, particularly when smaller training data sets were available. Third, it demonstrates that voting methods can be adapted to prioritize other metrics of interest of image classification (e.g., reduced false negative rates).

The code used to generate the synthetic images can be found at https://github.com/O-ARE/2D-Image-Generation-HCOMP.

Acknowledgments

The authors thank all participants in this study, which received institutional IRB approval prior to deployment. All authors gratefully acknowledge funding support from the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-03-04. The lead PI of the project (the fifth author) and two of the students (the first and second authors) also gratefully acknowledge support from the National Science Foundation (Award 1850355).

References

Barrington, L.; Ghosh, S.; Greene, M.; Har-Noy, S.; Berger, J.; Gill, S.; Lin, A. Y.-M.; and Huyck, C. 2012. Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics* 54(6).

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of computational social choice*. Cambridge University Press.

Cheplygina, V.; de Bruijne, M.; and Pluim, J. P. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* 54: 280–296.

Cheplygina, V.; and Pluim, J. P. 2018. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 105–111. Springer.

de Herrera, A. G. S.; Foncubierta-Rodrıguez, A.; Markonis, D.; Schaer, R.; and Müller, H. 2014. Crowdsourcing for medical image classification. In *Annual congress SGMI*, volume 2014.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Downs, J. S.; Holbrook, M. B.; Sheng, S.; and Cranor, L. F. 2010. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2399–2402.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

Escobedo, A. R.; Moreno-Centeno, E.; and Yasmin, R. 2021. An Axiomatic Distance Methodology for Aggregating Multimodal Evaluations. *Optimization-Online preprint 8223* .

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2): 303–338.

Foncubierta Rodríguez, A.; and Müller, H. 2012. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, 9–14.

Foody, G.; See, L.; Fritz, S.; Moorthy, I.; Perger, C.; Schill, C.; and Boyd, D. 2018. Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. *ISPRS International Journal of Geo-Information* 7(3): 80.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640.

Görzen, T.; Laux, F.; et al. 2019. Extracting the Wisdom from the Crowd: A Comparison of Approaches to Aggregating Collective Intelligence. Technical report, Paderborn University, Faculty of Business Administration and Economics.

Griffin, D.; and Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive psychology* 24(3): 411–435.

Grofman, B.; Owen, G.; and Feld, S. L. 1983. Thirteen theorems in search of the truth. *Theory and decision* 15(3): 261–278.

Gurari, D.; Theriault, D.; Sameki, M.; Isenberg, B.; Pham, T. A.; Purwada, A.; Solski, P.; Walker, M.; Zhang, C.; Wong, J. Y.; et al. 2015. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. In *2015 IEEE winter conference on applications of computer vision*, 1169–1176. IEEE.

Hara, K.; Le, V.; and Froehlich, J. 2012. A feasibility study of crowdsourcing and google street view to determine sidewalk accessibility. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, 273–274.

Hastie, R.; and Kameda, T. 2005. The robust beauty of majority rules in group decisions. *Psychological review* 112(2): 494.

He, J.; van Ossenbruggen, J.; and de Vries, A. P. 2013. Do you need experts in the crowd? A case study in image annotation for marine biology. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, 57–60.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385. URL http://arxiv.org/abs/1512.03385.

Hsing, P.-Y.; Bradley, S.; Kent, V. T.; Hill, R. A.; Smith, G. C.; Whittingham, M. J.; Cokill, J.; Crawley, D.; volunteers, M.; and Stephens, P. A. 2018. Economical crowdsourcing for camera trap image classification. *Remote Sensing in Ecology and Conservation* 4(4): 361–374.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010a. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010b. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67. ACM.

Irshad, H.; Oh, E.-Y.; Schmolze, D.; Quintana, L. M.; Collins, L.; Tamimi, R. M.; and Beck, A. H. 2017. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. *Scientific reports* 7(1): 1–10.

Jeannin, S.; and Bober, M. 1999. Description of core experiments for MPEG-7 motion/shape. *MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG99 N* 2690.

Kemmer, R.; Yoo, Y.; Escobedo, A.; and Maciejewski, R. 2020. Enhancing Collective Estimates by Aggregating Cardinal and Ordinal Inputs. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 73–82.

Khattak, F. K.; and Salleb-Aouissi, A. 2011. Quality control of crowd labeling through expert evaluation. In *Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds*, volume 2, 5.

Koriat, A. 2012a. The self-consistency model of subjective confidence. *Psychological review* 119(1): 80.

Koriat, A. 2012b. When are two heads better than one and why? *Science* 336(6079): 360–362.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25: 1097–1105.

Lee, K.-H.; He, X.; Zhang, L.; and Yang, L. 2018. CleanNet: Transfer Learning for Scalable Image Classifier Training With Label Noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Q.; and Varshney, P. K. 2017. Does confidence reporting from the crowd benefit crowdsourcing performance? In *Proceedings of the 2nd International Workshop on Social Sensing*, 49–54.

Litvinova, A.; Herzog, S. M.; Kall, A. A.; Pleskac, T. J.; and Hertwig, R. 2020. How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision* 7(3): 183.

Mannes, A. E.; Soll, J. B.; and Larrick, R. P. 2014. The wisdom of select crowds. *Journal of personality and social psychology* 107(2): 276.

Mao, A.; Procaccia, A. D.; and Chen, Y. 2013. Better human computation through principled voting. In *AAAI*.

Mason, W.; and Watts, D. J. 2009. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*, 77–85.

Matoulkova, B. K. 2017. *Wisdom of the Crowd: Comparison of the CWM, Simple Average and Surprisingly Popular Answer Method*. Master's thesis, Erasmus University Rotterdam, Rotterdam, Netherlands.

Mavandadi, S.; Dimitrov, S.; Feng, S.; Yu, F.; Sikora, U.; Yaglidere, O.; Padmanabhan, S.; Nielsen, K.; and Ozcan, A. 2012. Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. *PloS one* 7(5).

McDaniel, P.; Papernot, N.; and Celik, Z. B. 2016. Machine Learning in Adversarial Settings. *IEEE Security Privacy* 14(3): 68–72. doi:10.1109/MSP.2016.51.

Meyen, S.; Sigg, D. M.; von Luxburg, U.; and Franz, V. H. 2021. Group decisions based on confidence weighted majority voting. *Cognitive research: principles and implications* 6(1): 1–13.

Mitry, D.; Peto, T.; Hayat, S.; Morgan, J. E.; Khaw, K.-T.; and Foster, P. J. 2013. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of Images in the EPIC Norfolk Cohort on behalf of the UKBiobank Eye and Vision Consortium. *PloS one* 8(8): e71154.

Mitry, D.; Zutis, K.; Dhillon, B.; Peto, T.; Hayat, S.; Khaw, K.-T.; Morgan, J. E.; Moncur, W.; Trucco, E.; and Foster, P. J. 2016. The accuracy and reliability of crowdsource annotations of digital retinal images. *Translational vision science & technology* 5(5): 6–6.

Nguyen, T. B.; Wang, S.; Anugu, V.; Rose, N.; McKenna, M.; Petrick, N.; Burns, J. E.; and Summers, R. M. 2012. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology* 262(3): 824–833.

Nowak, S.; and Rüger, S. 2010. 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*, 557–566.

Oosterman, J.; Nottamkandath, A.; Dijkshoorn, C.; Bozzon, A.; Houben, G.-J.; and Aroyo, L. 2014. Crowdsourcing knowledge-intensive tasks in cultural heritage. In *Proceedings of the 2014 ACM conference on Web science*, 267–268.

Oyama, S.; Baba, Y.; Sakurai, Y.; and Kashima, H. 2013. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Papernot, N.; Mcdaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016*, Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016, 372–387. United States: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/EuroSP.2016.36. Publisher Copyright: © 2016 IEEE.; 1st IEEE European Symposium on Security and Privacy, EURO S and P 2016 ; Conference date: 21-03-2016 Through 24-03-2016.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Prelec, D.; Seung, H. S.; and McCoy, J. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541(7638): 532.

Rankin, W. L.; and Grube, J. W. 1980. A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology* 10(3): 233–246.

Rasp, S.; Schulz, H.; Bony, S.; and Stevens, B. 2020. Combining Crowdsourcing and Deep Learning to Explore the Mesoscale Organization of Shallow Convection. *Bulletin of the American Meteorological Society* 101(11): E1980–E1995.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211–252.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR* abs/1409.0575. URL http://arxiv.org/abs/1409.0575.

Rutchick, A. M.; Ross, B. J.; Calvillo, D. P.; and Mesick, C. C. 2020. Does the "surprisingly popular" method yield accurate crowdsourced predictions? *Cognitive research: principles and implications* 5(1): 1–10.

Saab, F.; Elhajj, I. H.; Kayssi, A.; and Chehab, A. 2019. Modelling cognitive bias in crowdsourcing systems. *Cognitive Systems Research* 58: 1–18.

Salek, M.; Bachrach, Y.; and Key, P. 2013. Hotspotting—a probabilistic graphical model for image object localization through crowdsourcing. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622.

Stevens, B.; Bony, S.; Brogniez, H.; Hentgen, L.; Hohenegger, C.; Kiemle, C.; L'Ecuyer, T. S.; Naumann, A. K.; Schulz, H.; Siebesma, P. A.; et al. 2020. Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society* 146(726): 141–152.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Swanson, A.; Kosmala, M.; Lintott, C.; Simpson, R.; Smith, A.; and Packer, C. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data* 2(1): 1–14.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2011. The multidimensional wisdom of crowds. Neural Information Processing Systems.

Yi, S. K. M.; Steyvers, M.; Lee, M. D.; and Dry, M. J. 2012. The wisdom of the crowd in combinatorial problems. *Cognitive science* 36(3): 452–470.

Yoo, Y.; Escobedo, A.; and Skolfield, K. 2020. A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings. *European Journal of Operational Research* 285(3): 1025–1041.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2021. Scaling vision transformers. *arXiv preprint arXiv:2106.04560* .

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning Deep Features for Scene Recognition using Places Database. *Advances in Neural Information Processing Systems* 27: 487–495.