International Journal of Advanced Smart Convergence Vol.9 No.2 90-104 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.2.90

IJASC 20-2-11

Understanding Interactive and Explainable Feedback for Supporting Non-Experts with Data Preparation for Building a Deep Learning Model

Yeonji Kim[1†], Kyungyeon Lee[1†], Uran Oh[2*]

[1]*Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea*
[2]*Assistant Professor, Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea*
E-mail: yeonjikim@ewhain.net, ruddus716@ewhain.net, uran.oh@ewha.ac.kr

Abstract

It is difficult for non-experts to build machine learning (ML) models at the level that satisfies their needs. Deep learning models are even more challenging because it is unclear how to improve the model, and a trial-and-error approach is not feasible since training these models are time-consuming. To assist these novice users, we examined how interactive and explainable feedback while training a deep learning network can contribute to model performance and users' satisfaction, focusing on the data preparation process. We conducted a user study with 31 participants without expertise, where they were asked to improve the accuracy of a deep learning model, varying feedback conditions. While no significant performance gain was observed, we identified potential barriers during the process and found that interactive and explainable feedback provide complementary benefits for improving users' understanding of ML. We conclude with implications for designing an interface for building ML models for novice users.

Keywords: *End-user Machine Learning, Interactivity, Explainability*

1. Introduction

Studies have shown that it is not straightforward for nonexperts to build ML models without prior knowledge [1, 2] and that they cannot build ML models in an efficient way [3, 4]. For instance, Yang et al. [4] conducted an interview study to understand how non-experts build ML solutions for themselves in real life and revealed several pitfalls. The major problem was that non-experts rarely tried to understand the internal mechanism of learning algorithms, hence had trouble improving the performance or even gave up their tasks. Also, most of them did not consider the overfitting problem, ending up having poor accuracy on a new dataset.

Based on these observations, researchers have worked on a variety of approaches to assist non-experts with building ML models. One of them is to provide interactivity in the user interface [5-7]. For example, Amershi et al. [5] showed that adopting interactive feedback while training an ML system enables people to use it naturally when training ML models and improves the quality of users' models. Also, Fiebrink et al. [7] found

Manuscript Received: April. 10, 2020 / Revised: April. 13, 2020 / Accepted: April. 17, 2020
*Corresponding Author: uran.oh@ewha.ac.kr †Both authors have contributed equally
Tel: +82-2-3277-6896
Assistant Professor, Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea

that with an interactive ML system, users not only made relevant judgments on evaluating their models to improve the trained models but also learned what types of models are easy to build and used this information to modify their trained models. Others showed that providing explanations on how ML systems operate can help users to build mental models quickly [1, 2, 8]. Kulesza et al. [2], for example, provided users a music recommender with a tutorial discussing how the recommender works and how various feedback controls would affect. As a result, participants responded positively to learning details about the system, and they were more satisfied with the recommender's output. Likewise, Hitron et al. [8] uncovered the process of ML system to children and discovered that this enhanced children's understanding of basic ML concepts. Kulesza et al. [1] also built a text classification tool that provides explainable feedback to participants, which increased their understanding of the ML system and allowed them to correct the system's wrong results (e.g., and improve the model performance by adjusting importance of features (words) with incorrect weight for labels) with fewer actions than participants using a traditional learning tool. While these studies found that interactive and explainable feedback are effective for nonexpert users who wish to train ML models, none of them explored how to provide appropriate feedback to novice users without expertise when building deep learning models, which is much more challenging because it is difficult to generate explanations on why the model behaves in a certain way, compared to traditional machine learning model. Moreover, it is not feasible to provide immediate feedback due to its extremely long training time.

To assess the effectiveness of interactive and explainable feedback when building a deep learning model for novice users, we conducted a user study with 31 participants who have no expert knowledge in ML, focusing on data preparation process where they were asked to refine the training dataset to improve the accuracy of an image-based dog breed classification model with and without interactive and explainable feedback. Based on the analysis of the accuracy of models built by our participants and their subjective responses, we confirmed that participants were able to improve the model performance similar to the accuracy achieved by experts. Although there was no significant difference across different feedback conditions, findings revealed that participants' understanding of building ML models increased the most when both interactive and explainable feedback were provided.

The contributions of this work are: (i) the assessment of how adopting interactive and explainable feedback during data preparation help improving people's understanding of ML systems, and (ii) the identification of the potential barriers of building ML models for non-experts.

2. Related Work

2.1 Interactive Feedback for Building ML Models

Since Fails and Olsen [9] first introduced the term interactive machine learning, which introduced iterative cycles help users quickly correct the result by giving feedback back to the system, many researchers have proposed interactive ML systems for creating ML models based on user feedback [5, 7, 8, 10-13, 29-30]. For example, Fogarty et al. [12] proposed CueFlik that provides users with best- and worst-matching examples that enable users to assess the quality of the model quickly so that they can build better models, rather than simply presenting the results of all data as in traditional machine learning. Likewise, Fiebrink et al. [7] found that iterative training of an ML model helps users to make relevant judgments on evaluating their models and achieve better performance as a result. Moreover, they found that the most frequently observed behavior for improving the model performance was modifying their training set among changing the features, learning parameters of an algorithm, or the algorithm itself. Inspired by the finding that interactive feedback in training ML models allows users to quickly examine the impact of their actions and adapt subsequent inputs to obtain

desired behavior [14], we examine the effectiveness of providing interactivity (i.e., rapid updates) for assisting non-expert users with building ML system by presenting the effect of their behavior so that users can make appropriate adjustment in the next cycle.

2.2 Explaining the Performance of ML Models

Another approach many researchers have explored for assisting non-experts in building ML models is uncovering the process of training or outcomes of an ML system with explanations [1, 2, 8, 15] presented in forms that are easy to understand for human, which is called explainable machine learning [16]. In gesture recognition, for instance, Hitron et al. [8] discovered that children's understanding of the basic concepts of ML, such as sample size and sample versatility, can be improved when they performed data labeling and model evaluation tasks for training the ML system. In the field of image recognition, methods for medical decision making have been studied with explainable AI [17-19, 27-28]. EluciDebug [1], another example for text data, was proposed as a system that introduces explanatory debugging to give users an explanation of how predictions were made. They confirmed that their participants could correct the system's mistakes, and the improvement of model accuracy was doubled at most compared to the accuracy when using a traditional learning system. However, studies [1, 8, 20] have investigated how interactive and explainable feedback work together in affecting end-users for building deep learning models, not how each feedback affect. Moreover, the barriers non-experts face when building their deep learning models have not been thoroughly enough discussed. Similar to EluciDebug, our goal is to examine how explainability during ML model training can benefit non-experts for constructing robust deep learning models while preventing overfitting problems.

3. User Study

To investigate the effectiveness of interactivity and explainability and to identify potential barriers that non-experts would experience while building their ML systems, we conducted a study with 31 participants who had little or no ML-related experience where they were asked to use our custom user interface designed for this study to train an image-based ML model. All of them were compensated for their participation.

3.1 Experimental Conditions

We had four conditions for this study varying types of feedback (i.e., interactive and explainable feedback) provided to users, as shown in Table 1. Interactive feedback allows users to interactively examine the effect of their current decision, and help them to refine their dataset for the next training phase. It includes two features: red-bordered box and interactive box. The red-bordered box was designed to inform misclassified images to users by having a bounding box with red borderlines (see Figure 1b1). Meanwhile, the interactive box shows the classification result with confidences of the current ML model instantly for a selected image, as shown in Figure 1c1 (e.g., 100.00% Yorkshire, 0.00% Bulldog, and 0.00% Labrador).

Table 1. Types of feedback provided for each of the four conditions examined in this study.

Feedback Condition	Default	Interactive	Explainable	Full
Interactive	no	yes	no	yes
Explainable	no	no	yes	yes

On the other hand, explainable feedback was designed to provide explanations of how predicted results are generated by the current model in understandable forms to novices. It consists of three features: activation map,

confidence table, and model description. The activation map was introduced to make Convolutional Neural Network (CNN)-based models more explainable as it visualizes the areas of input that are important for predictions as proposed in [21]. As shown in Figure 1c2, this map shows how an image is evaluated in the pre-trained ImageNet model [22]. The image in Figure 1c2 is highlighted on the face of a dog even though the dog and a person are together, suggesting that the current model recognized the image as a dog. The confidence table, on the other hand, informs users about Top 5 recognition results with the highest confidences based on the same model (see Figure 1c3) to convey the estimated performance of their model. Lastly, the model description is provided after users train their model — it specifies the number of misclassified images in each class.

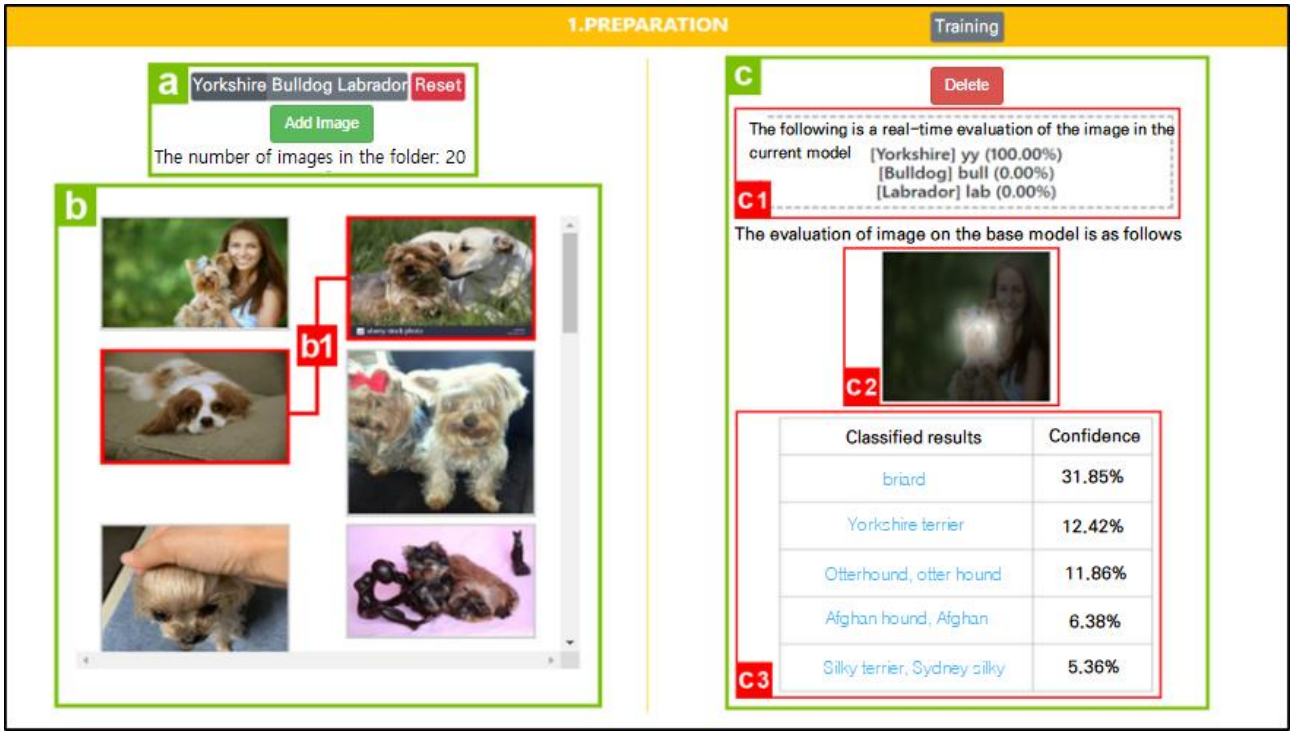


Figure 1. A screenshot example of the user interface for one of the feedback condition (full) custom-built for this study when the first image is selected. It supports data preparation for building a deep learning model by allowing participants to (a) import data for each of the predefined classes, (b) review uploaded images per class, and (c) receive estimated results per image before starting the actual training. Red labels indicate different feedback features: (b1) red-bordered box, (c1) interactive box, (c2) activation map, and (c3) confidence table.

3.2 Participants

We recruited participants through a local online community, university board, and word of mouth. Thirty-four participants applied, but two were excluded as they reported had studied AI or ML before or have worked on related projects (6 or 7 in Table 2) as our target participants were people who are non-experts in ML. The rest 32 participants were distributed between 1 and 5 (see Table 2) and reported 2.3 on average. Table 3 shows the demographics of 31 participants after removing one outlier, where the accuracy performance of the model that the participant built was beyond 3 standard deviations of the mean, throughout the study. Their age range

was between 18 to 56, and each participant was randomly assigned to one of the four conditions.

3.3 Apparatus

To observe how non-experts create ML systems, we implemented a web-based system that allows users to iteratively import, review, refine the training dataset, and evaluate their trained models to build a deep learning network for image classification. We designed the interface with a focus on data labeling and model evaluation out of the four building blocks of ML mentioned by Meng et al. [23], that were found to be more accessible to and less complicated for novices than other two blocks (i.e., extracting features or selecting models).

Table 2. Descriptions for collecting prior knowledge of AI or ML on a 7-point scale.

Score	Description
7	I can build AI or ML systems.
6	I have studied AI or ML through lectures or books.
5	I can explain in an abstract way how AI or ML works.
4	I can list examples of AI or ML applications.
3	I know the definition of the terms.
2	I have heard of the terms, but I know them as superficial.
1	I have never heard of the terms AI and ML.

Table 3. Participants' gender distribution, average age, and their prior knowledge about ML per condition. Standard deviations are presented in parenthesis.

Condition	Gender	Age	ML-Knowledge
Default (D1-D8)	3 M, 5 F	24.2 (4.4)	2.1 (1.4)
Interactive (I1-I7)	3 M, 4 F	24.6 (4.5)	2.9 (1.2)
Explainable (E1-E8)	3 M, 5 F	26.4 (12.3)	1.9 (0.6)
Full (F1-F8)	4 M, 4 F	25.8 (7.6)	2.5 (0.8)

Our apparatus consists of two ML models; the primary is MobileNetV2 [24] that is retrainable and additionally VGG16 [25]. We used MobileNetV2 as the primary model and applied transfer learning to create image classifiers and generate feedback as quickly as possible by training only a part of the pre-built ML model with smaller number of classes — we trained only the top-layer of MobileNetV2, and the classes we used were Bulldog, Yorkshire, and Labrador, which were selected from 1,000 different object classes on ImageNet [22]; training only the top-layer of MobileNetV2 model with modified dataset took 5-10 seconds and the accuracy result of the model was updated in real-time. Using this model, we implemented the interactive box, red-bordered box, and model description features. We also had another VGG16 model to visualize the activation map and show the confidence table since generating the activation map is time-consuming (50-60 seconds per image). While the MobileNetV2 model for providing interactive feedback was updated every time the model gets re-trained, parameters for VGG16 were pre-trained to shorten the time taken to train the model and produce an activation map and a confidence table. To train these two models, we used four Intel Skylake processors, 15GB memory and NVIDIA Tesla K80 GPU for hardware, and used Python 3.7 with Tensorflow for implementation.

3.4 Procedure

The study was conducted as a one-hour single-session study consisting of a pretest, a data preparation task for building a deep learning model, and a posttest followed by an interview. As a between-subject test, participants were assigned to one of the four conditions for the data preparation task. Participants' demographic information and their prior knowledge about artificial intelligence and ML on a 7-point scale were collected at the beginning of the study.

Pretest. Before performing the data preparation task for building a model, we presented two ML application scenarios (i.e., image classification, speech recognition), and asked participants to explain what kind of dataset would be needed to make such ML systems with high accuracy for each.

Data Preparation Task. For the main task, we first played video tutorials that explain the basic concepts of ML such as definitions of ML and class in a classification problem, the difference between training and test data, and how the accuracy of an ML model is calculated as well as the instructions on how to use our custom interface for performing the task. Before performing the actual task, participants were allowed to ask questions, if any. Then they were given 3 minutes to explore the interface themselves, such as importing and deleting images and training a model. After this brief introduction, participants were instructed to add or remove images per class using the web interface to improve the accuracy of the initial model as high as possible within 30 minutes. They were allowed to train and check the current model's accuracy up to three times, which can be done by clicking the "Training" button on the top right of the interface (see Figure 1). Once the training is completed, participants were able to check the results such as training accuracy and validation accuracy on a new page. To understand participants' mental model, we used a think-aloud protocol throughout the task and collected participants' feedback before and after checking the training results by asking how they prepared their data and what they think about the results.

Posttest and Interview. After the task completion, we had a posttest session where we asked participants the same questions again with the two scenarios from our pretest to examine if participants' understanding of ML systems has improved after performing our task. Then we had a wrap-up interview to understand the potential barriers for preparing a dataset for an ML model and how the presented interactive or explainable feedback can be used to ease the training process. We also asked participants about their satisfaction, usefulness, the difficulty of each feature they used as well as perceived task load (NASA-TLX) [26] on a 7-point scale.

4. Findings

Here we present the potential effects of interactive and explainable feedback in terms of accuracy, the understanding of ML, and subjective responses collected from participants for each feedback feature.

4.1 The Impacts of Feedback Conditions on Model Accuracy

To investigate to what extent participants can improve the accuracy of their models with and without interactive and explainable feedback, we assessed both leave-one-out (LOO) validation accuracy, where 10% randomly selected images from the entire data were used for testing and the rest 90% for training, and advanced validation accuracy, where the test data were composed of 15 images, 5 for each class, selected by researchers with the intention to include images that are difficult to classify accurately, as shown in Figure 2. Training accuracy results were mostly 100% and thus excluded in the analysis. The results are shown in Figure 3 and Figure 4, and each refers to the highest accuracy achieved during the task.

Initial vs. Best Accuracy. The LOO validation accuracy was compared with the accuracy of the initial model trained with the default dataset at the beginning of the task to confirm if non-expert users can improve

the performance of an ML model with and without interactive or explainable feedback. As a result, we found that the average accuracy of the model was increased across all conditions compared to that of the initial model; the LOO validation accuracy increased by about 35% on average ranged from 32.0% to 36.6%. Likewise, the advanced validation accuracy was also increased up to 19.1%, with a minimum of 16.0%. The differences between the initial accuracy and the best accuracy were found to be statistically significant with paired t-tests for both LOO validation accuracy and advanced validation accuracy ($p < .001$ for all, except for advanced validation accuracy for full condition where $p = .03$). While we expected to observe the effects of feedback conditions on model performance, no significant difference was found with one-way ANOVA tests.

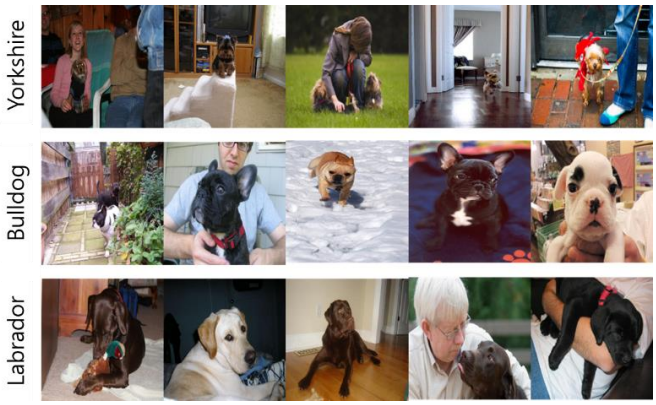


Figure 2. Fifteen test images used for computing advanced validation accuracy for each class.

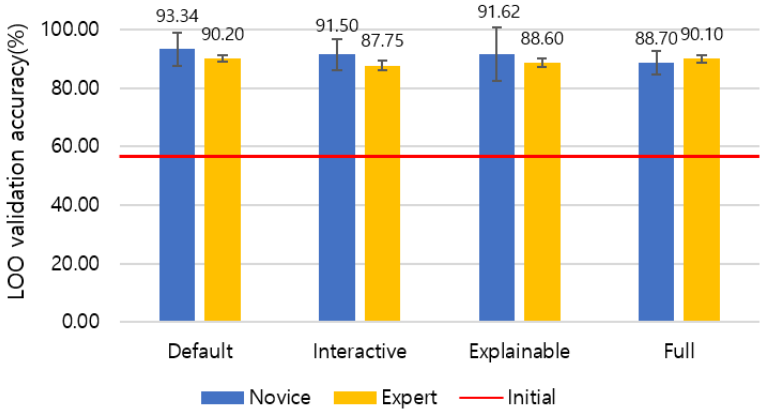


Figure 3. Average LOO validation accuracy of models trained by participants (novice) and experts per condition compared to the LOO validation accuracy of the initial model (56.7%) for all conditions. Error bars indicate standard deviations.

Experts vs. Novice. In addition, as a secondary analysis, we recruited three ML experts who had either taken ML courses or worked on an ML project to confirm the upper limit of the model's accuracy. Each of them was asked to train ML models with all four conditions where the presented order was randomized. Paired t-tests with Bonferroni corrections comparing the accuracy of experts' models and that of participants' model revealed that the best accuracy achieved by novice participants was not that significantly lower than that by

experts.

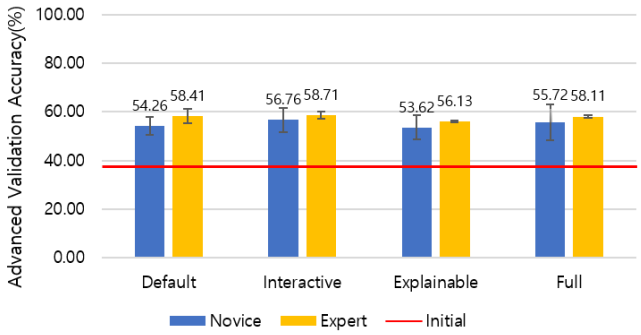


Figure 4. Average advanced validation accuracy of the initial model (37.7%) and models trained by participants (novice) and experts per condition. Error bars indicate standard deviations.

4.2 The Understanding of Building a Better ML Model

To examine if participants' understanding of constructing ML models has been improved after performing the task, one researcher graded participants' pretest and posttest responses where the participants' group was blinded. For each scenario per test, responses of participants were given a score between 0 and 4 depending on how much they are aware of the fact that increasing volume and variety of training data can help to improve the accuracy (2 scores each per metric). The maximum score a participant could get from a test was 4 (2 metrics × 2 points). The metric of volume, for instance, a participant would receive 0 score if they did not mention about increasing the number of training images at all, 1 if the participant mentioned the concept of the quantity of the training data vaguely, and 2 if the concept is clearly explained. As a result, we found that participants' understanding of how data should be prepared to improve the accuracy of an ML model has been improved for all conditions, as shown in Figure 5. Participants in full condition showed the biggest improvements on average, followed by explainable, default, and interactive conditions. Unlike the model accuracy, the difference between pretest and posttest scores were found to be statistically significant with Wilcoxon Signed Rank tests for full and explainable conditions ($p = .027$ and .020, respectively). Interestingly, while the performed task is about image classification, the scores suggest that building an image classification model can also help to deepen the understanding of the speech recognition model.

4.3 The Use and Subjective Assessments of Feedback Features

We further assessed how each feature for each feedback type assisted participants in preparing a dataset for building ML models by analyzing participants' subjective feedback such as contribution (i.e., how much this feature has contributed to their decision when preparing the data) and usefulness at the end of the study; see Figure 6. In addition, we asked them to vote for the type of feedback that helped them the most when performing the task (see Figure 7) and their reasons.

Red-bordered box. While the difference between red-bordered box and interactive box was not significant in terms of subjective ratings, this feature was considered to be most helpful for improving the model accuracy by 7 participants out of 15 across two conditions who had the access to this feature; all participants except one (I4) from interactive condition and one participant (F5) from full condition. Some participants (I2, I4, I6, F5) mentioned that this feature was intuitive since it lets them know which images were misclassified. Related,

half of the participants from default condition (N = 4 out of 8) wished to know which images are good or bad efficiently as they found it tedious to check every image to understand why the accuracy was dropped or increased after training. D3 said, "If a person has to check every image when adding data, it would be very hard to check all of them when there is a lot of data." Similarly, I2 said that the red-bordered box was helpful in finding problematic images quickly and saving time as a result. Indeed, we observed that more than half of the participants (N = 8) deleted the images quickly as soon as they saw red-bordered images without a second thought. Yet, other 5 participants tried to understand why some images were misclassified and deleted them if they thought the images were inappropriate for training, and the rest two ignored the flagged images and continued refining other data instead, hoping it would help to improve the accuracy.

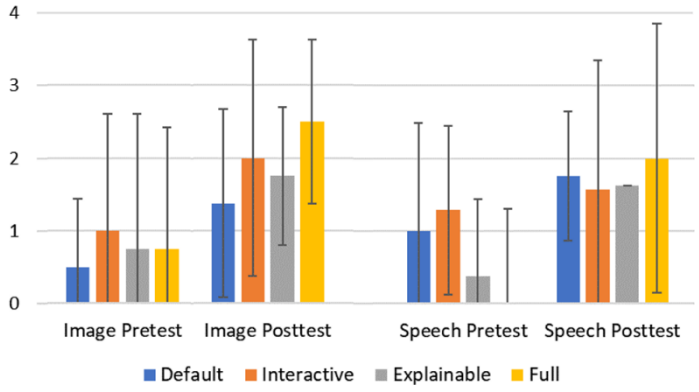
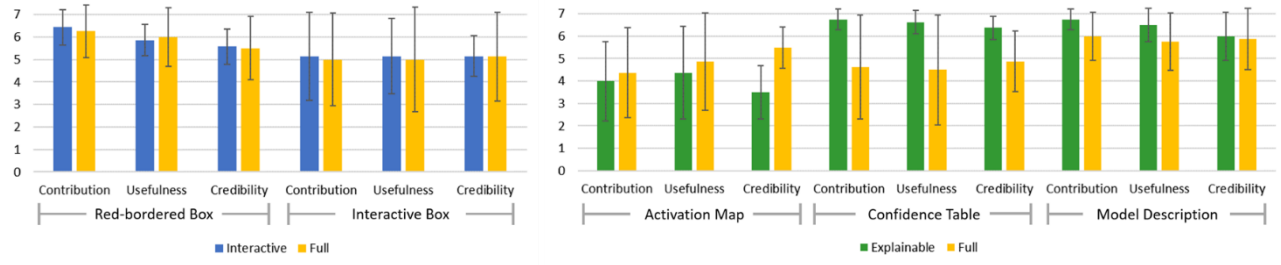


Figure 5. Average pretest and posttest scores per condition for both image classification and speech recognition where maximum score is 4. Error bars indicate standard deviations.



(a) Interactive Feedback (b) Explainable Feedback

Figure 6. Average assessment of various types of (a) interactive and (b) explainable feedback per condition on contribution, usefulness, and credibility, where 7 is best. Error bars indicate standard deviations.

Interactive Box. Only 4 out of 15 participants considered the interactive box was the most helpful feedback as they could check quantified results on how each image has been classified with a confidence score in percentages generated from the model. When participants were asked if they prioritized their own judgment over the provided feedback or vice versa, 3 participants from interactive condition (I2, I4, I7) and four from full condition (F1, F2, F3, F5) responded that they fully trusted the information in the box and adjusted their dataset accordingly. However, some (I1, F7, F8) used it as a reference for making their own judgments on if the selected image is good, while other four (I3, I5, F4, F6) did not use it at all since the presented information

was the result based on the last-trained model, or because they were busy looking at information generated by other features. The subjective ratings were almost the same for both interactive and full conditions, as shown in Figure 6a.

Activation Map. Three full participants utilized the activation map to interpret training results (F2, F7, F8). For example, F2 said, "If the map highlights a person in the image, not the dog, then I assumed that having a person in an image caused misclassification results." Participants from full condition gave higher ratings than those from explainable condition in general (Figure 6b); the difference was significant for credibility ($U = 6.5$, $p = .004$) with Mann-Whitney U tests. Three explainable participants (E1, E3, E6) replied that they did not refer to the map at all. E3 described, "I was able to perform the task without having to look at the image (activation map) because I've got accurate numbers from the (confidence) table."

Confidence Table. The confidence table was evaluated significantly higher for contribution, usefulness and credibility for explainable condition than those for full condition ($N = 8$ each) (see Figure 6b) with Mann-Whitney U tests ($U = 12.0, 14.0, 7.5$; $p = .014, .025, .003$, respectively). As shown in Figure 7, the majority of explainable participants ($N = 6$ out of 8) reported that they used the table the most as they can choose images if and only if its the object with the highest confidence is similar to the desired label (class name) along with its confidence value or if the desired label is ranked high. E6, for example, deleted images that showed confidence below 60% in the table. On the other hand, full participants, who also had access to interactive feedback, reported that they did not care much about the activation map nor the confidence table because these feedback were the results of pre-trained ML model for a quick preview which does not affect their model, and thus they felt the feedback was irrelevant to their model.

Model Description. Six participants reported that the model description, which informs users about the number of misclassified images per class, was helpful as it guided them in which class they should pay attention to when refining their dataset. For example, if the description notified that many Labrador images were misclassified, participants would mainly focus on improving the dataset of this class by removing problematic images and/or adding new Labrador images. We observed that full participants tend to use this feature with the red-bordered box; they first read the descriptions to understand which class they have misclassified images and then used the red-bordered box to handle those images. On the other hand, one participant (F7) rarely read the description because he could examine misclassified images directly using the red bordered box.

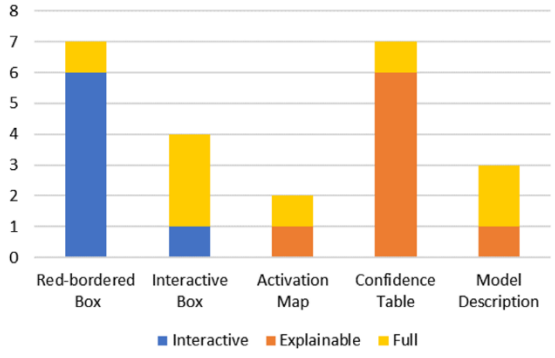


Figure 7. Vote counts for the feature that affected the task performance the most.

4.4 Perceived Task Loads and Observed Behaviors of Novices

Perceived Task Load. We used NASA-TLX questionnaires [26] to assess participants' perceived task load for performing the task of building an image classifier, and the result is shown in Figure 8. Similar to the model

accuracy, the responses were not significantly different across conditions when tested with one-way ANOVA. Also, when compared to the responses collected from experts, we have confirmed that mental demand is slightly higher for participants. On the other hand, other responses were similar, including effort. This suggests that novice users' perceived task load is not too different from experts except for mental demand.

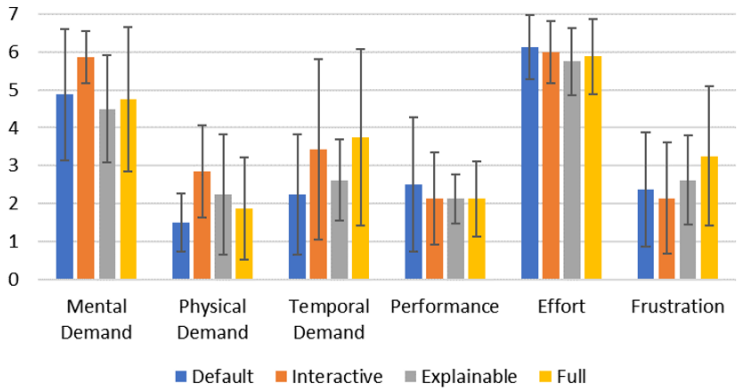


Figure 8. Average NASA-TLX (7-scale) for all four conditions, where 1 means low and 7 means high for all metrics except for performance, where 1 means good and 7 means poor. Error bars indicated standard deviations.

Amount of Data Novices Prepared. The number of images that participants used for training varied. More than half of the participants started with 20-40 images per class and maintained the number throughout the task (N = 17) while the other 6 participants gradually increase or decrease the number of training data. This is contrary to the behavior of experts who started training the model with a large volume of data. In the study, only the remaining eight added images as many as possible from the start, assuming that adding many images would increase the accuracy, especially for F8, who related the performance with the word 'big data.'

Tendency to Over-clean the Data. To understand users' mental models, we observed how participants built their ML models while they are performing the task. As a result, we found that participants had a tendency to clean the data as much as possible without knowing that this can cause an overfitting problem. To be specific, participants tended to remove images if dogs with different breeds or people are included for the selected class even if the correct dog is present. This behavior was more frequently observed from the participants in default condition, which had the minimum feedback without interactive nor explainable feedback. Moreover, when adding new pictures, most participants added images if and only if the images are representative of that class, such as images where the frontal face of a single dog with the correct breed is visible. This indicates that non-expert users believe that having only the representative data can lead to higher accuracy without knowing that their model would not perform well on real-world data with large variations (noises).

5. Discussion

5.1 Discrepancy Between Understanding of ML and Accuracy

Although we could not confirm if interactive and explainable feedback help increase the accuracy, the difference between the pretest and posttest scores suggests that the understanding of constructing better ML systems increases with explainable feedback (with or without interactive feedback). In other words, despite the increase in understanding of building better ML models the model performance showed no significant difference between conditions. This could be due to how the pre-trained model (MobileNetV2) was trained.

That is, as a transfer learning model, it is already well-trained so that it is not too difficult to improve the accuracy of the model by making modifications to the dataset regardless of the feedback conditions. Another possible explanation could be related to the choice of our task, which consists of data preparation and model evaluation. As these are relatively easy for novice users compared to the other two blocks of building a model — feature extraction and model selection [23], the feedback we provided might not have helped as expected. Further studies are needed to examine which feedback types are needed to support each block of an ML model building process.

5.2 Understanding of the Volume and Variety of Training Data

While all of the experts used almost all of the available training data, only about 25% of the participants attempted to do the same across conditions. Instead, most of the participants showed a pattern of keeping the number of images similar to the initial state when preparing a dataset for training an ML model. We also identified that the most frequent and popular ways to improve the accuracy among novices during the data preparation process for building ML models are filtering out misclassified images (e.g., deleting misclassified images that are classified as Labrador in a Bulldog class) and correctly classified images but with low confidence rather than including diverse images. This may be due to a lack of understanding that training with a large number of images with varieties can increase the accuracy of their model.

5.3 Trade-offs Between Feedback Types

Our findings imply that there are trade-offs between interactive feedback and explainable feedback. While interactive feedback has the advantage of providing immediate feedback to users, explainable feedback was found to be more beneficial for deepening users' understanding of building a better ML system. In addition, based on participants' use of both model description and red-bordered box, findings revealed that these two feedback can be complementary to one another. For instance, after training, users can read the model description to find the most problematic class and then use the red-bordered box to remove unwanted images within that class. Also, reviewing the activation map allows users to explore what features of images are considered as important so that they can try adding a more variety of data, which would help to prevent overfitting problems. For this reason, it may be desirable to provide both interactive and explainable feedback for users so that users can selectively utilize different feedback as needed.

5.4 Non-Experts' Misconceptions of Machine Learning Models

Many participants seemed to form their mental model of ML models for achieving high accuracy by checking how the modifications they made to their dataset changed the accuracy. However, before or during the process, we have observed participants' misconceptions related to ML, believing that their certain behavior was crucial in increasing accuracy while it might not be the main reason. For example, several participants added many images where the frontal face of a dog is clearly visible and believed that this is the main reason for the accuracy improvements while other behaviors they performed such as deleting wrong-labeled images (e.g., Bulldog image in Labrador class) could have a greater impact on the accuracy of their model. Without knowing exactly how to increase the accuracy, users may be led to believe that the consequences of their attempts are random or in misconceptions. Thus, it is essential to guide novices with proper feedback for improving the accuracy of ML models [4].

5.5 Risk of Providing Incomplete Feedback to Novice Users

In our study, non-expert participants tended to delete misclassified images without a doubt, assuming that

the feedback is always correct when it is only presenting the information based on estimated results, which could be wrong. Classifying dog breeds will not be a big problem if non-experts follow the system's wrong recommendation without confirmation, but it might be even dangerous in the case where accurate judgment is required, such as a medical decision. Thus, since non-experts may tend to over-trust ML system's feedback, it might be better not to provide feedback to non-experts with a warning that the feedback should be considered as a reference or a suggestion. Moreover, considering full participants' tendency to focus on single feedback that is perceived to be intuitive (i.e., red-bordered box) even when single/multiple feedback with rich information is provided such as relying more on red-bordered box although activation map provides more details, a system can also guide users to examine and utilize the multiple feedback provided with various aspects make their own judgment.

5.6 Feasibility of Providing Feedback for Educational Purpose

Our results showed that the understanding of ML concepts has increased with explainable feedback not only with image data but also with speech data. This suggests that a user interface for training a machine learning model with interactive and explainable feedback can also be used as an educational tool to convey the concept of ML with different types of data to users by doing. Considering that our participants felt that preparing the data for training a machine learning model was mentally demanding, providing appropriate guidance with feedback can also help users to learn how different quantities and variations of a dataset can affect the overall accuracy of their model as they go through the process.

6. Conclusion and Future Work

To support people with no expertise in ML for constructing deep learning models in the data preparation process, we provided interactive and explainable feedback to assist with the process. Based on the analysis of 31 participants' data collected from the user studies, we found that interactive and explainable feedback can help novice users to improve the accuracy of the model without having a significant gap compared to the accuracy trained by experts. Findings also revealed that interactive and explainable feedback are complementary to each other in increasing the understanding of constructing better ML models. Finally, we gained insights into how people map out their strategy to prepare their training data for their machine learning models, and what barriers they go through.

We hoped to continue exploring various feedback types to identify the most effective way to guide novice users for each of the phases for building a deep machine learning model in real-time with a larger sample size and longer exposure of the system. Based on the identified potential benefits and barriers of our feedback, we plan to extend our work to support various data types as well as more complex configuration of deep learning networks as future work.

References

[1] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in Proceedings of the 20th international conference on intelligent user interfaces. ACM, pp. 126–137, 2015. DOI: https://doi.org/10.1145/2678025.2701399

[2] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more?: the effects of mental model soundness on personalizing an intelligent agent," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1–10, 2012. DOI: https://doi.org/10.1145/2207676.2207678

[3] W. B. Knox and P. Stone, "Reinforcement learning from human reward: Discounting in episodic tasks," in 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. IEEE, pp. 878–885, 2012. DOI: https://doi.org/10.1109/roman.2012.6343862

[4] Q. Yang, J. Suh, N.-C. Chen, and G. Ramos, "Grounding interactive machine learning tool design in how non-experts actually build models," in Proceedings of the 2018 Designing Interactive Systems Conference. ACM, pp. 573–584, 2018. DOI: https://doi.org/10.1145/3196709.3196729

[5] S. Amershi, J. Fogarty, A. Kapoor, and D. Tan, "Examining multiple potential models in end-user interactive concept learning," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1357–1360, 2010. DOI: https://doi.org/10.1145/1753326.1753531

[6] ——, "Effective end-user interaction with machine learning," in Twenty- Fifth AAAI Conference on Artificial Intelligence, 2011. DOI:

[7] R. Fiebrink, P. R. Cook, and D. Trueman, "Human model evaluation in interactive supervised learning," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 147–156, 2011. DOI: https://doi.org/10.1145/1978942.1978965

[8] T. Hitron, Y. Orlev, I. Wald, A. Shamir, H. Erel, and O. Zuckerman, "Can children understand machine learning concepts?: The effect of uncovering black boxes," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI '19. New York, NY, USA: ACM, pp. 415:1–415:11, 2019. DOI: http://doi.acm.org/10.1145/3290605.3300645

[9] J. A. Fails and D. R. Olsen Jr, "Interactive machine learning," in Proceedings of the 8th international conference on Intelligent user interfaces. ACM, pp. 39–45, 2003. DOI: https://doi.org/10.1145/604045.604056

[10] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" Brain Informatics, vol. 3, no. 2, pp. 119–131, 2016. DOI: https://doi.org/10.1007/s40708-016-0042-6

[11] K. Patel, N. Bancroft, S. M. Drucker, J. Fogarty, A. J. Ko, and J. Landay, "Gestalt: integrated support for implementation and analysis in machine learning," in Proceedings of the 23nd annual ACM symposium on User interface software and technology. ACM, pp. 37–46, 2010. DOI: https://doi.org/10.1145/1866029.1866038

[12] J. Fogarty, D. Tan, A. Kapoor, and S. Winder, "Cueflik: interactive concept learning in image search," in Proceedings of the sigchi conference on human factors in computing systems. ACM, pp. 29–38, 2008. DOI: https://doi.org/10.1145/1357054.1357061

[13] Q. Yang, J. Zimmerman, A. Steinfeld, and A. Tomasic, "Planning adaptive mobile experiences when wireframing," in Proceedings of the 2016 ACM Conference on Designing Interactive Systems, pp. 565–576, 2016. DOI: https://doi.org/10.1145/2901790.2901858

[14] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," AI Magazine, vol. 35, no. 4, pp. 105–120, 2014. DOI: https://doi.org/10.1609/aimag.v35i4.2513

[15] S. L. Rosenthal and A. K. Dey, "Towards maximizing the accuracy of human-labeled sensor data," in Proceedings of the 15th international conference on Intelligent user interfaces. ACM, pp. 259–268, 2010. DOI: https://doi.org/10.1145/1719970.1720006

[16] D. Gunning, "Explainable artificial intelligence (xai)," Defense Advanced Research Projects Agency (DARPA), nd Web, vol. 2, 2017. *https://www.darpa.mil/attachments/XAIProgramUpdate.pdf*

[17] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe et al., "Human-centered tools for coping with imperfect algorithms during medical decision-making," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, p. 4, 2019. DOI: https://doi.org/10.1145/3290605.3300234

[18] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017. DOI: https://doi.org/10.1016/j.media.2017.07.005

[19] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," Scientific reports, vol. 6, p.26286, 2016. DOI: https://doi.org/10.1038/srep26286

[20] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong, "Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, p. 230, 2019. DOI: https://doi.org/10.1145/3290605.3300460

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017. DOI: https://doi.org/10.1109/iccv.2017.74

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp. 248–255, 2009. DOI: https://doi.org/10.1109/cvpr.2009.5206848

[23] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen et al., "Mllib: Machine learning in apache spark," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 1235–1241, 2016. https://dl.acm.org/doi/abs/10.5555/2946645.2946679

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520, 2018. DOI: https://doi.org/10.1109/cvpr.2018.00474

[25] D. Frossard, "Vgg in tensorflow," VGG in TensorFlow· Davi Frossard, 2017.

[26] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in Human Mental Workload, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, vol. 52, pp. 139 – 183, 1988. DOI: https://doi.org/10.1016/s0166-4115(08)62386-9

[27] Papanastasopoulos, Z., Samala, R. K., Chan, H. P., Hadjiiski, L., Paramagul, C., Helvie, M. A., & Neal, C. H. (2020, March). Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In Medical Imaging 2020: Computer-Aided Diagnosis (Vol. 11314, p. 113140Z). International Society for Optics and Photonics.

[28] Kashyap, S., Karargyris, A., Wu, J., Gur, Y., Sharma, A., Wong, K. C., ... & Syeda-Mahmood, T. (2020, April). Looking in the Right Place for Anomalies: Explainable Ai Through Automatic Location Learning. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1125-1129). IEEE.

[29] Ramos, G., Meek, C., Simard, P., Suh, J., & Ghorashi, S. (2020). Interactive machine teaching: a human-centered approach to building machine-learned models. Human–Computer Interaction, 1-39.

[30] Ishibashi, T., Nakao, Y., & Sugano, Y. (2020, March). Investigating audio data visualization for interactive sound recognition. In Proceedings of the 25th International Conference on Intelligent User Interfaces (pp. 67-77).