Integrating Crowd and Machine Learning in an Intelligent Interface: A Case Study of Oil Spill Detection in Satellite Images

Rifat Mehreen Amin
rifat.amin@ifi.lmu.de
LMU Munich
Munich, Bavaria, Germany

Feng Chen
feng.chen@campus.lmu.de
LMU Munich
Munich, Bavaria, Germany

Linda Hirsch
linda.hirsch@ifi.lmu.de
LMU Munich
Munich, Bavaria, Germany

Changkun Ou
research@changkun.de
LMU Munich
Munich, Bavaria, Germany

Tran-Vu La
tran-vu.la@list.lu
Luxembourg Institute of Science and Technology
Esch-sur-Alzette, Luxembourg

Andreas Butz
andreas.butz@ifi.lmu.de
LMU Munich
Munich, Bavaria, Germany

ABSTRACT

Object detection tasks still often require manual image analysis. Using Machine Learning (ML) instead creates accountability challenges, necessitating experts for model refinement, which is costly and takes time. We investigate integrating crowd knowledge as a cost-effective alternative. While human capabilities in recognizing complex patterns and perceiving variations can still outperform machines and improve an imperfect ML model, ML predictions can compensate for the crowd's lack of expertise. Our investigation (N=28 non-expert) in oil spill detection shows that adopting an ML-assisted UI elevates precision and recall by over 11% and increases efficiency by 29% compared to a non-assisted UI. Considering agreement among non-expert crowd workers further improved precision by 8% and recall by almost 5%, which is also substantially beyond pure ML performance. Our work contributes an approach for combining crowd knowledge and ML to advance human-AI collaboration in oil spill detection.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Environmental sciences**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

machine learning, human-AI collaboration, human-AI interaction design, intelligent interfaces, object detection

ACM Reference Format:
Rifat Mehreen Amin, Feng Chen, Linda Hirsch, Changkun Ou, Tran-Vu La, and Andreas Butz. 2024. Integrating Crowd and Machine Learning in an Intelligent Interface: A Case Study of Oil Spill Detection in Satellite Images. In *Proceedings of ACM AVI 2024 (AVI 2024).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

AVI 2024, June 3–7, 2024, Arenzano, Genoa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Oil spills occurring offshore and along coasts are considered hazards to wildlife, marine ecosystems, and coastal environments [5]. In recent years, the effectiveness of oil spill response operations, which involve efforts to decontaminate affected areas, has been enhanced due to a dense network of Earth Observation (EO) satellites [3]. The data acquired by many satellite sensors enable quick identification and delineation of the areas affected by oil slicks [41]. To process the significant database of satellite data, ML models, specifically object detection models, have been widely used in recent years [20, 42]. These models assist humans in diverse risk detection, enabling them to take quick counteractions against economic and ecological consequences [20].

However, current object detection models are often unreliable [35]. Overcoming the limitations of these models requires additional (human) expert involvement for model refinement, which is expensive and time-intensive [10]. Additionally, in some domains, experts are difficult to find [21]. Another approach is engaging crowd workers for manual labeling, creating, e.g., labeled datasets for ML models' supervised learning [9]. Such a crowd – here referred to as an untrained (non-expert) group of people – has proven to comprise valuable contributors in numerous fields, such as surveillance and data gathering [32] and is easy to recruit through online platforms. Depending on the area of expertise, crowd workers have also proven to be well-suited substitutes for experts in specialized fields (e.g., IT tasks) [22]. Further, including non-expert crowd workers as collaborators offer a potentially more scalable and cost-effective solution than including experts. The knowledge gathered from these crowd workers is referred to as crowd knowledge (CK). According to Blesik et al. [4]: "*Crowd knowledge is a collaborative aggregation of context-dependent information contributed and used by participants that is stored in an artifact and provided to fulfill a purpose.*"

Building on the advantages of engaging CK, our work explores a novel approach that fosters crowd collaboration and an ML object detection model for oil spills in satellite images. In a within-subject study, we convey this approach with 28 non-experts as crowd workers, comparing an ML-assisted with a non-assisted interface. In more detail, we assessed precision, recall, and efficiency for each approach using a state-of-the-art object detection model. Our work is guided by the following research questions:

RQ1 *Does the collaboration between the crowd and ML result in more precise and efficient identification of oil spills in satellite images than the crowd's manual detection?*

RQ2 *How does the crowd complement the oil spill detection results of ML?*

We found that the crowd-ML collaboration substantially improves precision and recall in the oil spill detection tasks compared to non-assisted crowd knowledge and increases the overall accuracy with crowd consensus. The work emphasizes the potential of crowd-ML collaboration to refine object detection models as an alternative approach to expert involvement and highlights its challenges, such as e.g., generating consistent crowd engagement. With this, our study contributes to the field of environmental science and human-AI collaboration with an intelligent interface that employs an imperfect ML model to assist the crowd in evaluating oil spill detection in satellite images, enabling a faster and more precise human-AI collaborative labeling process.

2 RELATED WORK

This section provides an overview of CK applications related to machine learning and insights into oil spill detection algorithms and intelligent interfaces for human-AI collaboration.

2.1 Crowd Knowledge in Machine Learning

Through collective intelligence and aggregated information, CK offers an alternative approach that complements and sometimes even surpasses the insights provided by individual experts [36]. It has been used in citizen science projects such as Galaxy Zoo [16] to classify galaxies by observing their shapes from sample images. Another example is the eBird citizen-science project, where bird enthusiasts worldwide report their bird observations [34]. In both cases, the collective information surpasses the expertise of individual experts since the vast numbers of participants provide broader coverage of observation, diverse perspectives, and larger sample sizes. CK has also been applied in ML and can, according to Wang et al. [36], be included in all three stages of a standard ML process: 1) data preparation [18], 2) feature discovery and learning [44], and 3) model assessment and refinement [15]. Consequently, a collaborative approach combining CK with ML can complement an imperfect ML model and compensate for the crowd's lack of expertise.

2.2 Deep Learning based Oil Spill Detection

Deep learning (DL) based object detection supports identifying and localizing objects of interest in digital images or videos [45]. Some object detection algorithms are CNN, Fast RCNN [17], Faster R-CNN [31], Mask R-CNN [19], and YOLO [30]. These detection models have been researched in many applications contexts, including medical image processing for, e.g., cancerous tumor detection [25], industrial activities for anomaly or fault detection [1]. In maritime surveillance, DL techniques are applied to detect ships and oil spills [20]. Liu et al. [27] utilizes a CNN algorithm for ship detection whereas Nieto-Hidalgo et al. [29] utilizes CNN for SAR images to identify oil spills. Nieto-Hidalgo et al. [29] pointed out the feasibility of detecting oil spills with object detection algorithms. Emna et al. [13] and Yekeen and Balogun [40] employed Mask-Region-Based Convolutional Neural Network (Mask-RCNN) for detecting oil slicks. Mask-RCNN combines object detection and semantic segmentation. However, for the development of a near real-time (NRT) oil spill detection system, the focus shifts to highly efficient one-stage object detection algorithms, such as You Only Look Once (YOLO) [39]. Considering these findings, we apply the latest YOLOv8[1] and explore its efficiency compared to being enhanced through CK.

2.3 Intelligent Interfaces for Human-AI collaboration

ML models, such as deep learning models, are used for aiding humans in annotations; tools like CVAT [11] and Annotator [6] offer the ability to generate annotations using pre-trained models automatically. Google Fluid [2] aims to achieve full image annotation in a single pass instead of breaking it down into a series of microtasks, such as identifying object presence or drawing polygons or boxes around objects [18]. Moreover, significant efforts have been directed towards developing intelligent interfaces supporting clinicians when working with imperfect ML models in decision-making [7, 25]. Further, for annotating audio-visual data using a deep learning model, Zhang et al. [43] propose a collaborative tool named Peanut. Peanut's human-AI collaborative pipeline separates the multi-modal task into two single-modal tasks. These advancements highlight the importance of intelligent interfaces, emphasizing their crucial role in enhancing collaboration and decision-making.

Drawing from previous work, we aim to investigate the collaboration between a crowd and an ML model in the context of oil spill detection. Specifically, we hypothesized that:

H1: *The collaboration between the crowd and ML enhances the performance efficiency in detecting oil spills in satellite images compared to their individual performance.*

H2: *Crowd workers can identify errors of the object detection model*, such as missed oil spills (*False Negatives*) or wrong detections of oil spills (*False Positives*).

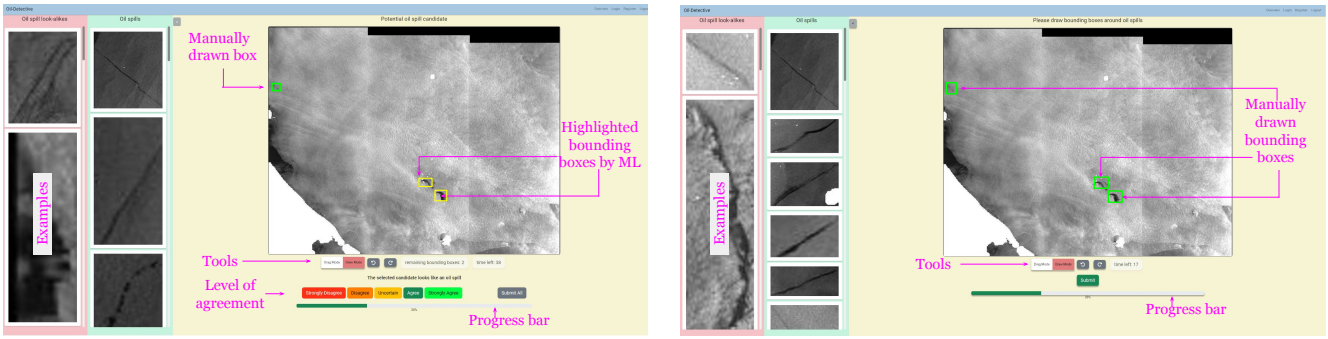
3 SYSTEM DESIGN

To test our hypotheses, we developed two web interfaces for oil spill detection, one with and one without ML assistance, using a similar User Interface (UI) (see Figure 1).

3.1 Object Detection Model Development

We employed the YOLOv8 object detection model for oil spill detection since it has the most balanced tradeoff between detection accuracy and speed. For efficient computation and memory management, we chose a batch size of 8, which allows the model to process several images simultaneously during each training iteration. In this paper, we used 1445 Sentinel 1 Band A/B (S1-A/B) labeled images. These images cover the Gulf of Mexico, the Indian Ocean, and the East and South China Set between January 2015 and May 2021 [20]. These images were divided into two groups of 80% (1156) and 20% (289). The first set was used for training, and the second was used for testing. The validation IoU (Intersection over Union) threshold for the model was set to 0.5 to merge highly overlapping bounding boxes and reduce redundancy in the output

[1] https://github.com/ultralytics/ultralytics, last accessed April 4, 2024



(a) ML-assisted user interface **(b) Non-assisted user interface**

Figure 1: Interfaces of Study Conditions: a) the ML-assisted interface, including ML oil spill suggestions and agreement scale in the form of buttons. b) the non-assisted GUI.

detections. In the trained detection model, the F1-score peaks at 0.72 within a confidence range of 0.3 to 0.4. This threshold achieves a balanced performance with high Precision (0.78) and Recall (0.67). This precision implies that 78% of the positive predictions are correct, reducing false positives (FP). Recall signifies that the model identifies 67% of actual positives, minimizing false negatives (FN).

3.2 Back End

We developed a Node.js server for our study and used a MongoDB database to store users' personal attributes, satellite image URLs, ground truth information, and predictions from the ML model. We used Cloudinary[2] for storing the SAR satellite images. The web application for both interface conditions runs on the server, where the ML predictions are executed. Figure 2 shows the system architecture and approach for both experiment conditions.

3.3 Front End and Interaction Concept

Both UIs share a common set of features while introducing elements specific to their assistance mechanisms. In both cases, examples and counterexamples (lookalikes) of oil spills are presented on the left in a scrollable list (see Figure 1). In total, 100 randomly selected images, each for oil spills and lookalikes, serve as reference points for the crowd to identify potential oil spills. Both UIs offer two distinct modes: **Drag Mode** for image navigation and **Draw Mode** to add bounding boxes for object detection by the crowd. These drawn bounding boxes are fully reversible via the **undo** button. In case of an inadvertent reversal, the **redo** button aids in recovering deleted bounding boxes. In the ML-assisted UI, yellow dotted bounding boxes represent suggestions made by the ML model. Further, users must indicate the agreement level with the ML suggestions on a 5-point Likert scale from 1:Strongly Disagree to 5:Strongly Agree before moving to the next image. Both UIs incorporate a 60-second timer, encouraging but not limiting users to complete the detection task for each displayed satellite image within the specified timeframe.

[2]https://cloudinary.com/, last accessed April 4, 2024

4 STUDY DESIGN AND APPROACH

We conducted a within-subject online study on Zoom, which was approved by the university's ethics board. The study involved 28 participants who compared both oil spill detection interfaces developed by us described in subsection 3.3.

4.1 Dataset Preparation

We used 40 SAR images chosen from the ML test set for the study. One of the authors, an expert in the field of oil spill detection for over five years, categorized the images into three distinct sets based on their difficulty in detecting oil spills, considering factors such as the presence of oil spill lookalikes, the level of noise in the image, and other environmental complexities. Subsequently, a subset of images was randomly selected for the main user study, consisting of 20 easy images, 10 medium images, and 10 hard images.

4.2 Experimental Setup

4.2.1 Independent Variables. Our independent variable is the machine learning assistance with two levels: *ML-assisted* crowd detection and *non-assisted* crowd detection.

4.2.2 Dependent Variables. As dependent variables, we measured precision and recall in detecting oil spills and efficiency following prior work on object detection [18, 23]. We also assessed the user experience using the Technology-Acceptance Model version 4 according to Lewis [24] to identify potential interface design impacts.

4.2.3 Supplementary Open-ended and Single Choice Questions. We added open-ended questions about whether participants had experienced specific challenges or difficulties during task completion and whether they could imagine continuing the collaboration for oil spill detections. Furthermore, we collected suggestions for improvement and motivational factors for further usage. Additionally, we added three single-choice questions about which interface participants would choose regarding time efficiency, facilitating accurate judgment and for future oil spill detections.

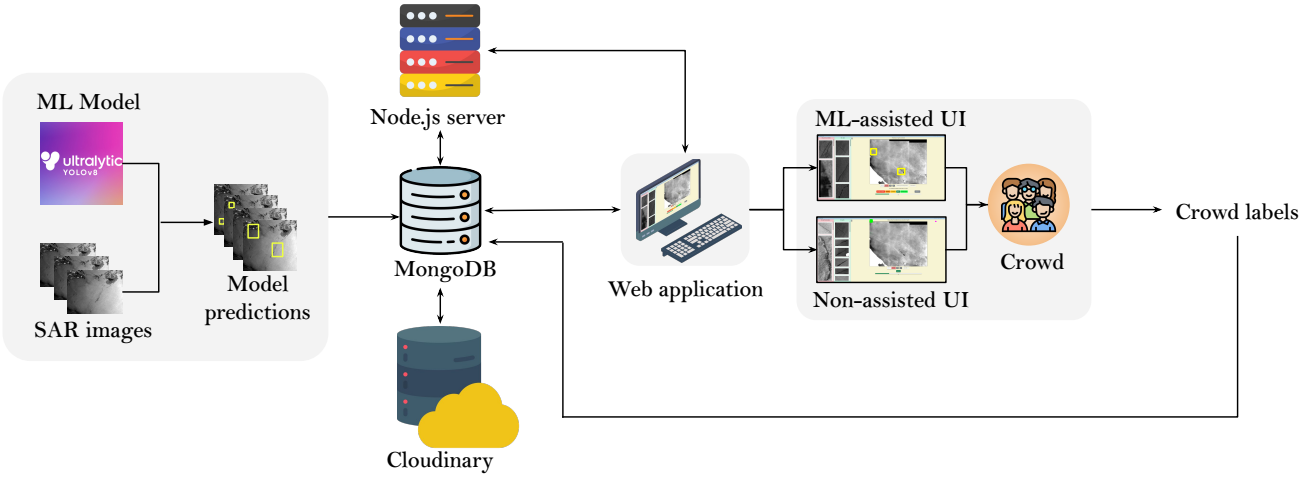


Figure 2: Oil spill detection approach and system architecture: both UIs ran on the same server and were presented using a web application to participants. Depending on the interface, crowd labels were stored in a separate model in the MongoDB database.

4.2.4 Study Approach. We tested our study design in a pilot study with five participants. In the pilot study, participants labeled images containing up to 50 oil spills, which resulted in prolonged completion time fatiguing participants. Thus, only images containing fewer than five oil spills were considered further to balance dataset diversity and study efficiency. The main study comprised 40 images and took approximately 60 minutes. Before each task, participants were given instructions, task descriptions, and expert tips and completed a trial run with each interface to reduce novelty effects. In the main study, participants were evenly distributed to either Group 1 or Group 2. Each participant evaluated all 40 SAR images, comprising 10 easy, five medium, and five hard images in each condition, with a predetermined order (easy to hard) to account for a learning effect in the process [8]. Group 1 started with ML-assisted labeling, evaluating 20 images, then non-assisted labeling with another 20 images, while Group 2 began with non-assisted labeling and then ML-assisted labeling. Both groups encountered the images in an identical sequence. This guaranteed that both groups assessed the same images in the same sequence but within distinct tasks, enabling a fair comparison of the results between the two approaches. Moreover, it ensured that each image received 14 evaluations from the non-assisted labeling task and 14 evaluations from the ML-assisted task, enabling a comprehensive comparison between these methodologies for each image. In the ML-assisted condition, we asked participants to rate the system's confidence level of correctly labeled oil spills and mark additional oil spills. In the non-assisted condition, participants marked all oil spills. They also had the option to submit images without drawing bounding boxes around any oil spills in either condition. We integrated a soft timer in the UI set to 60 seconds for each image to manage study duration. However, participants received extra time if needed. After completing a condition, we asked participants to fill out a questionnaire.

4.3 Participants

We recruited 28 participants between the ages of 18 and 44 years (M = 27.73, SD= 4.38) using convenience sampling, of which 23 self-identified as male and five as female. They gave informed consent and participated voluntarily without compensation. None of the participants had any experience in oil spill detection.

5 RESULTS

We applied inferential statistics to compare precision, recall, F1-score, and efficiency across conditions. All approaches were evaluated using the same dataset of 40 SAR images containing 103 oil spills. We measured the quality of the created annotations using Intersection over Union (IoU) with existing ground truth labels.

5.1 Oil Spill Detection

We analyze the oil spill detection performance based on the metrics true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The trained YOLO model detected 83 of 103 oil spills correctly (TP). However, the model missed detecting 14 oil spills (FN) and falsely predicted another 20 oil spills (FP) and provided us with a precision of 0.856 and a recall of 0.806, leading to an F1-score of 0.830. In the non-assisted condition, 1071 bounding boxes were drawn in total by all participants, including 773 correct labels (TP), 298 incorrect labels (FP), and 611 oil spills that were not detected (FN), resulting in a precision of 0.755 and recall of 0.576, which led an F1-score of 0.653. For the ML-assisted UI, we used a confidence-similar mapping to transform the ordinal scale to a numeric scale to ensure that the data obtained from the user study could be effectively analyzed to measure the overall detection performance and compare the results across different tasks. To assess the performance of the ML-assisted task, the verification of oil spill predictions and the additionally drawn bounding boxes were considered. The prediction of the ML model was confirmed by the crowd's agreement responses. The level of agreement was calculated for each box separately. We analyze the confidence via

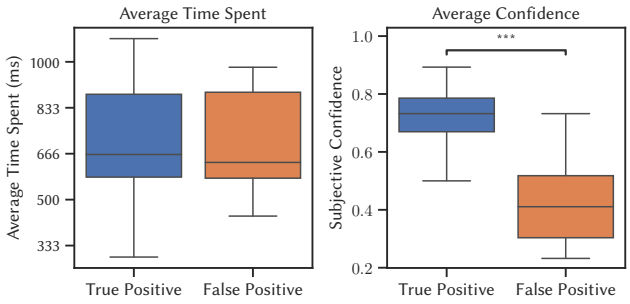


Figure 3: Comparison between TP and FP results. There is insufficient evidence to show a significant difference in average time spent. Instead, participants reported significantly higher confidence in TP compared to FP.

a 5-point Likert scale. In the ML-assisted condition, 223 bounding boxes were drawn, including 107 correct labels (TP) and 116 incorrect labels (FP). Regarding the oil spill prediction verification, the participants verified 820 oil spills correctly (TP). However, 56 ML predictions that are lookalikes are identified as oil spills (FP). In summary, 927 oil spills are correctly identified, and 182 are mistakenly classified as oil spills in the ML-assisted condition. Furthermore, 515 oil spills were not detected during this task. In this case, we got an increased precision of 0.843 and recall of 0.642 compared to the non-assisted UI, leading to a higher F1-score of 0.729.

5.2 Crowd Contributions

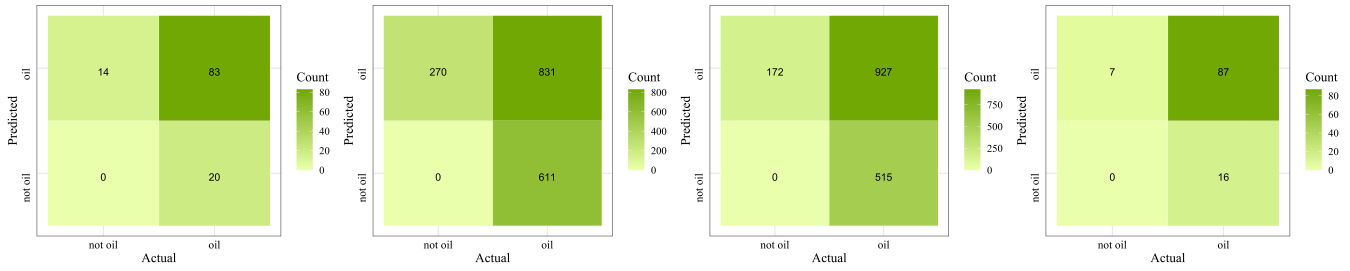
5.2.1 Annotation Performance. We assessed the normality of distribution of the average subjectively reported confidence using the Shapiro-Wilk test. For TPs, the test indicated a significant deviation from normality ($W = 0.948, p = 0.002$). Conversely, for FP cases, the subjective confidence was found to adhere to a normal distribution ($W = 0.913, p = 0.202$). In addition, the homogeneity of variances was evaluated using the Levene test, revealing significant differences in the variances between TP and FP for subjective confidence ($p = 0.004, W = 8.559$). Furthermore, a comparison of the distributions for TP ($M = 0.732$) and FP ($M = 0.411$) using the Mann-Whitney U test demonstrated a statistically significant difference ($u = 988.000, p < .001$). Similarly, we also evaluated the distribution of average completion time. For TPs, there was a significant deviation from a normal distribution ($W = 0.965, p = 0.022$). On the other hand, for FPs, the average completion time conformed to a normal distribution ($W = 0.891, p = 0.102$). The Levene test confirmed the equality of variances between TP and FP for average completion time ($p = 0.698, W = 0.152$). Additionally, when comparing the distributions of TP and FP using the Mann-Whitney U test, no significant difference was observed ($u = 572.000, p = 0.366$), as shown in Figure 3. These results, thereby, support hypothesis **H2**.

In the evaluation process of the ML-assisted detection task, the crowd demonstrated capabilities to complement the ML model. Firstly, they identified a significant portion of FPs through a low confidence score, which confirms our **H2** for FPs. This corrective ability significantly enhanced the overall accuracy of the combined model (collective crowd evaluation combined with ML model predictions), as shown in Table 1. However, throughout the tasks, the crowd also made mistakes. Particularly, falsely drawn bounding boxes were found during the non-assisted task, which led to a lower precision of 0.755 (see Table 1). However, the integration of the ML assistance proved beneficial in mitigating FPs by 61% (see 4c). In addition, with the help of ML assistance, the number of correctly detected oil spills was increased by 11.6%. Furthermore, in several instances, the crowd successfully identified undetected oil spills (FN) by the ML model **(see Figure 5)**. Additionally, the crowd extended the identified oil spill areas beyond the ML predictions (see Figure 5). This capability adds value to the detection process, occurring around 30 times, even without explicit instructions.

5.2.2 Crowd Consensus. For adapting this crowd consensus concept, we introduce a confidence threshold, $c_{\text{threshold}}$ as a validation mechanism to mitigate individual errors for the decisions on the ML predictions. If the participants' average confidence for an ML-predicted oil spill is equal to or above $c_{\text{threshold}}$, the prediction is considered an oil spill and vice versa. A similar approach is applied to the additionally drawn bounding boxes. Only areas labeled with a sufficient number of participants are considered. Therefore, the ratio of participants who drew a bounding box in the same area to the overall number of participants who evaluated the image is considered. The labeled area is considered an oil spill if this ratio exceeds or equals $c_{\text{threshold}}$. In our case, setting $c_{\text{threshold}} = \mathbf{0.5}$ strikes the best equilibrium between precision and recall as we have a higher F1-score at that point. This configuration maximizes the detection of true positives while keeping false positives to a minimum (refer to 7b). As mentioned in subsection 4.2, each image had 14 evaluations for both the ML-assisted and non-assisted UI. Therefore, for manual labeling in both UIs, setting $c_{\text{threshold}} = 0.5$ gives us $0.5 * 14 = 7$ agreements for considering an object as an oil spill. After applying the $c_{\text{threshold}} = 0.5$, the crowd consensus approach assisted by the ML model identified 87 oil spills correctly (TP), while seven areas were mistakenly identified as oil spills (FP), and 16 oil spills were missed (FN). This provided us with a higher precision of 0.926 and a recall of 0.845, eventually leading to the highest F1-score of 0.884. These metrics provide valuable insights into the performance of the crowd consensus contribution combined with the ML model in terms of the effectiveness of the assisted object detection tool (see 4b).

5.3 Quality and Efficiency

We conducted a Shapiro-Wilk normality test for the efficiency and quality of data and found that our data was not normally distributed. We then performed a Wilcoxon test and found that participants completed the ML-assisted task significantly faster than the non-assisted ($M = 46, SD = 30.02$) task ($M = 38.72, SD = 22.64, z = 7.35, p < .001$), which supports our hypothesis **H1**. Considering only the first 20 images, non-assisted ($M = 56.41, SD = 31.77$), assisted ($M = 40.31, SD = 23.31, z = 11.56, p < .001$), the efficiency gap widens to 56 seconds for the non-assisted task and 40 seconds for the ML-assisted task on average. Consequently, an efficiency improvement of 29% can be achieved by integrating ML assistance for less experienced users. Additionally, the participants' efficiency demonstrated an upward trend throughout the study, indicating a

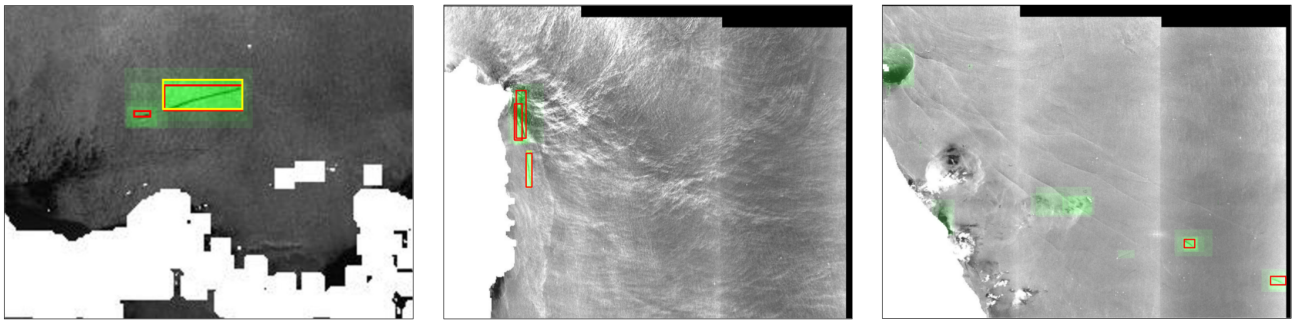


(a) ML model (YOLO) (b) Non-assisted (c) ML-Assisted (d) ML-Assisted + Consensus

Figure 4: Confusion matrices of (a) the YOLO model, (b) non-assisted crowd detection, (c) ML-assisted detection, and (d) validated ML-assisted detection on the user study image set.

Table 1: Accuracy metrics table of the YOLO model, non-assisted detection task, ML-assisted detection task, and validated ML-assisted detection task.

	YOLO model	non-assisted detection task	ML-assisted detection	Validated ML-assisted detection (crowd consensus-ML approach) ($c_{\text{threshold}} = 0.5$)
Precision	0.856	0.755	0.843	0.926
Recall	0.806	0.576	0.642	0.845
F1-score	0.830	0.653	0.729	0.884



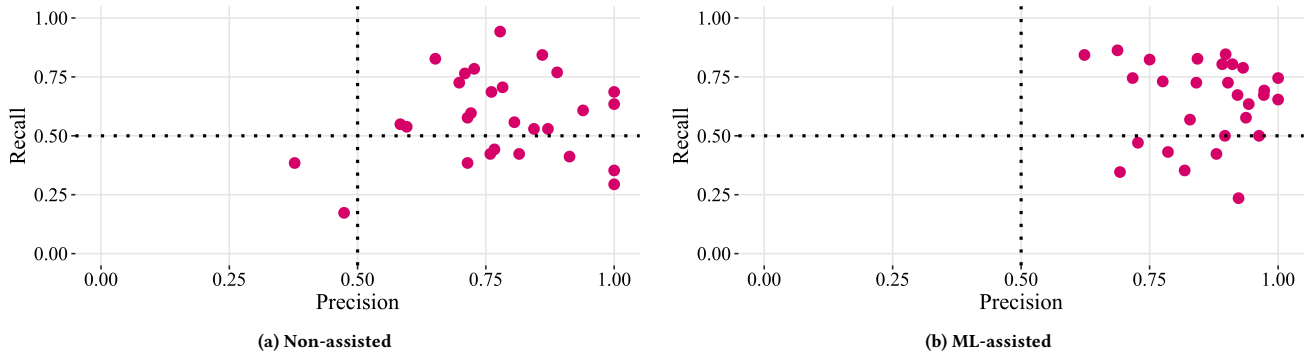
(a) Crowd Extension (b) ML-assisted UI (c) Non-assisted UI

Figure 5: (a) Examples of crowd extensions to cover the entire oil spill occurrence. Red bounding boxes are ground truth labels. Yellow bounding boxes are ML predictions. Green-filled boxes are evaluations from the crowd. Here, bright green-filled rectangles signify areas where the collective crowd expresses high confidence in the proposed area. Faded green-filled rectangles represent areas where the confidence is lower or where an individual participant marked the region as a potential oil spill area. (b) Identifying more oil spills not detected by the ML. (c) Mistakes by the crowd.

learning process as they became more familiar with the interface and detection procedures. However, after the task switch, the participants' performance becomes influenced, making an unbiased analysis feasible only for the first 20 images. After switching the tasks, participants of both groups spent a similar amount of time on the tasks (see Figure 7). Furthermore, our findings also show that using the ML-assisted UI increased the quality of the crowd based on individuals' performance on precision and recall (see Figure 6).

5.4 User Experience and Technology Acceptance

For user experience and preferences, we assessed the open-ended, single-choice, and TAM questions. For the TAM, we followed the evaluation as described in Davis [12] by comparing the Perceived Usefulness (PU) and Perceived Ease-of-Use (PEU) factors for both conditions through a Wilcoxon-Signed Rank test. The non-assisted interface ($M = 5.685, SD = 1.194$) was higher rated than the ML-assisted one ($M = 5.333, SD = 1.43$) for PU, resulting in a significant



(a) Non-assisted (b) ML-assisted

Figure 6: Quality of the crowd based on individuals' performance on precision and recall.

difference, $Z = -2.481, < 0.05$. Comparing the results for PEU between the non-assisted interface ($M = 6.071, SD = 1.382$) and the ML-assisted interface ($M = 6.024 SD = 1.524$) did not reveal any significant results ($Z = -0.144, > 0.05$). These results indicate greater technology acceptance for the non-assisted UI. In contrast, all participants would choose the ML-assisted interface for future oil detection tasks. Further, the majority found the ML-assisted more time efficient (n=26) and easier for making accurate judgments (n=27). Participants appreciated the oil spill examples: *"I looked at the example oil spills and lookalikes. For me, the example oil spills helped more than the lookalikes because the image quality of the lookalikes was very poor [...]."*, P8. Overall, participants found both interfaces easy to use but hard to differentiate between oil spills and example images. However, we also noted two people with more negative attitudes toward the ML-assisted UI, who further rated it considerably lower in the TAM questionnaire. Suggestions regarding UI improvements concerned enabling different types of boxes to increase precision in outlining oil spills or providing better example images and more context information about them. Further, only seven participants would reuse and continue to support refining the ML algorithm, particularly if some reward was offered. The remaining participants did not feel qualified, lacked the time, or did not want to spend their free time on such a task. Motivational factors could be feedback on users' performance or a gamified approach; e.g., *"Making it like a game and having some points and ranking system."*, P9. This highlights another challenge for crowd-ML collaboration concerning participants' engagement and reward system.

6 DISCUSSION

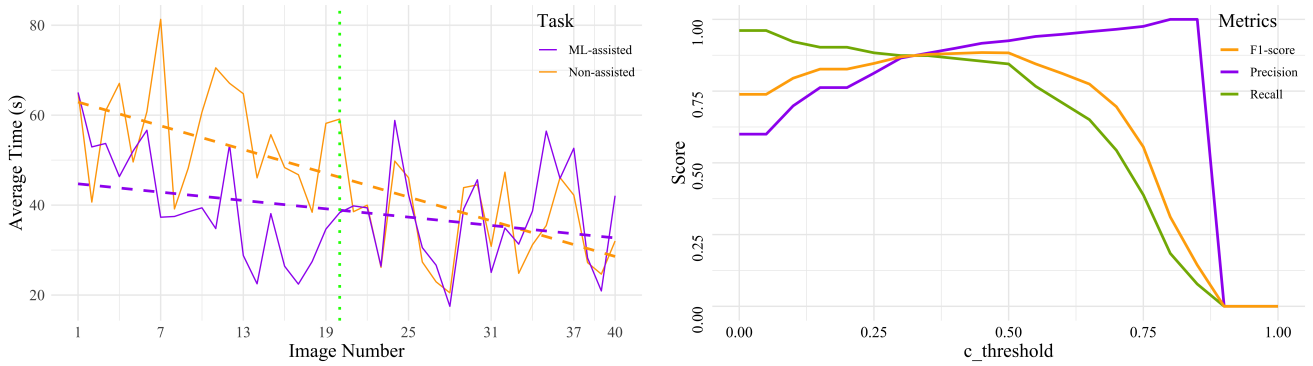
Our work compared two UIs for oil spill detection in satellite images, one applying human-AI collaboration (ML-assisted CK) and one based on CK only. We assessed the differences in precision, recall, efficiency, and performance between the UIs (**RQ1**). The results support our hypotheses, **H1** about ML-assisted CK performs more efficiently than the non-assisted UI, and **H2** the crowd detects ML model's errors. Further, we will discuss the takeaways and synergies we found for crowd-ML collaboration (**RQ2**) below.

Crowd-ML collaboration significantly improves precision and recall compared to only CK. Supporting our hypothesis **H1**, compared to the non-assisted approach, the ML-assisted UI enhances the participants' ability to assess oil spills, improving precision and recall by approximately 11%. This means that more oil spills are correctly identified while errors such as undetected oil spills and incorrect labeling of non-oil spill areas are reduced. Therefore, humans or crowds can contribute to ML by bringing human expertise and perception to enhance overall performance. It shows the synergy potential generated through crowd-ML collaboration.

Displaying ML predictions to users accelerate the crowd-learning process. The ML-assisted UI decreased participants' evaluation time, enhancing their efficiency. Particularly, within the initial 20 images before the task switch, the average evaluation time decreased as participants gained familiarity with the UI and the oil spill labeling process. It is worth noting that the suggestions from the ML model accelerated the learning process, and multiple participants acknowledged the significance of these suggestions compared to the oil spill lookalike images. This indicates a crowd-learning effect from the information provided by the ML model.

Crowd consensus can significantly improve ML precision and recall. The collective crowd contributions identified prediction errors from the object detection model. While individual non-experts may not achieve the accuracy of the ML model or can make precise expert predictions, the cumulative crowd input by using the confidence threshold $c_{\text{threshold}}$ allows for forming a more conclusive and precise prediction significantly improving precision and recall compared to ML model predictions. This makes crowd-ML collaboration also interesting for other use cases beyond the scope of our work, such as finding errors or faults in industrial activities [1, 37] or in sports analytics [38].

Crowd complements the ML model by extending bounding boxes and identifying errors. The crowd also identified errors such as missed objects and wrong detections, supporting **H2**. The crowd also exhibited a notable skill in extending the proposed oil spill areas, pinpointing regions the model had missed. The crowd's ability to discern and extend oil spill areas not captured by the ML model demonstrates their capacity to contribute supplementary insights and also can be used to label datasets of oil spills as there



(a) All Images (b) Confidence threshold

Figure 7: (a) Average evaluation time for each image for the non-assisted task (orange), and for the ML-assisted task (purple). The task switch is indicated with the vertical green line. The orange and purple dotted lines display the general direction (trendline) of the completion time. (b) Confidence-Precision Curve (purple), Confidence-Recall Curve (green), and Confidence-F1 Curve (orange)

is a scarcity of real oil spill data [14], by working as an alternative to experts.

Integrating CK with ML models through an intelligent interface helps mitigate environmental hazards. Our study reveals that when the crowd collaborates with ML algorithms through an intelligent interface, there's an increase in accuracy for identifying oil spills compared to relying solely on ML models. By displaying ML predictions, we expedite the learning process for the non-expert crowd, who can leverage the model's suggestions to enhance their assessments. Moreover, collective insights from the crowd serve to strengthen the model's performance, highlighting its potential to augment human expertise in environmental monitoring. Additionally, the crowd's ability to identify missed oil spills and errors complement the ML model, contributing to more comprehensive and accurate assessments of environmental hazards.

Including explainable AI techniques could improve the crowd-ML relationship. The black-box nature of ML is commonly identified as a threat to user trust and accountability [25]. As a result, users may find it challenging to fully understand the reasoning behind the ML predictions. This lack of transparency can reduce confidence in the collaborative approach, which might explain the lower TAM results. To address this issue, we suggest including explainable AI techniques in future crowd-ML collaborations that could provide valuable insights into an ML's decision-making process [26].

7 LIMITATIONS AND FUTURE WORK

We address current limitations while emphasizing new research opportunities. Encouraging active participation from a large crowd can be challenging. To address this and sustain participants' motivation, implementing gamification or a reward system could prove effective [28]. Further, the concept of crowd-ML collaboration might face limitations when dealing with complex tasks, such as identifying cancer cells in microscopic images, which is a highly intricate task that may be exceptionally challenging for non-experts [25]. Therefore, it's crucial to assess the suitability of the collaboration for specific object detection tasks based on their complexity and the required expertise level. Furthermore, this work only focuses on the use of non-experts; the results should be compared with respect to the findings of experts and qualitative feedback on refining the tool. In the future, we aim to explore how crowd-ML performance could complement experts' labeling or how to design a system that facilitates such collaboration. Additionally, we intend to make the annotation tool open source for broader application across various sectors, drawing inspiration from the work of Schilling et al. [33].

8 CONCLUSION

Our work introduces a crowd-ML collaboration intelligent interface that helps assess non-experts' performance in an oil spill object detection task with and without the assistance of an ML model. The CK utilizing the ML-assisted UI outperformed the non-assisted condition by achieving higher oil spill detection accuracy and mitigated errors. Notably, the collaborative effort of the crowd identifies a considerable number of errors wrongly proposed by the ML model. Furthermore, our work shows that the crowd can complement the capabilities of an imperfect ML model. Our findings further show the potential of including non-experts in responding to natural catastrophes from remote by training object detection models. Thus, our work is relevant for the field of human-AI and environmental science, contributing to human-AI collaborations. For future research, the detections of the crowd-ML model can serve as potential valuable labels for further model training and provide an alternative to expert-driven labeling.

ACKNOWLEDGMENTS

This research was funded by Elitenetzwerk Bayern.

REFERENCES

[1] Hafiz Mughees Ahmad and Afshin Rahimi. 2022. Deep learning methods for object detection in smart manufacturing: A survey. *Journal of Manufacturing Systems* 64 (2022), 181–196.

[2] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari. 2018. Fluid annotation: a human-machine collaboration interface for full image annotation. In *Proc. of the 26th ACM international conference on Multimedia*. 1957–1966.

[3] Niyazi Arslan, Meysam Majidi Nezhad, Azim Heydari, Davide Astiaso Garcia, and Georgios Sylaios. 2023. A Principal Component Analysis Methodology of Oil Spill Detection and Monitoring Using Satellite Remote Sensing Sensors. *Remote Sensing* 15, 5 (2023), 1460.

[4] Till Blesik, Markus Bick, and Tyge-F. Kummer. 2022. A Conceptualisation of Crowd Knowledge. *Information Systems Frontiers* 24, 5 (Oct. 2022), 1647–1665. https://doi.org/10.1007/s10796-021-10176-y

[5] JAMES L Bodkin, Daniel Esler, STANLEY D Rice, Craig O Matkin, and BRENDA E Ballachey. 2014. The effects of spilled oil on coastal ecosystems: lessons from the Exxon Valdez spill. *Coastal conservation* 19 (2014), 311.

[6] Justin Brooks. 2019. COCO Annotator. https://github.com/jsbroks/coco-annotator/.

[7] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300234

[8] Chia-Ming Chang, Yi He, Xi Yang, Haoran Xie, and Takeo Igarashi. 2022. Dual-Label: secondary labels for challenging image annotation. In *Graphics Interface 2022*.

[9] Anne Cocos, Ting Qian, Chris Callison-Burch, and Aaron J Masino. 2017. Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. *Journal of biomedical informatics* 69 (2017), 86–92.

[10] Mary Missy Cummings. 2014. Man versus machine or man + machine? *IEEE Intelligent Systems* 29, 5 (2014), 62–69. https://doi.org/10.1109/mis.2014.87

[11] CVAT. [n. d.]. Computer Vision Annotation Tool (CVAT). https://github.com/opencv/cvat/.

[12] Fred Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13 (09 1989), 319–. https://doi.org/10.2307/249008

[13] Amri Emna, Benoit Alexandre, Philippe Bolon, Migebielle Véronique, Conche Bruno, and Oppenheim Georges. 2020. Offshore oil slicks detection from sar images through the mask-rcnn deep learning model. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[14] Jianchao Fan and Chuan Liu. 2023. Multitask GANs for Oil Spill Classification and Semantic Segmentation Based on SAR Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 2532–2546. https://doi.org/10.1109/JSTARS.2023.3249680

[15] Rebecca Fiebrink, Perry R Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proc. of the SIGCHI conference on human factors in computing systems*. 147–156.

[16] Lucy Fortson, Karen Masters, Robert Nichol, EM Edmondson, C Lintott, J Raddick, and J Wallin. 2012. Galaxy zoo. *Advances in machine learning and data mining for astronomy* 2012 (2012), 213–236.

[17] Ross Girshick. 2015. Fast r-cnn. In *Proc. of the IEEE international conference on computer vision*. 1440–1448.

[18] Tom Haider and Florian Michahelles. 2021. Human-machine collaboration on data annotation of images by semi-automatic labeling. In *Proc. of Mensch und Computer 2021*. 552–556.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proc. of the IEEE international conference on computer vision*. 2961–2969.

[20] Xudong Huang, Biao Zhang, William Perrie, Yingcheng Lu, and Chen Wang. 2022. A novel deep learning method for marine oil spill detection from satellite synthetic aperture radar imagery. *Marine Pollution Bulletin* 179 (2022), 113666. https://doi.org/10.1016/j.marpolbul.2022.113666

[21] Jane Hung, Deepali Ravel, Stefanie C. P. Lopes, Gabriel Rangel, Odailton Amaral Nery, Benoit Malleret, Francois Nosten, Marcus V. G. Lacerda, Marcelo U. Ferreira, Laurent Rénia, Manoj T. Duraisingh, Fabio T. M. Costa, Matthias Marti, and Anne E. Carpenter. 2019. Applying Faster R-CNN for Object Detection on Malaria Images. arXiv:1804.09548 [cs.CV]

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

[23] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez I Badia. 2022. Towards Efficient Annotations for a Human-AI Collaborative, Clinical Decision Support System: A Case Study on Physical Stroke Rehabilitation Assessment. In *27th International Conference on Intelligent User Interfaces*. ACM, Helsinki Finland, 4–14. https://doi.org/10.1145/3490099.3511112

[24] James Jim Lewis. 2019. Comparison of Four TAM Item Formats: Effect of Response Option Labels and Order. *Journal of Usability Studies* 14, 4 (2019), 224–236. https://doi.org/tam-formats-effect-response-labels-order/

[25] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid Assisted Visual Search: Supporting Digital Pathologists with Imperfect AI. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 504–513. https://doi.org/10.1145/3397481.3450681

[26] Guoyang Liu, Jindi Zhang, Antoni B. Chan, and Janet Hsiao. 2023. Human Attention-Guided Explainable AI for Object Detection. *Proc. of the Annual Meeting of the Cognitive Science Society* 45 (2023). https://doi.org/uc/item/9r53b44n

[27] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang. 2017. Rotated region based CNN for ship detection. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 900–904.

[28] Benedikt Morschheuser, Juho Hamari, and Jonna Koivisto. 2016. Gamification in Crowdsourcing: A Review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 4375–4384. https://doi.org/10.1109/HICSS.2016.543

[29] Mario Nieto-Hidalgo, Antonio-Javier Gallego, Pablo Gil, and Antonio Pertusa. 2018. Two-stage convolutional neural network for ship and spill detection using SLAR images. *IEEE Transactions on geoscience and remote sensing* 56, 9 (2018), 5217–5230.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proc. of the IEEE conference on computer vision and pattern recognition*. 779–788.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[32] Kaja Scheliga, Sascha Friesike, Cornelius Puschmann, and Benedikt Fecher. 2016. Setting up crowd science projects. *Public Understanding of Science* 27, 5 (2016), 515–534. https://doi.org/10.1177/0963662516678514

[33] Marcel P. Schilling, Svenja Schmelzer, Lukas Klinger, and Markus Reischl. 2022. KaIDA: a modular tool for assisting image annotation in deep learning. *Journal of Integrative Bioinformatics* 19, 4 (2022), 20220018. https://doi.org/doi:10.1515/jib-2022-0018

[34] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, Andre A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M. Hochachka, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Will Morris, Tina B. Phillips, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, Andrea Wiggins, David W. Winkler, Weng-Keen Wong, Christopher L. Wood, Jun Yu, and Steve Kelling. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation* 169 (2014), 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

[35] Yongqiang Tian, Shiqing Ma, Ming Wen, Yepang Liu, Shing-Chi Cheung, and Xiangyu Zhang. 2021. To what extent do DNN-based image classification models make unreliable inferences? *Empirical Software Engineering* 26, 5 (2021), 84.

[36] Jiangtao Wang, Yasha Wang, and Qin Lv. 2019. Crowd-Assisted Machine Learning: Current Issues and Future Directions. *Computer* 52, 1 (Jan 2019), 46–53. https://doi.org/10.1109/MC.2018.2890174

[37] Daniel Weimer, Bernd Scholz-Reiter, and Moshe Shpitalni. 2016. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP annals* 65, 1 (2016), 417–420.

[38] Yinda Xu and Yonggang Peng. 2020. Real-Time Possessing Relationship Detection for Sports Analytics. In *2020 39th Chinese Control Conference (CCC)*. IEEE, 7373–7378.

[39] Yi-Jie Yang, Suman Singha, and Roberto Mayerle. 2022. A deep learning based oil spill detector using Sentinel-1 SAR imagery. *International Journal of Remote Sensing* 43, 11 (2022), 4287–4314.

[40] ST Yekeen and A-L Balogun. 2020. Automated marine oil spill detection using deep learning instance segmentation model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43 (2020), 1271–1276.

[41] Shamsudeen Temitope Yekeen and Abdul-Lateef Balogun. 2020. Advances in remote sensing technology, machine learning and deep learning for marine oil spill detection, prediction and vulnerability assessment. *Remote Sens* 12, 20 (2020), 1–31.

[42] Shamsudeen Temitope Yekeen, Abdul-Lateef Balogun, and Khamaruzaman B Wan Yusof. 2020. A novel deep learning instance segmentation model for automated marine oil spill detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), 190–200.

[43] Zheng Zhang, Zheng Ning, Chenliang Xu, Yapeng Tian, and Toby Jia-Jun Li. 2023. PEANUT: A Human-AI Collaborative Tool for Annotating Audio-Visual Data. In *Proc. of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[44] James Zou, Kamalika Chaudhuri, and Adam Kalai. 2015. Crowdsourcing feature discovery via adaptively chosen comparisons. In *Proc. of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 3. 198–205.

[45] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proc. of the IEEE* 111, 3 (2023), 257–276. https://doi.org/10.1109/jproc.2023.3238524