HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230073

Human Factors in Interactive Online Machine Learning

Agnes TEGEN [a,b], Paul DAVIDSSON [a,1] and Jan A. PERSSON [a]
[a] *Department of Computer Science and Media Technology, Internet of Things and People Research Center, Malmö University, Sweden*
[b] *Swedish Defense Research Agency (FOI), Stockholm, Sweden*
ORCiD ID: Agnes Tegen https://orcid.org/0000-0002-3155-8408, Paul Davidsson https://orcid.org/0000-0003-0998-6585, Jan A. Persson https://orcid.org/0000-0002-9471-8405

Abstract. Interactive machine learning (ML) adds a human-in-the-loop aspect to a ML system. Even though the input from human users to the system is a central part of the concept, the uncertainty caused by the human feedback is often not considered in interactive ML. The assumption that the human user is expected to always provide correct feedback, typically does not hold in real-world scenarios. This is especially important for when the cognitive workload of the human is high, for instance in online learning from streaming data where there are time constraints for providing the feedback. We present experiments of interactive online ML with human participants, and compare the results to simulated experiments where humans are always correct. We found combining the two interactive learning paradigms, active learning and machine teaching, resulted in better performance compared to machine teaching alone. The results also showed an increased discrepancy between the experiments with human participants and the simulated experiments when the cognitive workload was increased. The findings suggest the importance of taking uncertainty caused by human factors into consideration in interactive ML, especially in situations which requires a high cognitive workload for the human.

Keywords. interactive machine learning, online learning, human factors

1. Introduction

Interactive machine learning (ML) aims to utilize human feedback to improve performance and minimize the cost of labelling data in ML. The two main paradigms of interactive ML are active learning [1], where the system queries a human user for feedback, and machine teaching [2], where the human decides when and which feedback is provided. Often in interactive learning, the human is assumed to always provide correct feedback. In active learning the term *oracle* is frequently used for the feedback provider, indicating that it is never wrong. These assumptions do often not reflect real-world scenarios with human participants, where the feedback can be uncertain because of human factors, e.g. due to errors or missing data. How human factors affect the feedback and performance is

[1] Corresponding Author: Paul Davidsson, paul.davidsson@mau.se.

especially important for online learning as the data is presented in a single-pass streaming manner [3], where time constraints typically influence the interaction process, and when the cognitive workload for the human is high. For instance, consider a scenario where a human operator is monitoring multiple video streams in parallel and they are expected to provide feedback on what is shown in the video streams [4]. In these scenarios it is not always possible to store data (e.g. due to privacy regulations such as GDPR) and the interaction has to be made on the fly.

We present experiments on interactive online ML with human participants. The results are compared to simulated experiments with oracles where all data points are labelled correctly. We study how experiments with simulated participants and their assumptions differs from experiments with human participants, as well as how an increased cognitive workload for the human participants affects performance of the learning system. The aim is to validate the interactive learning strategies, as well as the results from the simulated experiments, by including the uncertainty which real human users bring. The results show a difference in the performance between the experiments with human participants and the simulated experiments, showcasing the importance of including human factors in interactive ML models.

2. Related work

Most work in interactive online ML do not include human participants in their experiments or discuss how uncertainty caused by human factors can affect the performance and learning process. It is often assumed that the teacher always will provide correct feedback and that there is no human error. These assumptions might not hold in a real-world scenario however, due to the complex interactive process of a system with human-in-the-loop [5,6,7]. Human factors such as providing incorrect feedback, not replying to a query or in some other way not acting in accordance with what is expected, can affect the feedback provided.

In previous work we studied how human factors can affect performance in interactive online learning [8]. Experiments were performed where the reliability of the teacher is varied. Further, active learning experiments where the oracle might not reply to a query and might provide an incorrect label, to make the scenario more realistic have been performed by Donmez and Carbonell [9]. In a previous study, we evolve the notion of labelling cost in interactive learning. Typically, the number of labels provided are used as the single measure of labelling cost, but here the amount of attention needed from the teacher is also investigated [10]. Vembu and Zilles introduce an interactive ML method where multiple teachers provide feedback. The disagreement among the annotators is used to estimate how meaningful a training example is [11]. Yan et. al. study a scenario where the teacher might provide incorrect labels as well as abstain from providing labels at all. A theoretical analysis is made regarding the estimation error given a labelling budget [12]. Lughofer presents an active learning method for single-pass streaming data that combines *conflict*, data instances that are on the border between multiple classes, and *ignorance*, data instances that are in yet unexplored parts of the feature space [13]. Krawczyk addresses the issue of activity recognition in data streams by introducing an active and adaptive ensemble classifier as well as verifying it with experiments on six real-life datasets [14]. In these publications, no human factors were included in the experiments and most of them do not address a problem scenario with single-pass streaming data.

Du et al. adds example-dependent noise to oracles to make them more human-like [15]. They study when and how an annotator provides incorrect answers. Based on the findings they present an active learning algorithm which they then test in simulated experiments. Huang et al. presents an "oracle epiphany model" to make the interactions between active learning algorithms and oracles more realistic [16]. Instead of only ignoring from providing labels that is uncertain for the annotator, they can temporarily abstain from giving a label until enough examples have been presented to them and they have an "epiphany" of how to label. They illustrate their model in simulated experiments.

None of the above articles include human participants in the experiments. Simulated experiments are very useful as they are easier to control and can generate a lot of knowledge. However, experiments with human participants is an important complement to simulated experiments, as assumptions and simplifications always are made in simulations. Some of these works employ a definition of machine teaching where the human is the learner [17,18]. Miu et. al. present an online active learning framework for collection in real-time of annotations of human activity recognition tasks [19]. The framework was also implemented as a mobile app and tested with human participants. The human user could only reply when queried by the active learning strategy and not proactively provide labels themselves. Jauhri et. al. present an interactive learning framework which uses human feedback to improve the behaviour of an agent [20]. The method is evaluated with both simulated experiments and using a real robot-arm with non-expert human participants as teachers. He et. al also study the use of human feedback to improve the behaviour of an agent. They introduce a feedback model which takes human uncertainty into account [21].

Another aspect of uncertainty in labelling is that each example might be associated with more than one label or that the annotator is uncertain between two or more labels. Collins et al. presents soft labels and Geng introduces label distribution learning, where an annotator can provide more than one label per data instance as well as probabilities for each of them [22,23]. These works focus on the uncertainty that the annotator is aware of, not mistakes or limitations of a human.

Our work presents a study of how human factors affect performance for interactive online ML by comparing results from experiments with human participants to simulated experiments. We have not found any previous study that focuses on the uncertainty arising from the human teacher in interactive online ML. Interactive ML typically does not discuss how human factors can affect results, even less include it in the experiments. Moreover, the aspect of streaming data is not well explored within interactive ML (typically batch learning or iterative learning is considered) but is a main part of our work. Also, in a scenario with streaming data, the effects of the uncertainty due to human factors on performance are larger as time constraints affect the performance of the human.

3. Experiments

In the experiments, the participants were asked to annotate images presented in a single-pass streaming manner. The aim was to compare different interactive ML strategies and study how the amount of cognitive workload for the human user affected their ability to provide correct feedback and how this uncertainty affects the performance of the ML model. The participants partook anonymously in the experiments using a local computer and were asked to use a computer mouse (i.e. not touch pad), while the processing and

Figure 1. GUI for the interactive online ML experiments. This example is of the step with three parallel image data streams. The participant can select one of the two classes 'cat' or 'dog' for each image. The blue highlights show the suggested classification from the ML system, which can be changed by the participant if she/he think they are incorrect. In this example, two images are correctly classified, but the middle image is not.

storage of data was done on a server hosted in Amazon Web Services (see code for details). The participants were informed that data they generated would be used in a study, but the data stored does not contain personally identifiable information. The participants were volunteers and they did not receive any compensation for their participation. We consulted the Ethics Council at the university regarding the need for an ethical review but since it did not contain any personally identifiable information and the experiment only consisted of labelling images a review was not considered necessary. In total, the results include data from 18 participants.

The experiment was divided into four steps of increasing difficulty and therefore gradually higher cognitive workload for the participants. The number of data streams run in parallel were increased in each step, i.e. the number of images displayed at the same time. The steps contained one, three, six and nine parallel data streams respectively. Figure 1 contains an example of the GUI showcasing the second step, containing three data streams. In this example two are chosen correctly as 'cat' and 'dog' respectively, but the image in the middle was inaccurately classified as 'dog' by the ML algorithm. Each data stream consisted of a sequence of 30 randomly selected images of cats and dogs. The participant had four seconds to annotate each setup of images, regardless if it was one image in the first step or nine images in the last step. The number of images in combination with the time constraint were selected based on initial testing to represent four levels of difficulty and cognitive workload, *easy*, *medium*, *difficult* and *nearly impossible*.

For each step, two separate interactive learning strategies were tested, the *MT strategy* and the *ALMT strategy*. In the first strategy, machine teaching was employed. All images were shown and the human participant had to be proactive in deciding what to annotate. When the system had gathered enough annotated data from the participant, a prediction was displayed, which they could either change or not change. If the participant changed the label, the data point with the selected label was returned to the system for training. If the pre-selected label was not changed, due to a conscious choice or human factors, the data point with the pre-selected choice from the systems prediction was still returned to the system as an annotated data point. In the second strategy, *ALMT strategy* active learning in combination with machine teaching were employed. The active learning strategy analyzed the images and only displayed a selection for the human participant based on how certain it was regarding the classifications of the images (the active learning strategy is described further in the next subsection). The two different approaches were evaluated to solve the same problem. In both approaches, the human user and the machine learner work together to to solve it. The two different strategies differ in the sense that for *MT strategy* all decisions concerning the annotation process are made by the human, while for *ALMT*

strategy this is divided between the human and the machine learner. A real world example of this setup is e.g. a surveillance room where a security operator is monitoring a number of parallel video streams to identify undesired states, e.g. a burglary or a fire. The idea is to teach the system to be better at identifying such states automatically.

One experiment session contained eight subsessions, where each subsession had a unique combination of number of parallel data streams and interactive learning strategy. For each data stream, a ML algorithm was trained from scratch with the sequence of images and human feedback presented, i.e. in a cold start scenario. When the algorithm had enough training data to do predictions, these were presented to the participant as pre-selected choices of the images in the GUI. The participant could change this pre-selected label if it was incorrect. However, this became increasingly more difficult with more images displayed in parallel, as the time to provide feedback was kept constant even though the number of data streams increased. The full text of instructions are included in appendix A.

3.1. Interactive machine learning strategies and algorithms

We used the Python library creme for the implementation of the ML [24]. It focuses on incremental learning, which is suitable when training is done iteratively as more data samples are gradually being presented. In our experiments, the ML algorithm is updated with each new labelled data point it receives. The ML algorithm employed was logistic regression. It performs the learning task well and is not computationally heavy [25], which was important for the learning to be done in real-time. A cold start scenario was employed in the experiments, which means the classifier must learn quickly and relatively fast reach an acceptable performance level. Logistic regression had also been previously shown to be well suited for the data used. While it is essential to have a ML algorithm suitable for the problem scenario, the aim of this work was not to find the best performing one, but rather to study aspects of human factors in relation to ML.

The active learning strategy used was entropy [25] with a varying threshold [26]. Entropy is a measurement of whether the classifier is uncertain regarding its own classifications. If the classifier is uncertain, it will query the human participant. The variable threshold adjusts the level of what is considered uncertain over time, as this might change e.g. when the classifier improves or when new data is presented. The results from the experiments with human participants were compared to results from simulated experiments with an oracle. In the simulated experiments, the same set of experiment sessions with the same sequences of images were run, but always given correct feedback. The two strategies are titled *MT oracle strategy* and *ALMT oracle strategy*. This is similar to how experiments on interactive ML, especially active learning, often simulates human behaviour and interaction. Well-known interactive learning strategies and ML algorithms were chosen because the aim of the experiments was not to invent or improve on ML algorithms or interactive learning strategies, but rather to study how well different types perform, how results from human participants compare to simulated results, and how human factors affects results. The code for the experiments and the data collected can be found by the following link.[2]

[2]https://github.com/iotap-center/human_in_the_loop_experiment

3.2. Dataset

The dataset used for the experiments was Kaggle Dogs vs. Cats and contains 25000 images of cats and dogs[3]. From the 25000 images, 1000 images were extracted, 500 of each category, and used in the experiments. When choosing a dataset for these experiments, it was important that it was easily recognizable for the participants, i.e. the expertise or knowledge of the participants should not decide if they could provide feedback or not. We also choose an easier classification task with only binary classification because of the time constraint in combination with cold start of the ML algorithm. Examples of images from the dataset are displayed in Figure 1.

4. Results

Figure 2 displays the accumulated accuracy over number of images processed. The accumulated accuracy measurement calculates the average accuracy of all data points up until the given point. Accuracy was chosen as the performance metric because we have a binary classification problem where the two classes are equally important and equally represented in the dataset. The figure shows results from each subsession for both the experiments with human participants and the two simulated baselines. The results are averaged over all data streams with the given combination of step and strategy and over all participants of the experiments and the shaded area is the one-standard error.

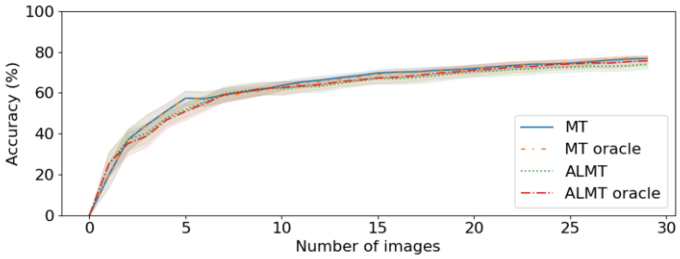
Apart from the results obtained during the learning phase through accumulated accuracy, the resulting ML models were also evaluated on a test set. The test set contained 200 images from the dataset, 100 of each class, which had been excluded from the training process. Figure 3 shows results where the averaged accuracy is plotted over number of data streams displayed during the given step. Also here both the experiments with human participants and the two simulated baselines are included.

Figure 4 displays how accurate the feedback from the human user was, given the number of images processed for each subsession. The results are averaged over all data streams with the given combination of step and strategy. The average number of images displayed for the *ALMT strategy* was 62%. As discussed earlier, an off-the-shelf and popular active learning strategy was chosen, because the aim of this study is not to develop new strategies. However, there exists no active learning strategy that is optimal for all problems and sometimes a random strategy is equally good or even better. To study how well the active learning strategy chose which images to not display, the accuracy of the predictions for the images displayed and not displayed was calculated respectively. The final accumulated accuracy for the images the active learning algorithm decided to not show was on average 85 %, which can be compared to the average final accumulated accuracy of the images that were shown of 64%.

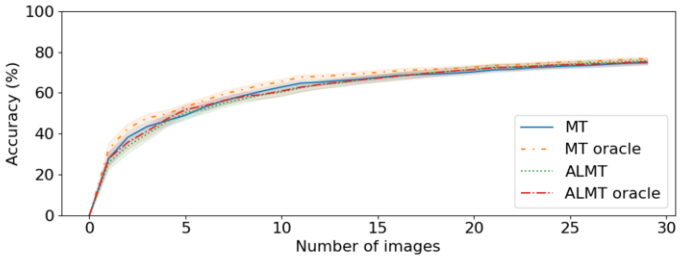
5. Discussion

When there are few data streams (and low workload), the two strategies perform equal overall, or the *MT strategy* performs slightly better. When the number of data streams are

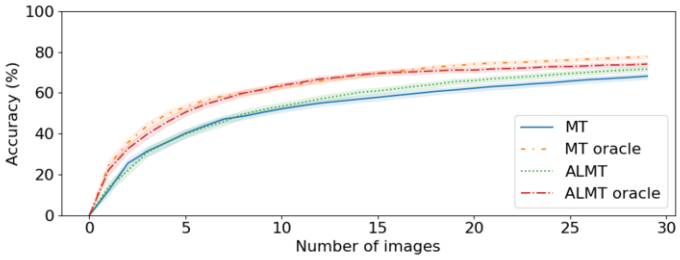
[3]The dataset can be found at https://www.kaggle.com/c/dogs-vs-cats/overview



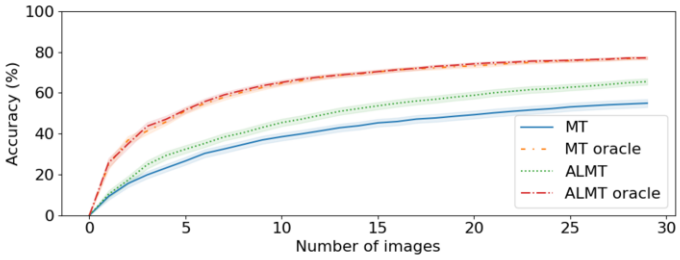
(a) 1 image/data stream



(b) 3 images/data streams



(c) 6 images/data streams



(d) 9 images/data streams

Figure 2. Average accuracy over number of images processed for the four steps with increasing number of data streams.

increased however, there is an increased performance for the *ALMT strategy* compared to the *MT strategy*. This can be seen in Figure 2 for the accumulated accuracy, as the

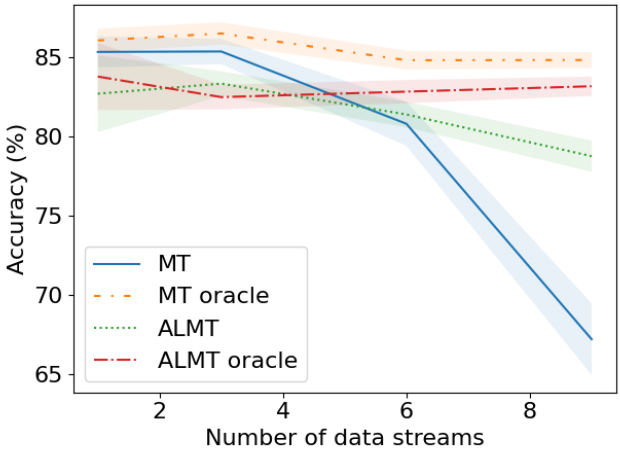
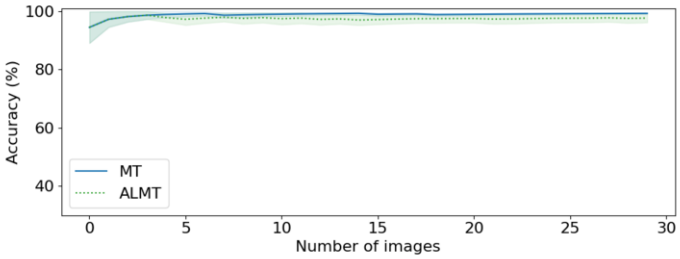


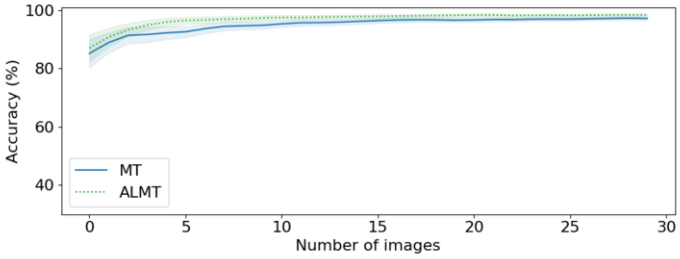
Figure 3. Accuracy on test set over number of images processed.

difference between the strategies increases when the number of data streams increases. It is even more evident in the evaluation using the test set. Figure 3, shows that while the performance of the *MT strategy* steadily decreases, the performance of the *ALMT strategy* is relatively steady with an increased number of data streams. As the number of data streams increases, the uncertainty caused by human factors also increases, leading to a decreased quality of the labels provided. With a larger number of images to label in the same time frame, the cognitive workload will increase and there will not be enough time consider to all images. This can also be seen in Figure 4, displaying the accuracy of the human feedback over time for the different number of data streams. For one and three data streams, the two strategies are relatively similar. The *MT strategy* is actually slightly better at the beginning, which is especially visible in Figure 3. Because a human user typically has time to annotate all images at these steps it might only be a disadvantage to not utilize the full cognitive capacity of the participant. Still, the difference between the two strategies is not large. For six images, the *ALMT strategy* is worse in the very beginning, but improves over time. As all images are not displayed all the time, the human can focus on the ones that are currently shown. Even though the accuracy is not perfect at the beginning, the performance can gradually over time improve as the ML algorithm improves and starts to pre-select correct labels and the human can focus on the inaccurate labels. This is in comparison to the *MT strategy*, where performance does not increase in the same way. For nine images the accuracy does not reach the same level, even for the *ALMT strategy*, but it is still better than the *MT strategy* and results in a better performance of the classifications.

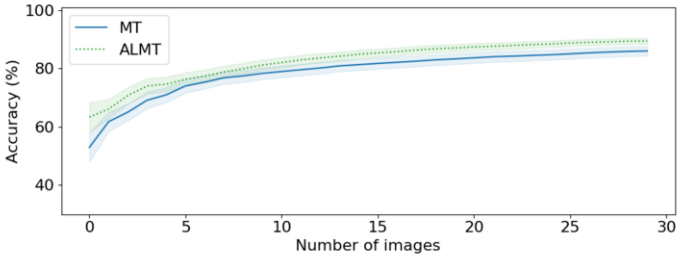
The baseline strategies were added to the experiment to represent how interaction from the human user often is simulated in interactive ML experiments without human participants, i.e. where the uncertainty of human factors are not included. In Figure 2, the baselines are similar to the other strategies when there is one image or when there are three images displayed in parallel. At six images in parallel, the baselines perform better especially at the start, but the other strategies catch up towards the end of the subsessions. For nine images, there is a distinction not only between the *ALMT strategy* and the *MT*



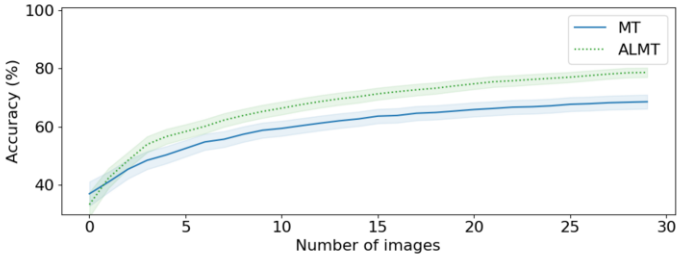
(a) 1 image/data stream



(b) 3 images/data streams



(c) 6 images/data streams



(d) 9 images/data streams

Figure 4. Average accuracy of human feedback over number of images processed for the four steps with increasing number of data streams.

strategy, but also between them and their respective baselines. This is supported by the results displayed in Figure 4, where the accuracy of the human feedback is relatively good

at one data stream and three data streams, but starts to drop at six data streams and gets even worse for nine data streams. The result from the baselines were not included in this figure, as it would always have been at 100 % accuracy.

As with most image processing applications the results can be used in an unethical way in e.g. surveillance applications, data collection etc. By utilizing online learning techniques and the streaming manner in which the data arrives, data do not need to be stored, which is more protective of privacy. Furthermore, we propose a solution where both the human and the machine cooperate, i.e. neither one is the only part with all control. If the human has all control, there is a risk of human error or even malicious intent from unethical decisions. If the machine has all control, there is risk of the model being skewed, resulting in e.g. discriminatory consequences. By giving control to both human and machine, these risks are reduced, yet should always be considered in applications.

It is understandable that human participants are not always included in interactive ML experiments, as simulated experiments are easier to conduct and the knowledge gained from such experiments can still be valuable. However, in a significant amount of previous work in interactive ML, the way in which the uncertainty caused by human factors can affect results or the interactive behaviour is not even discussed. The experiments described in this paper show that this might lead to inaccurate conclusions.

6. Conclusions and future work

In this work, we presented the results from interactive online ML experiments with human participants. Different interactive learning strategies were compared on how they affect the performance of the ML algorithm and the uncertainty caused by human factors from increased cognitive workload, e.g. providing incorrect feedback. The experiments with human participants were also compared to simulated experiments where the feedback was always correct. The results show that with increased cognitive workload, the *ALMT strategy*, combining active learning and machine teaching, performed better than the *MT strategy*, only including machine teaching. The difference between the results from the simulated experiments and results from experiments with human participants increased as well with higher cognitive workload. The results show that including human factors in interactive ML is important, especially in complex scenarios with high cognitive workload.

While we can draw conclusions from our results, we do acknowledge the limitations of this study. In this work we studied one problem scenario, including one dataset and one ML algorithm. To verify the findings and expand the knowledge on how human factors can affect interactive ML, further experiments would be useful. The classification task (cats vs. dogs) was kept on an easy level for the participants, but future work could explore different datasets where the difficulty is varied. It would also be of interest to expand the experiments to include different types of ML algorithms and interactive learning strategies. Another interesting aspect for future work would be to increase the participants' cognitive workload by letting them solve another task at the same time, or in parallel to the annotation task. This would mean that the participants are not always focused on the annotation task as they are distracted. Most importantly, future work on interactive ML and similar areas should discuss and further explore how the uncertainty in human feedback can affect the learning process and performance.

References

[1] Settles B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences; 2009.

[2] Zhu X. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29; 2015. p. 4083–4087.

[3] Lughofer E. On-line active learning: a new paradigm to improve practical useability of data stream modeling methods. Information Sciences. 2017;415:356-76.

[4] Wickens CD, Goh J, Helleberg J, Horrey WJ, Talleur DA. Attentional models of multitask pilot performance using advanced display technology. Human factors. 2003;45(3):360-80.

[5] Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, Huang J, et al. Learning something from nothing: Leveraging implicit human feedback strategies. In: 23rd IEEE International Symposium on Robot and Human Interactive Communication; 2014. p. 607-12.

[6] MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, Roberts DL, et al. Interactive learning from policy-dependent human feedback. In: International Conference on Machine Learning; 2017. p. 2285-94.

[7] Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems. 2022.

[8] Tegen A, Davidsson P, Persson JA. The effects of reluctant and fallible users in interactive online machine learning. In: Interactive Adaptive Learning 2020, Ghent, Belgium, September 14th, 2020. CEUR Workshops; 2020. p. 55-71.

[9] Donmez P, Carbonell JG. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM; 2008. p. 619-28.

[10] Tegen A, Davidsson P, Persson JA. Active Learning and Machine Teaching for Online Learning: A Study of Attention and Labelling Cost. In: IEEE International Conference on Machine Learning and Applications (ICMLA); 2021. .

[11] Vembu S, Zilles S. Interactive learning from multiple noisy labels. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2016. p. 493-508.

[12] Yan S, Chaudhuri K, Javidi T. Active learning from noisy and abstention feedback. In: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE; 2015. p. 1352-7.

[13] Lughofer E. Single-pass active learning with conflict and ignorance. Evolving Systems. 2012;3(4):251-71.

[14] Krawczyk B. Active and adaptive ensemble learning for online activity recognition from data streams. Knowledge-Based Systems. 2017;138:69-78.

[15] Du J, Ling CX. Active learning with human-like noisy oracle. In: 2010 IEEE International Conference on Data Mining. IEEE; 2010. p. 797-802.

[16] Huang TK, Li L, Vartanian A, Amershi S, Zhu J. Active learning with oracle epiphany. Advances in neural information processing systems. 2016;29.

[17] Mac Aodha O, Su S, Chen Y, Perona P, Yue Y. Teaching categories to human learners with visual explanations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 3820-8.

[18] Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S. Counterfactual visual explanations. In: International Conference on Machine Learning. PMLR; 2019. p. 2376-84.

[19] Miu T, Missier P, Plötz T. Bootstrapping personalised human activity recognition models using online active learning. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. IEEE; 2015. p. 1138-47.

[20] Jauhri S, Celemin C, Kober J. Interactive imitation learning in state-space. arXiv preprint arXiv:200800524. 2020.

[21] He X, Chen H, An B. Learning behaviors with uncertain human feedback. In: Conference on Uncertainty in Artificial Intelligence. PMLR; 2020. p. 131-40.

[22] Collins KM, Bhatt U, Weller A. Eliciting and learning with soft labels from every annotator. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 10; 2022. p. 40-52.

[23] Geng X. Label distribution learning. IEEE Transactions on Knowledge and Data Engineering. 2016;28(7):1734-48.

[24] Halford M, Bolmier G, Sourty R, Vaysse R, Zouitine A. creme, a Python library for online machine learning; 2019. Available from: https://github.com/MaxHalford/creme.

[25] Yang Y, Loog M. A benchmark and comparison of active learning for logistic regression. Pattern Recognition. 2018;83:401-15.
[26] Žliobaitė I, Bifet A, Pfahringer B, Holmes G. Active learning with drifting streaming data. IEEE transactions on neural networks and learning systems. 2013;25(1):27-39.

A. Appendix

The text for instructions to the participants are included in this appendix. Each subsection presents the text displayed at that step.

A.1. Welcome text and step 1

Thank you for agreeing to participate in this experiment on interactive machine learning!

During the experiment you will need to give your full attention to the task at hand. The data collected is done anonymously. Because of this, it is important that you don't close the window (or tab) until the experiment is completely finished.

Your task is to teach the system to recognize cats and dogs. You will be shown images and are asked to tell the system whether the images contain cats or dogs by clicking a button. In some cases, the system has already made a guess and then one button is preselected. If you think that the guess is correct, you do not have to do anything.

Please use a web browser on a computer and not your mobile phone for these experiments. The experiment consists of different parts. After each part there is a break before the next part starts. For each part the number of images will increase. Even if you do not have time to annotate all, please continue. The total time of the experiment will be about 20-25 minutes.

In the first part of the experiment you will see 1 image at a time and you will have 3 seconds to provide input.

If you have any questions regarding the experiment before you start or after, you are welcome to contact anonymous@anonymous.com[4].

A.2. Step 2

In the next part you will see 1 image at a time again, but this image will not be displayed all the time (when it is not displayed you cannot provide feedback).

A.3. Step 3

In the next part you will see 3 images at a time.

A.4. Step 4

In the next part you will see 3 images at a time again, but the images will not be displayed all the time (when it is not displayed you cannot provide feedback).

A.5. Step 5

In the next part you will see 6 images at a time.

[4] real name an mail concealed for review

A.6. Step 6

In the next part you will see 6 images at a time again, but the images will not be displayed all the time (when it is not displayed you cannot provide feedback).

A.7. Step 7

In the next part you will see 9 images at a time.

A.8. Step 8

In the next part you will see 9 images at a time again, but the images will not be displayed all the time (when it is not displayed you cannot provide feedback).

A.9. Final text

That was the last part, thank you for your participation in the experiments!