CrowdLearn: A Crowd-AI Hybrid System for Deep Learning-based Damage Assessment Applications

Daniel (Yue) Zhang, Yang Zhang, Qi Li, Thomas Plummer, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame, IN, USA
{yzhang40, yzhang42, qli8, tplumme2, dwang5}@nd.edu

***Abstract*—Artificial Intelligence (AI) has been widely adopted in many important application domains such as speech recognition, computer vision, autonomous driving, and AI for social good. In this paper, we focus on the AI-based damage assessment applications where deep neural network approaches are used to automatically identify damage severity of impacted areas from imagery reports in the aftermath of a disaster (e.g., earthquake, hurricane, landslides). While AI algorithms often significantly reduce the detection time and labor cost in such applications, their performance sometimes falls short of the desired accuracy and is considered to be less reliable than domain experts. To exacerbate the problem, the black-box nature of the AI algorithms also makes it difficult to troubleshoot the system when their performance is unsatisfactory. The emergence of crowdsourcing platforms (e.g., Amazon Mechanic Turk, Waze) brings about the opportunity to incorporate human intelligence into AI algorithms. However, the crowdsourcing platform is also black-box in terms of the uncertain response delay and crowd worker quality. In this work, we propose the CrowdLearn, a crowd-AI hybrid system that leverages the crowdsourcing platform to troubleshoot, tune, and eventually improve the black-box AI algorithms by welding crowd intelligence with machine intelligence. The system is specifically designed for deep learning-based damage assessment (DDA) applications where the crowd tend to be more accurate but less responsive than machines. Our evaluation results on a real-world case study on Amazon Mechanic Turk demonstrate that CrowdLearn can provide timely and more accurate assessments to natural disaster events than the state-of-the-art AI-only and human-AI integrated systems.**

I. Introduction

The recent advances in artificial intelligence (AI) has transformed many important domains of modern life (e.g., transportation, finance, education, healthcare, and entertainment) and its encroachment is expected to intensify [1]–[4]. In this paper, we focus on an important AI application - *deep learning-based disaster damage assessment (DDA)* where deep neural network approaches are used to automatically identify damage severity of impacted areas from imagery reports in the aftermath of a disaster (e.g., earthquake, hurricane, landslides) [5], [6]. The assessments are then delegated to emergency response agencies (e.g., FEMA and police departments) to adopt appropriate countermeasures. Traditionally, the damage assessment was primarily done by domain experts, which suffers from the apparent limitation of the heavy labor cost of labeling and low efficiency in the presence of massive amount of data [7]. A set of AI algorithms have recently been developed to automatically label the damage severity in images from social media posts without the presence of domain experts [5], [6].

While AI algorithms can significantly reduce the labor cost and improve the detection efficiency in DDA applications, they are prone to various failure scenarios (see Figure 1). For example, the AI algorithms mistakenly report a severe damage for the images in Figure 1(a) and 1(b) and report no damage for images in Figure 1(c) and 1(d) (please refer to the detailed discussions under Figure 1). One main reason for the above failure scenarios is that the AI-based DDA algorithms can only capture the low level features of the images (e.g., color, layout, shapes) but fail to "understand" the high level context of the images (e.g., the story behind the image). Such failure may lead to severe consequences (e.g., the rescue team may be sent to the wrong places while places where people's lives are at stake are not responded). In contrast, human intelligence (HI) is often more accurate in such failure scenarios [8]. For example, humans can reliably assess the damage severity by identifying fake or irrelevant images (e.g., Figure 1(a) and 1(b)) and observing the actual events happening in the images (e.g., Figure 1(c) and 1(d)).

(a) Fake Image

(b) Close Up

(c) Low Resolution

(d) Implicit

Image (a) is a fake image showing a car falling from a huge cleavage of a road. Image (b) is a close-up of a crack on a road. The AI algorithms mistakenly return false detection result of "severe damage" for both images. Image (c) shows a disaster scene image with low resolution. Image (d) shows kids were injured and taken away from a damaged area. The AI algorithms mistakenly return false detection result of "no damage" for both images.

Figure 1: Examples failures of AI algorithm of DDA

Motivated by the unique and complementary advantages and

limitations of AI and HI, we propose CrowdLearn, a crowd-AI hybrid system that leverages HI to troubleshoot, tune and eventually improve the performance of AI-based DDA application. To acquire HI, we leverage the crowdsourcing platform (i.e., Amazon Mechanical Turk or MTurk) that provides a massive amount of freelance workers with low cost. However, two critical pitfalls exist by leveraging crowdsourcing platform: 1) the freelance workers may not be able to provide responses that are as accurate as domain experts due to the lack of experience/expertise; 2) the delay of the crowd workers can be potentially too high to be acceptable for DDA applications. These two pitfalls are further exacerbated by the black-box challenges of both the AI and crowdsourcing platform that are not well addressed by the existing literature in human-AI systems [9], [10]. We elaborate the challenges below.

Black-box AI Challenge: the first challenge in combining HI and AI lies in the black-box nature of AI algorithms. In particular, the lack of interpretability of the results from AI algorithms makes it extremely hard to diagnose the failure scenarios such as performance deficiency - why the AI model fails? Is this due to lack of training data or the model itself? Such questions make it hard for the crowd to effectively improve the black-box AI model. The interpretability issue has been initially identified in [10], [11] where accountable AI solutions were proposed to leverage humans as annotators to troubleshoot and correct the outputs of AI algorithms. However, these solutions simply use humans to verify the results of AI and ignore the issue where human annotators can be both slow and expensive. There also exist some human-AI systems that use crowdsourcing platforms to obtain labels or features to retrain the model [12], [13]. However, these systems do not address the problem where the AI algorithms themselves are problematic in which no matter how many training samples are added, the AI performance will not increase. Given the black-box nature of AI, the research question we address here is: *how do we accurately identify the failure scenarios of AI that can be effectively addressed by the crowd?*

Black-box Crowdsourcing Platform Challenge: the second unique challenge lies in the black-box nature of the crowdsourcing platform, which is characterized by two unique features. First, the requester (the DDA application that queries the platform) often cannot directly select and manage the workers in the crowdsourcing platform. In fact, the requester can only submit tasks and define the incentives for each task. The lack of control makes the incentive design for the crowdsourcing platform very difficult since we cannot cherry-pick the highly reliable and responsive workers to complete the tasks. For this reason, the current incentive design solutions that assume the full control of the crowd workers cannot be applied to our problem [14]–[18]. Second, the time and quality of the responses from the crowd workers are highly dynamic and unpredictable and their relationships to incentives are not trivial to model. Existing solutions often assume that more incentives will lead to less response time and high response quality [13], [19]. However, we found the quality of the responses from the crowd workers is diversified and does not simply depend on the level of incentives provided in our experiments (e.g., the quality can be high even with low incentives provided). Similarly, we observe the response delay from crowd is not simply proportional to the incentive level. With these unique features, the research question to tackle here is: *how to effectively incentivize the crowd to provide reliable and timely responses to improve AI performance*?

In this work, we design a CrowdLearn framework that leverages human feedback from the crowdsourcing platform to troubleshoot, calibrate and boost the AI performance in DDA applications. In particular, CrowdLearn address the black-box challenges of AI and the crowdsourcing platform by developing four new schemes: 1) a query set selection (QSS) scheme to find the best strategy to query the crowdsourcing platform for feedback; 2) a new incentive policy design (IPD) scheme to incentivize the crowd to provide timely and accurate response to the query; 3) a crowd quality control (CQC) scheme that refines the responses from the crowd and provides trustworthy feedback to the AI algorithms; 4) a machine intelligence calibration (MIC) scheme that incorporates the feedback from the crowd to improve the AI algorithms by alleviating various failure scenarios of AI. The four components are integrated into a holistic closed-loop system that allows the AI and crowd to effectively interact with each other and eventually achieve boosted performance for the DDA application. The CrowdLearn framework was evaluated using Amazon Mechanical Turk (MTurk) and a real-world DDA application. We compared CrowdLearn with the state-of-the-art baselines in both AI-only algorithms and human-AI frameworks. The results show that our scheme achieves significant performance gain in terms of classification accuracy in disaster damage assessment with reasonably low response time and costs.

II. Related Work

A. Human-AI Systems

Humans have traditionally been an integral part of artificial intelligence systems as a means of generating labeled training data [3], [11], [20]. Such a paradigm has been proven to be effective in supervised learning tasks such as image classification [21], speech recognition [22], autonomous driving [23], social media mining [24], and virtual reality [25]. However, it also suffers from two key limitations. First, some applications (e.g., damage assessment) may require a large amount of training data to achieve reasonable performance, which could be impractical due to the labor cost [5], [9]. Second, the AI models are often black-box systems and it is difficult to diagnose in the event of failure and unsatisfactory performance. To address these limitations, a few human-AI hybrid frameworks have been developed in recent years. For example, Holzinger *et al.* proposed the notion of interactive human machine learning ("iML") where humans directly interact with AI by identifying useful features that could be incorporated into the AI algorithms [26]. Branson *et al.* invented a human-in-the-loop visual recognition system to accurately classify the objects in the picture based on the descriptions

of the picture from humans [12]. Nushi *et al.* developed an accountable human-AI system that leverages workers on MTurk to identify the limitations of the AI algorithms [10] and provide suggestions to improve them. However, the above solutions largely ignored the innate limitations of the AI algorithms that cannot be simply improved by retraining the model with more data. In contrast, CrowdLearn proactively identifies the innate limitations of AI and develops a set of machine intelligence calibration strategies to address various failure cases. Moreover, the above human-AI systems also ignore the black-box nature of crowdsourcing platform and adopt a fixed-incentive strategy that randomly assigns data for the crowd to label. Such an approach could cause significant delay in acquiring the human labels. In contrast, CrowdLearn incorporates a context-aware reinforcement learning scheme to ensure quick and reliable response from the crowd.

B. Active Learning Frameworks

Active Learning (AL) is a common technique to combine machine and human intelligence in human-AI systems [13]. In an active learning framework, an AI algorithm actively asks for the labels of some instances from domain experts [27]. The major benefit of AL is that it selects a "subset" of data samples to be labelled and significantly reduces the labeling costs and improves the efficiency. For example, Ambati *et al.* proposed Active Crowd Translation (ACT), a new machine translation paradigm where active learning technique is applied to dynamically query the crowd for annotations of texts. The annotations are then used to train a AI model to automatically translate low resource languages [28]. Laws *et al.* proposed an active learning framework using a retraining technique for supervised learning tasks - the algorithm iteratively identify instances for the crowd to obtain the labels and retrain the model using the newly obtained labels [13]. However, these solutions could not handle scenarios where AI algorithms fail due to the flaws in their model design instead of insufficient training data. In contrast, CrowdLearn is able to diagnose the model and query the crowd to directly take over the AI algorithm in such failure scenarios. We compare our scheme with representative active learning frameworks in Section V.

C. AI-based Disaster Response

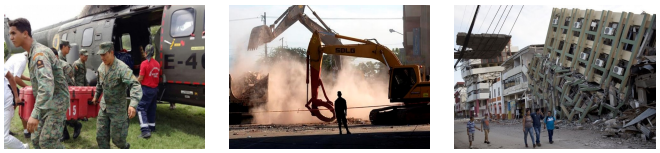
Disaster response is a critical application to ensure immediate resolution to emergent and hazardous events [29]–[33]. A critical step in disaster response is to perform damage assessment (e.g., determine the severity of the damage caused by a disaster based on imagery data). Traditionally, the damage assessment models were built on remote sensing data (e.g., satellite images). For example, Facebook recently proposed an AI framework to identify the areas that were severely affected by a disaster using convolutional neural networks (CNNs) on satellite imagery [34]. In a more recent work, Nguyan *et al.* developed a deep CNN model with domain-specific fine-tuning (referred to as VGG16) to effectively detect the level of damage from social media images [6]. Li *et. al* further extends the VGG16 model to accurately locate the damage area by combining CNN and Grad-CAM to generate a damage heatmap of a given image [5]. However, the above AI-driven solutions are incapable of providing accurate damage assessments in cases that deal with low-resolution or deceptive images. In this paper, we propose a novel scheme to significantly improve the performance of AI algorithm by welding the crowd wisdom with AI. To the best of our knowledge, CrowdLearn is the first Crowd-AI hybrid system in this application domain.

III. Problem Formulation

In this section, we first introduce the AI and crowd models respectively and then formally define our problem.

A. AI-based Disaster Damage Assessment Model

We first introduce the AI-based Disaster Damage Assessment (DDA) model. In a DDA application, images posted from social media related to a disaster event are dynamically crawled and classified based on the levels of the damage reported in the image. Figure 2 shows an example of different levels of damage from images in an DDA application. The damage assessment provides the critical information for emergency responses (e.g., sending out the rescue teams, allocating resources). The DDA application is constantly running and the images of the disaster are periodically crawled and analyzed. We refer to the updating period as a *sensing cycle*, which is formally defined below.



(a) No Damage (b) Moderate Damage (c) Severe Damage

Figure 2: Examples Output Labels of DDA

DEFINITION 1. Sensing Cycle (Ω): a period of time where new (unseen) data samples are collected.

We assume a DDA application has a total of T sensing cycles. The input data samples to the DDA algorithm is a set of N images, denoted as $I_1^t, I_2^t, ..., I_N^t$, where I_i^t denotes the i^{th} input image at the t^{th} sensing cycle. Each image I_i^t is associated with a ground truth label O_i^t and an estimated label (i.e., classification result) from the AI algorithm $\tilde{O}_i^t$.

As discussed in the introduction, we make a few observations about the deep learning-based DDA algorithms below.

1) *Black-box:* the DDA algorithms are black-box deep neural network models and the classification results in general lack interpretability.
2) *Failure accountability:* the AI-based DDA algorithms can fail (i.e., providing wrong classification labels for images) and the failure scenarios cannot be easily diagnosed without human scrutiny [11].

The above observations are critical in the design of the CrowdLearn scheme. To alleviate the performance deficiency

of the AI algorithms, we meld AI and crowd intelligence into a holistic system by leveraging the crowdsourcing platform. We elaborate the crowdsourcing platform model below.

B. Crowdsourcing Platform Model

Crowdsourcing platforms are well known for its cost efficiency and the massive amount of freelance workers [35]. We first define the key terms used in our crowdsourcing platform.

DEFINITION 2. Crowd Query (q_x^t): a set of questions assigned to the crowdsourcing platform.

DEFINITION 3. Query Response (r_x^t): the corresponding answers provided to the crowd query q_x^t.

An example query is shown in Figure 3. We assume a set of $X(t)$ queries are sent at each sensing cycle t to the crowdsourcing platform $\mathcal{Q}(t) = \{q_1^t, q_2^t, ..., q_X^t(t)\}$ where q_x^t denotes the x^{th} query submitted to the crowd at the t^{th} sensing cycle. Each query q_x^t is associated with an incentive provided by the application, denoted as b_x^t. We assume the application has a total budget of B for the crowdsourcing platform.

Figure 3: An Example of Crowd Query on MTurk

The responses to the queries are denoted as $\mathcal{R}(t) = \{r_1^t, r_2^t, ..., r_X^t(t)\}$ where r_x^t denotes the answer to q_x^t. For each query, two items are solicited from the crowd: the *label of the image* and *a set of questions*. The questions collect the contextual information observed by humans that cannot be easily extracted from the AI. For example, we ask humans whether the image is fake and what is actually happening in that image (e.g., car damage or bridge falling down). Such contextual information cannot be easily captured by AI but is crucial in determining the damage severity. We leverage the contextual information to decide the actual label of the image (more details in Section IV-C). Each response is associated with a response delay denoted as d_x^t. We also make a few observations about the crowdsourcing platform:

1) *Black-box:* the crowdsourcing platform is a black-box where the requester cannot directly control or pick the workers for the queries.
2) *Unreliable Workers:* the crowd workers are not perfectly reliable and can provide responses based on their own biases and personal opinions.
3) *Non-trivial incentive-delay-quality Relationship:* the relationship between incentives and the delay and quality of the response from the crowd cannot be simply modeled as linear relationships (e.g., the quality is proportional to the incentives and the delay is inversely proportional to the incentives). Instead, such relationship can be complex, dynamic and context-dependent.

The above observations are elaborated in Section IV-B. These observations of the crowdsourcing platform are critical in the design of the incentive mechanism and quality control schemes in CrowdLearn to ensure timely and accurate responses from the crowd.

Given the above definitions and assumptions, the goal of our CrowdLearn system can be formulated as a constrained multi-objective optimization problem. In particular, CrowdLearn targets maximizing the classification accuracy of the AI-based DDA algorithms, while minimizing the average delay from the crowd for a given budget on the crowdsourcing platform. The accuracy maximization objective ensures that the crowd can help AI assess the damage severity with a high accuracy in the absence of domain experts. The delay minimization objective ensures that the crowd provides feedback to AI in a timely manner. The resource constraints make sure the CrowdLearn framework does not incur unexpected excessive costs to the DDA application. Formally we have:

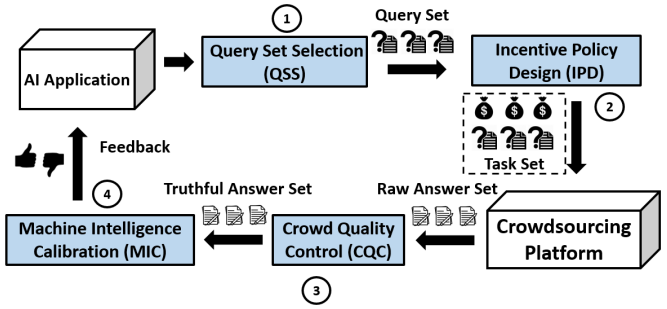
$$\begin{aligned} \text{max:} \quad & Pr(O_i^t = \tilde{O}_i^t | \mathcal{R}(t), B), \forall 1 \le i \le N, 1 \le t \le T \\ \text{min:} \quad & \frac{\sum_{x=1}^{X} d_x^t}{X}, \forall 1 \le x \le X(t), 1 \le t \le T \\ \text{s.t.:} \quad & \sum_{t=1}^{T} \sum_{x=1}^{X(t)} b_x^t \le B, 1 \le x \le X(t), 1 \le t \le T \end{aligned} \tag{1}$$

IV. The CrowdLearn Framework

An overview of the CrowdLearn framework is shown in Figure 4. The CrowdLearn is designed as a crowd-AI hybrid system that consists of four main modules: i) a Query Set Selection (QSS) scheme that identifies failure instances in AI algorithms and send queries to the crowd; ii) an Incentive Policy Design (IPD) scheme that takes the query set from QSS and assigns effective incentives for the queries to achieve the desired response delay; iii) a Crowd Quality Control (CQC) scheme that derives truthful answers from the crowd response; iv) a Machine Intelligence Calibration (MIC) scheme that incorporates query answers from CQC to improve the accuracy of the AI algorithms. We present them in detail below.

A. Query Set Selection (QSS)

The design of QSS is motivated by the two common failure scenarios of AI algorithms: 1) the lack of sufficient training data, and 2) the innate problem of the AI algorithm itself (e.g, oversimplified assumptions, inappropriate models). In CrowdLearn, we address the first failure scenario by actively asking the crowd to provide more reliable labels (e.g., the damage severity of specific images in DDA application) and use the labels to retrain the model. With more training samples that are judiciously selected by QSS, the performance of AI is expected to improve. We address the second failure scenario by



System workflow - ① QSS selects a set of data samples to query the crowd. ② IPD takes in the query set and generates a monetary incentive for each query. The query set with incentives are submitted to the crowdsourcing platform as a set of tasks. ③ The workers take the tasks and provide answers to the queries. CQC takes the answers from the crowd and provides quality control to generate truthful answers. ④ MIC compares the crowd answers with the results of the AI algorithms and improve their accuracy.

Figure 4: System Architecture of CrowdLearn

directly offloading the inference tasks to the crowd - i.e., ask the crowd to take over the AI algorithm. In both cases, we need to first identify the subset of data samples to be labeled from the crowdsourcing platform. Note that it is often impractical to send all data samples for the crowd to label due to budget and time constraints [7], [13]. The QSS module is designed to find the subset of data samples to query the crowd that can effectively address the failure scenarios of AI algorithms.

To identify the query subset, the key strategy is to identify the data samples that the AI algorithm is uncertain about - i.e., cannot confidently decide the label of the sample. Take the DDA as an example, if the AI algorithm cannot distinguish which damage level best describes the image, then it is better to send the image for the crowd to label. Based on this intuition, we first design a Query by Committee (QBC)-based active learning (AL) scheme to derive the uncertainty of the AL algorithms. In the QBC scheme, a set of relevant AI algorithms vote which new data sample needs to be queried from the crowd. Such technique has been proven to be robust by removing the bias of a single classifier [36]. We define a few key terms for our QBC-based model.

DEFINITION 4. Committee: a committee is set of AI algorithms for our DDA application.

DEFINITION 5. Expert: an expert is an AI algorithm selected into the committee.

In the DDA application, we choose a diverse set of M representative DDA algorithms $AI_1, AI_2, ..., AI_M$ that all take images as inputs. At a given sensing cycle, each expert (a deep neural network DDA algorithm) independently labels all the unseen data samples. The output of each expert is defined as an "expert vote".

DEFINITION 6. Expert Vote: an expert vote is the output of the AI algorithm, which is a probabilistic distribution of all possible class labels estimated by the algorithm.

We use $\mathcal{V}(AI^t_{m,i})$ to denote the vote of AI_m at a given data sample I^t_i by AI_m. For each algorithm AI_m, we define an expert weight - w^t_m as follows.

DEFINITION 7. Expert Weight (w^t_m): the level of trustworthiness of the algorithm in determining the final label of the data sample. The higher the weight, the more trustworthy the algorithm's classification result is.

In CrowdLearn, the expert weights are dynamically adjusted based on the feedback from the crowd. We discuss the adjustment process in the Section IV-D. Given the expert weights and votes of the experts, the committee decides the final label of the data sample (referred to as "committee vote"), which is the weighted sum of the label distributions of all committee members. Formally, we calculate the committee vote ρ for sample I^t_i as:

$$\rho^t_i = \sum_{m=1}^{M} w^t_m \times \mathcal{V}(AI^t_{m,i}) \tag{2}$$

The label distribution is further normalized with a sum of 1 to emulate an aggregated probabilistic distribution of the inference results. To quantify the uncertainty of AI algorithms in a committee, we define a new metric *Committee Entropy*.

DEFINITION 8. Committee Entropy ($\mathcal{H}$): the committee's overall uncertainty of labeling a data sample.

We derive $\mathcal{H}^t_i$ as the entropy of the normalized ρ^t_i.

$$\mathcal{H}^t_i = -\sum_{\rho \in \rho^t_i} Pr(\rho) \times \log Pr(\rho) \tag{3}$$

Given the committee entropy of each data sample, the intuitive query set selection strategy is to pick data sples with the highest committee entropy. However, such a stratamegy would fail when all classifiers confidently report the same wrong result. For example, in the DDA application, if all classifiers fail to capture fake images and report the fake images as "severe damage" with high confidence, the QSS will never pick the fake images for the crowd to check. Therefore, the QSS scheme also has to occasionally include the data samples with low committee entropy in the query set. This turns out to be an exploitation-exploration problem in reinforcement learning. We adopt an $\epsilon - greedy$ strategy [37] in our QSS scheme to solve this problem. We summarize the detailed procedure of QSS scheme in Algorithm 1.

B. Incentive Policy Design (IPD)

It is critical to provide timely and high quality responses from the crowd in the DDA application. Therefore, the CrowdLearn will decide how to incentivize the crowd after QSS selects a query set. We found the design of an incentive policy is a non-trivial problem due to two canonical challenges: 1) modeling the relationship between the incentives and the quality and delay of the crowd response is a non-trivial problem for the black-box crowdsourcing platform; 2) the quality and delay are context-dependent (e.g., the response

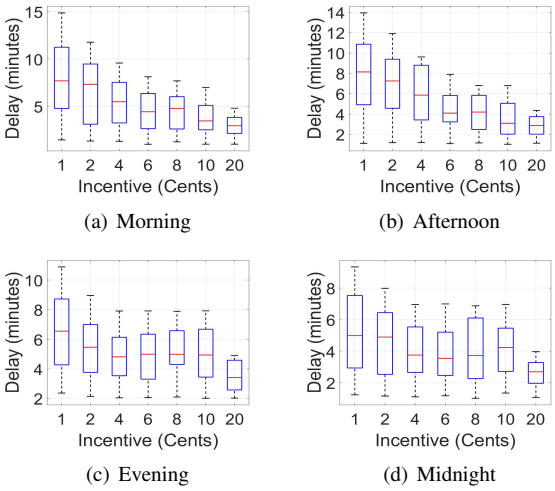
Algorithm 1 QSS Scheme

Input - Size of query set Y, Images $I_1^t, I_2^t, ..., I_N^t$
Initialize: $Committee$ = $AI_1, AI_2, ..., AI_M$, $votes$ = $newArray[M]$, $CommitteeEntropy$ = null, $output = newArray[Y]$
for $1 \leq t \leq T$ **do**
  for $1 \leq i \leq N$ **do**
    for $1 \leq m \leq M$ **do**
      $votes[m] \leftarrow \mathcal{V}(AI_{m,i}^t)$
    end for
    calculate $CommitteeEntropy$ using Equations (2) - (3).
  end for
  build sorted list s_list based on $CommitteeEntropy$ (high to low)
  for $1 \leq t \leq Y$ **do**
    $output.append(s\_list.pop())$ w.p. $(1-\epsilon)$
    $output.append(s\_list[rand(1, sizeof(s\_list))])$ w.p. ϵ
  end for
end for
return $output$

delay has different characteristics at different time of the day). To address the above challenges, we design a new reinforcement learning-based Incentive Policy Design (IPD) scheme to incentivize the black-box crowdsourcing platform for timely responses to the queries from the crowd.

1) Characterizing The Influence of Incentive on Response Delay and Quality: We first perform a pilot study to understand the effect of changing the level of incentives w.r.t. response delay and quality on our crowdsourcing platform MTurk. For the pilot study, we chose 7 incentive levels (1 cent, 2 cents, 4 cents, 6 cents, 8 cents, 10 cents, and 20 cents) and four different temporal contexts (morning, afternoon, evening and midnight). At each (incentive level, temporal context) combination, we assign a total of 100 HITs (Human Intelligence Tasks) to the MTurk platform: we issue a total of 20 queries and each query is allowed to be answered by 5 workers. In the existing literature, the response time of the human workers is often assumed to be proportional to the incentives provided [10], [13], [38]: the higher the incentive per task, the faster a crowd worker will provide a response. While this assumption is intuitive, our pilot study shows it may not always hold on MTurk. Figure 5 shows the response time of the crowd across different temporal contexts and incentive levels. We observe that the response time does seem to decrease as the incentive increases in the morning and afternoon. However, we also observe that most incentive levels (except for the lowest and highest) have very similar response time during evening and midnight. We attribute this observation to the fact that MTurk workers are often more active at night (e.g., after work) so that a query can always find a set of workers who are willing to take the HITs. However, during the day time, workers are less active and appear to be more "selective" in taking HITs. Such dynamics indicate the importance of considering the temporal contexts in the design of the incentive scheme for CrowdLearn.

We also study the quality of the annotated labels from the crowd with respect to the incentives provided. The results are shown in Figure 6. We observe that while very low incentives (e.g., 1 cent and 2 cents per HIT) generate relatively low label quality, increasing the incentives does not always significantly



(a) Morning (b) Afternoon (c) Evening (d) Midnight

Figure 5: Crowd Response Time vs. Incentives on MTurk

increase the quality. By performing Wilcoxon Test [39] (a statistical hypothesis test commonly used to compare two related samples), we found no statistical significance (significant if $p \leq 0.05$) when the incentive level increases from 2 to 4 (p = 0.12), from 4 to 6 (p = 0.45), from 6 to 8 (p=0.77), and from 8 to 10 (p=0.25)). We attribute the above observation to the fact that the image labeling tasks for the workers are relatively intuitive - workers often do not need to exert much effort/expertise to accurately label the images notwithstanding the incentives.

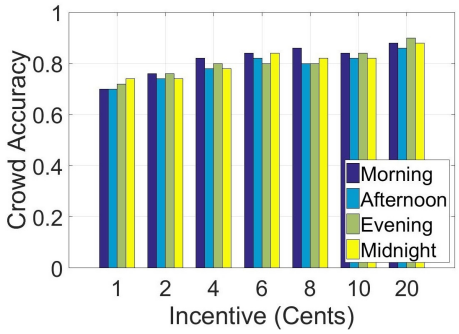


Figure 6: Label Quality vs. Incentives on MTurk

The above results offer us with the following design principles for the IPD module: 1) the context information must be incorporated into the policy design; 2) the dynamic response delay must be explicitly considered; 3) it may not be wise to increase the incentives merely for improved annotation quality. With the above principles, we design a reinforcement learning-based framework that targets at minimizing the response delay of the crowd workers. Note that we do not intend to use incentive mechanism to control the response quality considering the third design principle. Instead, we design another module to perform the crowd quality control in Section IV-C.

2) Constrained Contextual Multi-armed Bandit for Incentive Policy: We found that the incentive design problem can be nicely mapped to a constrained contextual multi-armed bandit (CCMB) problem in reinforcement learning. The key reason for choosing the contextual bandit is that it can incorporate the

context information into the incentive policy design (e.g., the contextual bandit can design fine-grained policy assignments under different temporal contexts). Also, the reinforcement learning framework allows the CCMB to dynamically adapt to the uncertain black-box crowdsourcing environments and derive the optimal incentive policy. We formally define the key terms in CCMB and its mapping to IPD below.

DEFINITION 9. Uncertain Environment: the uncertain environment in IPD refers to the black-box crowdsourcing platform that has the non-trivial incentive-delay tradeoff.

DEFINITION 10. Context: the context in IPD refers to the temporal context for the crowdsourcing platform. We choose four contexts - morning, afternoon, evening and midnight.

DEFINITION 11. Action: the action in IPD refers to the choice of incentive levels for a query.

DEFINITION 12. Payoff: the payoff in IPD refers to the additive inverse of the average delay of the query answers. The less delay, the higher payoff.

DEFINITION 13. Action Cost and Resource Budget: the action cost is the incentive set for each query. The resource budget is the total cost for using the crowdsourcing platform.

We formally define our CCMB model below. We consider a CCMB with a context set $\mathcal{X} = \{1, 2, ..., Z\}$ and an action set $\mathcal{A} = \{1, 2, ..., K\}$. An example context set is $\mathcal{X} = \{morning, afternoon, night, midnight\}$ and an example action set is $\mathcal{A} = \{1, 2, 4, 6, 8, 10, 20\}$ where each entry in $\mathcal{A}$ denotes the amount of money (in cents). We assume the crowdsourcing platform is associated with a specific context and each action $k \in \mathcal{A}$ generates a non-negative payoff p_k^t with cost c_k at each sensing cycle. We assume the conditional expectation $\mathbb{E}[p_k^t | \mathcal{X}^t = z]$ is unknown to the application. We use $\mathcal{C}^t$ to denote all the costs incurred at the t^{th} cycle. We assume the context $\mathcal{X}^t$ is observable at the beginning of a cycle. However, the payoff of the action taken by the agent is only revealed at the end of the cycle (i.e., you do not know the delay until the responses are submitted by the crowd).

The goal of CCMB is to derive an optimal incentive policy that decides to perform which action at which context to maximize the payoffs while keeping the total action cost within the resource budget. The CCMB problem is a decision making process that maps the historical observations $\{\mathcal{X}^1, \mathcal{A}^1, \mathcal{P}^1; \mathcal{X}^2, \mathcal{A}^2, \mathcal{P}^2; ...; \mathcal{X}^{t-1}, \mathcal{A}^{t-1}, \mathcal{P}^{t-1}\}$ and the current context $\mathcal{X}^t$ to an action $\mathcal{A}^t \in \mathcal{A}$. The objective of the CCMB problem is to maximize the expected total payoff for a given resource budget constraint as follows:

$$\begin{aligned} &\underset{\mathcal{A}^t}{\operatorname{argmax}} \sum_{t=1}^{T} P^t, 1 \le t \le T \text{ (payoff maximization)} \\ &\text{s.t.:} \sum_{t=1}^{T} \mathcal{C}^t \le B, 1 \le t \le T \text{ (budget constraint)} \end{aligned} \tag{4}$$

This objective function can be solved using the adaptive linear programming approach in [40]. The detailed discussion of training process of IPD is discussed in Section V.

C. Crowd Quality Control

A key challenge of the crowdsourcing platform is that the quality of the answers vary and some workers can provide wrong answers due to their limited knowledge or subjective opinions. In fact, our pilot studies show the average labeling accuracy of the crowd workers is not perfect (i.e., around 80% in Figure 6). Several existing solutions are developed to address this issue. For example, majority voting is a common technique (**Voting**) where the the aggregated result is simply the one returned by the majority of the workers. This approach is known to be suboptimal when workers have different reliability [41]–[43]. More principled approaches such as truth discovery (**TD-EM**) [29] is able to jointly derive the truthful label of the queries as well as the reliability of the workers. However, this technique does not work well when the number of responses per worker is low [44]. Another commonly used technique is worker quality filtering (**Filtering**) [13], which blacklists the workers with a record of poor labeling quality. However, this approach may fail when the workers are new to the platform and do not have sufficient labeling history. There also exist some expertise-aware worker assignment schemes [38], [45]–[47] that directly assign queries to workers with high quality. However, they assume the application has full control of the worker pool [48], which does not apply to the black-box crowdsourcing platform we study.

In light of the knowledge gap of existing crowd quality control schemes, we devise a new idea: we not only ask the crowd to provide direct labels of data samples, but also provide their evidence. The evidence is captured by a set of questionnaires (Figure 3). For example, in the DDA application, we ask the workers to answer "Is the image photoshopped (i.e., fake image)?", "Does this image show a damage of road?". Note that we use the format of fixed-form questionnaire rather than free-form input (e.g., ask the worker to describe the image) to eliminate the challenge of parsing natural language. The questionnaire collected a set of extra features that can help derive the truthful labels of the images.

Given the labels and features provided by the workers, we train a supervised classifier that takes both the labels and the questionnaire answers of a query as inputs and outputs the truthful label of the image. We choose the state-of-art gradient boosting model (XGBoost) [49] as our classifier. The combination of labels and questionnaire answers allows CQC to achieve at least 5.75% higher accuracy than existing approaches (shown in Table I). The accurate truthful labels generated by CQC module provide us with a good basis to evaluate and calibrate the AI algorithms. We elaborate the details of calibration process next.

D. Machine Intelligence Calibration (MIC)

The MIC module is designed to calibrate and improve the AI algorithms based on the labels provided by the crowd workers. The MIC module includes three complementary

Table I: Aggregated Label Accuracy

	Morning	Afternoon	Evening	Midnight	Overall
CQC	**0.93**	**0.92**	**0.94**	**0.94**	**0.9350**
Voting	0.82	0.83	0.85	0.87	0.8425
TD-EM	0.86	0.85	0.85	0.89	0.8625
Filtering	0.84	0.86	0.88	0.90	0.8775

calibration strategies that are performed simultaneously right after the execution of CQC module within each sensing cycle.

Dynamic Expert Weights Update Strategy: Recall that a set of AI algorithms form a committee in QSS module to collectively decide the classification result of a data sample and each algorithm is assigned an expert weight. The expert weight is crucial in determining the performance of the AI algorithms. We design a dynamic expert score update strategy that can learn the performance of each expert in the committee as the feedback is collected from the crowd. The proposed strategy builds a feedback control process using the crowd feedback as the control signal. In particular, for each AI algorithm AI_m, we compute a loss function based on the discrepancy of its classification result and the truthful label from the crowd as:

$$\mathcal{L}_m^t = \sum^{i \in \mathcal{Q}^t} 1 - \delta(KL^{sym}(\mathcal{D}(AI_{m,i}^t), \mathcal{D}(TL_i^t)) \quad (5)$$

where $\mathcal{Q}^t$ denotes the set of images chosen by QSS for MTurks at the t^{th} sensing cycle. $\mathcal{D}(TL_i^t)$ is the probabilistic distribution of the labels obtained from CQC module. $KL^{sym}(\mathcal{D}(AI_{m,i}^t), \mathcal{D}(TL_i^t))$ is the symmetric KL-Divergence between the two label distributions. δ is a normalization process to map the divergence to a [0,1] scale. Intuitively, the more different that the output from the AI algorithm is from the truthful label from the crowd, the higher the loss is. Given the loss function, we dynamically update the expert weights of the AI algorithms at each sensing cycle using a classical exponential weight update rule [50]. The updated weights reflects the reliability of each expert at the current sensing cycle. We use the updated weights to derive the final labels of the input images as discussed in Section IV-A.

Model Retraining and Crowd Offloading Strategies: The model retraining strategy is to address the failure case of AI algorithms that is caused by insufficient training samples. Similar to existing hybrid AI-human frameworks [13], we use the truthful labels provided by the crowdsourcing platform to retrain the AI models for the next sensing cycle. The crowd offloading strategy is implemented to tackle the cases where the AI algorithms may have innate flaws (e.g., failure to handle fake images in DDA applications). In this strategy, the truthful labels derived from the CQC is used to directly replace the classification labels of the query set from QSS in the current sensing cycle. The query set contains two categories of images that AI potentially fails: 1) the images that the AI algorithms in the committee do not agree with each other on their labels (captured by the committee entropy) ; and 2) the images that the AI algorithms happen to make the same wrong decision (captured by the $\epsilon - greedy$ strategy). By replacing both categories of images with human labels, the crowd offloading strategy not only prevents the AI from giving uncertain classification labels but also addresses the failure case when AI algorithms make a common mistake.

V. Evaluation

In this section, we present an extensive evaluation of our CrowdLearn scheme. We first discuss the evaluation setup and baselines for comparison. We then present the evaluation results using a real-world deep learning-based damage assessment (DDA) application in the aftermath of a disaster event. The results show that CrowdLearn achieves significant performance gains in terms of classification accuracy and crowd delay compared to the state-of-the-art baselines.

A. Baselines

We choose the following state-of-the-art DDA schemes and hybrid human-AI solutions as our baselines. For the QSS module in CrowdLearn, we use VGG16, BoVW, and DDM as the committee.

- **VGG16**: A DDA scheme that uses deep Convolutional Neural Networks (CNN) [6].
- **BoVW**: A DDA scheme that uses handcrafted features (e.g., scale invariant feature transform, histogram of oriented gradients) to train a neural network classifier [51].
- **DDM**: A DDA scheme that combines CNN and Gradient-weighted Class Activation Mapping (Grad-CAM) to produce a damage heatmap of a given image, which is used to derive the damage severity [5].
- **Ensemble**: An aggregation of the above algorithms (VGG16, BoVW, DDM) using a boosting technique [52].
- **Hybrid-Para**: a human-AI hybrid system where humans and AI independently label the images and their results are integrated using a complexity index [53].
- **Hybrid-AL**: a crowdsourcing-based active learning framework for AI algorithms where the annotated labels collected from MTurk are used to re-train the AI algorithm for the performance improvement [13].

B. Experiment Setup

We use a total of 960 social media images with golden ground truth labels about the Ecuador Earthquake in 2016 from Instagram and Twitter [6]. In our experiments, the dataset is split into a *training set* and a *test set*. The training set is used to 1) perform the pilot study to characterize the black-box MTurk platform; 2) train the reinforcement learning-based IPD module as described in Section IV-B; and 3) train the AI-based DDA algorithms. The training set contains a total of 560 images ad the test set has a total of 400 images that emulates unseen data dynamically generated during each sensing cycle.

We run the application over 40 sensing cycles during 4 different temporal contexts (i.e., morning, afternoon, evening, midnight) - 10 cycles for each temporal context. Each sensing cycle lasts 10 minutes and has a set of 10 images from the test set. The input and output to CrowdLearn and all baselines

schemes are the same: each scheme takes an image as input and output severity level of the image (including *severe*, *moderate* and *no damage*). Note that Hybrid-Para, Hybrid-AL and CrowdLearn are different from other baselines in the sense that they all leverage humans from MTurk. To that end, we allow the three hybrid schemes to query the same amount of images to MTurk (i.e., 5 images per sensing cycle).

C. Performance Evaluation

1) Classification Accuracy: In the first set of experiments, we focus on the overall performance of all schemes in terms of classification accuracy, which is evaluated using the classic metrics for multi-class classification: *Accuracy*, *Precision*, *Recall* and *F1-Score*. Similar to [5], [6], these scores are *macro-averaged* since the dataset has balanced class labels.

Table II: Classification Accuracy for All Schemes

Algorithms	Accuracy	Precision	Recall	F1
CrowdLearn	**0.877**	**0.904**	**0.885**	**0.894**
VGG16	0.770	0.845	0.744	0.791
BoVW	0.670	0.707	0.744	0.725
DDM	0.807	0.891	0.765	0.823
Ensemble	0.815	0.892	0.778	0.831
Hybrid-Para	0.797	0.849	0.795	0.821
Hybrid-AL	0.823	0.883	0.803	0.841

The results are reported in Table II. We observe the CrowdLearn consistently outperforms other baselines. In particular, CrowdLearn achieved 5.3% improvement on F-1 Score compared to best-performing baseline (i.e., Hybrid - AL). The reason is that the CrowdLearn can effectively integrate human intelligence into the DDA algorithm. In particular, the MIC module leverages human intelligence to improve the results by fine tuning the expert weights of candidate AI algorithms to outperform the AI-only schemes. Compared to other hybrid human-AI systems, CrowdLearn actively troubleshoots and eventually fixes the failure scenarios of AI algorithms. In contrast, Hybrid-parallel only takes the crowd as a source of annotations and does not directly interact with the AI algorithm. Hybrid-AL only leverages crowd annotations for retraining and cannot address the innate failure mode of the AI algorithm. We further plot the ROC curves of all schemes in Figure 7. We observe that CrowdLearn continues to outperform other baselines when we tune the classification thresholds.

2) Delay Analysis: Next, we evaluate the delay of all compared schemes in terms of 1) execution time, and 2) delay of query answered by the crowdsourcing platform. In a practical setting, the complete life cycle of the DDA application should include both of these delays. The experiment is conducted on a PC with Nvidia RTX 2070 GPU and Intel i7-8700K 6-core CPU and 16G of RAM. The average delay of algorithm execution time and crowd delay of all schemes are listed in Table III. We observe that the execution delay of CrowdLearn

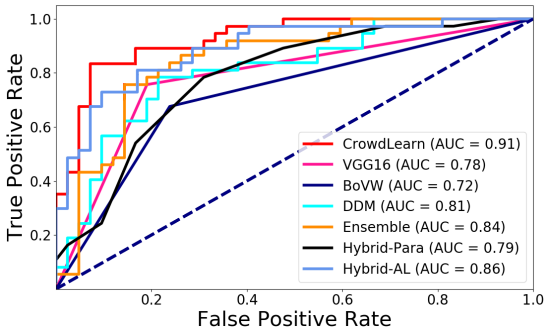


Figure 7: Macro-average ROC Curves for All Schemes

is higher than AI-only schemes (i.e., VGG16, BoVW and DDM) because CrowdLearn incorporates all three AI-only algorithms as its committee and runs extra modules to generate incentives and perform quality control to the crowd. We also observe that that response delay from the crowdsourcing platform is the major contributor to the overall delay of human-AI hybrid systems including CrowdLearn. This observation further demonstrates the importance of designing an effective incentive policy to minimize the response delay from the crowd and provide timely response to the application. The results show that CrowdLearn scheme significantly reduces the crowd delay by 35% compared to Hybrid-Para and Hybrid-AL that both adopt a fixed incentive policy. We attribute such a performance gain to our novel IPD module that leverages a context-aware reinforcement learning scheme to dynamically identify the optimal incentive strategy to reduce the response delay from the crowd.

Table III: Average Delay (in Seconds) per Sensing Cycle

Algorithms	Algorithm Delay	Crowd Delay
CrowdLearn	**55.62**	**342.77**
VGG16	47.83	N/A
BoVW	37.55	N/A
DDM	52.57	N/A
Ensemble	85.82	N/A
Hybrid-Para	94.28	588.75
Hybrid-AL	53.54	527.61

To further examine the crowd response delay, we show the delay across different temporal contexts. In addition to the fixed incentive policy adopted by Hybrid-Para and Hybrid-AL, we also compare to another heuristic baseline where the incentives are randomly assigned. For the fixed incentive strategy, we use the maximum incentive for each query (i.e., the total budget divided by the number of queries). The results are shown in Figure 8. We observe that the IPD module in CrowdLearn achieves the lowest delay with the least variations across different contexts compared to both fixed and random incentive mechanisms. This is because CrowdLearn can adjust its incentives based on how responsive the crowd is. For example, if the crowd is less responsive (e.g., in the morning),

CrowdLearn would provide higher incentives to stimulate timely responses. On the other hand, CrowdLearn would decrease the incentives when the crowd is very proactive (e.g., in the evening). The results show that CrowdLearn is robust against the change of contexts and provides consistently faster responses from the crowd than alternative strategies.

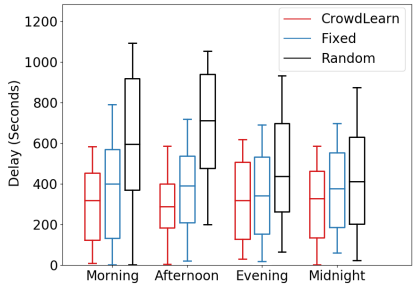


Figure 8: Crowd Delay at Different Temporal Contexts

3) Impact of Human Intelligence: A key parameter in our problem setting is the size of the query set (i.e., the amount of images that are sent to the MTurk for query). We tune the size of the query set from 0 percent of the images at each sensing cycle (AI only) to 100 percent (crowd only) to examine its effect on the classification performance. We only compare CrowdLearn with hybrid human-AI baselines that include the crowd component and the best-performing AI-only base line (i.e., Ensemble) as as a reference point. The results are shown in Figure 9. We observe that the performance gain of CrowdLearn compared to the baselines increases as we increase the size of the query set, which demonstrates the benefit of incorporating human intelligence into the AI algorithms. Interestingly, we note the the performance of other hybrid human-AI systems (i.e., Hybrid-AL and Hybrid-Para) are rather stable even with the increase of the number of queries to the crowd. We attribute this observation to the fact that these baselines did not really fix the innate problem of AI as we discussed in the classification accuracy section. We also observe that the performance of CrowdLearn degrades to Ensemble when there is no HI (i.e., 0% query set). However, our scheme still outperforms Hybrid-Para and Hybrid-AL when there is no AI (i.e., 100% query set). This is because the CQC component in CrowdLearn can provide much more accurate human annotations than the two baselines that simply use majority voting for the quality control.

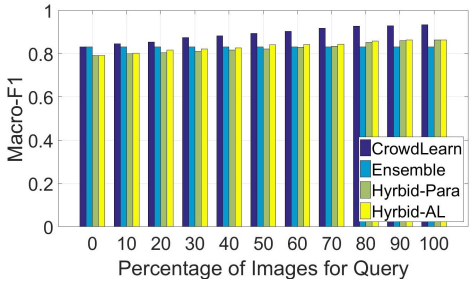


Figure 9: Size of Query Set vs. Classification Performance

4) Impact of Budget: In our last set of experiment, we study the impact of the resource budget on the classification accuracy and delay of CrowdLearn. We tune the total budget from 2 USD (1 cent per task on average) to 40 USD (20 cents per task on average). The classification accuracy and delay are reported in Figure 10 and Figure 11 respectively. We observe that that the classification performance of CrowdLearn is worse with low incentives as compared to higher ones. However, the performance becomes stable as long as a reasonable budget is provided (e.g., above 6 USD or 3 cents per task on average). For example, the F1 score only increases by 0.18 from the budget of 8 USD to the budget of 40 USD. We observe similar impact of budget on the crowd response delay. The above results show that the CQC scheme and IPD schemes in CrowdLearn are robust to the changes of the budget and can consistently perform well with reasonable budgets.

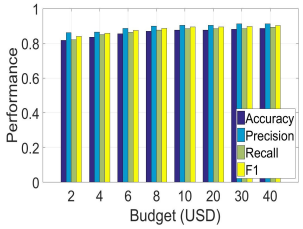


Figure 10: Budget vs. F1

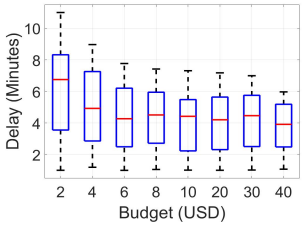


Figure 11: Budget vs. Delay

VI. Conclusion and Future Work

This paper presents the CrowdLearn framework to addresses fundamental challenges in melding *black-box AI* and *black-box crowdsourcing platform* in boosting the performance of deep learning-based DDA applications. The CrowdLearn framework leverages the crowd intelligence to troubleshoot, calibrate and improve the AI performance in DDA. CrowdLearn also designs a new incentive policy and quality control scheme to ensure timely and high quality responses from the crowd. Evaluation results on a real-world DDA application and MTurk show that CrowdLearn significantly outperforms existing AI-only baselines and state-of-the-art human-AI frameworks.

We envision CrowdLearn is a general crowd-AI hybrid approach that can be extended to applications beyond DDA (e.g., object recognition [54], autonomous driving [55], and event detection [56]). It will be interesting to explore the unique failure scenarios of AI algorithms in these applications and investigate how the crowd can help boost their performance.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CNS-1831669, CBET-1637251, CNS-1566465 and IIS-1447795, Army Research Office under Grant W911NF-17-1-0409, Google 2017 Faculty Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[2] T. J. Bench-Capon and P. E. Dunne, "Argumentation in artificial intelligence," *Artificial intelligence*, vol. 171, no. 10-15, pp. 619–641, 2007.

[3] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[4] Y. Zhang, Y. Lu, D. Zhang, L. Shang, and D. Wang, "Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1544–1553.

[5] X. Li, D. Caragea, H. Zhang, and M. Imran, "Localizing and quantifying damage in social media images," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 194–201.

[6] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 569–576.

[7] Y. Zhang, D. Zhang, N. Vance, and D. Wang, "Optimizing online task allocation for multi-attribute social sensing," in *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2018, pp. 1–9.

[8] D. Y. Zhang, L. Shang, B. Geng, S. Lai, K. Li, H. Zhu, M. T. Amin, and D. Wang, "Fauxbuster: A content-free fauxtography detector using social media comments," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 891–900.

[9] A. Holzinger, M. Plass, K. Holzinger, G. C. Crişan, C.-M. Pintea, and V. Palade, "Towards interactive machine learning (iml): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach," in *International Conference on Availability, Reliability, and Security*. Springer, 2016, pp. 81–95.

[10] B. Nushi, E. Kamar, E. Horvitz, and D. Kossmann, "On human intellect and machine failures: Troubleshooting integrative machine learning systems." in *AAAI*, 2017, pp. 1017–1025.

[11] B. Nushi, E. Kamar, and E. Horvitz, "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure," *arXiv preprint arXiv:1809.07424*, 2018.

[12] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *European Conference on Computer Vision*. Springer, 2010, pp. 438–451.

[13] F. Laws, C. Scheible, and H. Schütze, "Active learning with amazon mechanical turk," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1546–1556.

[14] Y. Wang, X. Jia, Q. Jin, and J. Ma, "Quacentive: a quality-aware incentive mechanism in mobile crowdsourced sensing (mcs)," *The Journal of Supercomputing*, vol. 72, no. 8, pp. 2924–2941, 2016.

[15] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. Chiang, "Quality-aware collaborative question answering: methods and evaluation," in *Proceedings of the second ACM international conference on web search and data mining*. ACM, 2009, pp. 142–151.

[16] W. Dai, Y. Wang, Q. Jin, and J. Ma, "Geo-qti: A quality aware truthful incentive mechanism for cyber–physical enabled geographic crowdsensing," *Future Generation Computer Systems*, vol. 79, pp. 447–459, 2018.

[17] E. Mitsopoulou, I. Boutsis, V. Kalogeraki, and J. Y. Yu, "A cost-aware incentive mechanism in mobile crowdsourcing systems," in *2018 19th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2018, pp. 239–244.

[18] D. Zhang, Y. Ma, Y. Zhang, S. Lin, X. S. Hu, and D. Wang, "A real-time and non-cooperative task allocation framework for social sensing applications in edge computing systems," in *2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2018, pp. 316–326.

[19] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–8.

[20] J. Marshall and D. Wang, "Mood-sensitive truth discovery for reliable recommendation systems in social sensing," in *Proceedings of International Conference on Recommender Systems (Recsys)*. ACM, 2016, pp. 167–174.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[22] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[23] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *Journal of Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.

[24] D. Y. Zhang, L. Song, Q. Li, Y. Zhang, and D. Wang, "Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 901–910.

[25] J. Shu, S. Kosta, R. Zheng, and P. Hui, "Talk2me: A framework for device–to–device augmented reality social network," 2018.

[26] A. Holzinger, M. Plass, K. Holzinger, G. C. Crisan, C.-M. Pintea, and V. Palade, "A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop," *arXiv preprint arXiv:1708.01104*, 2017.

[27] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Aaai*, vol. 6. Boston, MA, 2006, pp. 1000–1005.

[28] V. Ambati, S. Vogel, and J. G. Carbonell, "Active learning and crowd-sourcing for machine translation." in *LREC*, vol. 1. Citeseer, 2010, p. 2.

[29] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, Apr. 2012, pp. 233–244.

[30] D. Y. Zhang, J. Badilla, Y. Zhang, and D. Wang, "Towards reliable missing truth discovery in online social media sensing applications," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 143–150.

[31] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[32] D. Y. Zhang and D. Wang, "Heterogeneous social sensing edge computing system for deep learning based disaster response: demo abstract," in *Proceedings of the International Conference on Internet of Things Design and Implementation*. ACM, 2019, pp. 269–270.

[33] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[34] J. Doshi, S. Basu, and G. Pang, "From satellite imagery to disaster insights," *arXiv preprint arXiv:1812.07033*, 2018.

[35] X. Chen, E. Santos-Neto, and M. Ripeanu, "Crowdsourcing for on-street smart parking," in *Proceedings of the second ACM international symposium on Design and analysis of intelligent vehicular networks and applications*. ACM, 2012, pp. 1–8.

[36] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 287–294.

[37] M. Tokic and G. Palm, "Value-difference based exploration: adaptive control between epsilon-greedy and softmax," in *Annual Conference on Artificial Intelligence*. Springer, 2011, pp. 335–346.

[38] X. Zhang, Y. Wu, L. Huang, H. Ji, and G. Cao, "Expertise-aware truth analysis and task allocation in mobile crowdsourcing," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 922–932.

[39] E. A. Gehan, "A generalized wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, no. 1-2, pp. 203–224, 1965.

[40] H. Wu, R. Srikant, X. Liu, and C. Jiang, "Algorithms with logarithmic or sublinear regret for constrained contextual bandits," in *Advances in Neural Information Processing Systems*, 2015, pp. 433–441.

[41] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[42] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Transactions on Big Data*, 2018.

[43] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *The 33rd International Conference on Distributed Computing Systems (ICDCS'13)*, July 2013.

[44] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 1076–1081.

[45] P. Welinder and P. Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 25–32.

[46] C. Harris, "You're hired! an examination of crowdsourcing incentive models in human resource tasks," in *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. Hong Kong, China, 2011, pp. 15–18.

[47] J. Wang, J. Tang, D. Yang, E. Wang, and G. Xue, "Quality-aware and fine-grained incentive mechanisms for mobile crowdsensing," in *Distributed Computing Systems (ICDCS), 2016 IEEE 36th International Conference on*. IEEE, 2016, pp. 354–363.

[48] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2008, pp. 453–456.

[49] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[50] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[51] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[52] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine learning*, vol. 37, no. 3, pp. 297–336, 1999.

[53] J. Jarrett, I. Saleh, M. B. Blake, R. Malcolm, S. Thorpe, and T. Grandison, "Combining human and machine computing elements for analysis via crowdsourcing," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on*. IEEE, 2014, pp. 312–321.

[54] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[55] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.

[56] X. Feng, B. Qin, and T. Liu, "A language-independent neural network for event detection," *Science China Information Sciences*, vol. 61, no. 9, p. 092106, 2018.