Studying Collaborative Interactive Machine Teaching in Image Classification

Behnoosh Mohammadzadeh, Jules Françoise, Michèle Gouiffès, Baptiste Caramiaux

Studying Collaborative Interactive Machine Teaching in Image Classification

Behnoosh Mohammadzadeh
behnoosh.mohammadzadeh@lisn.upsaclay.fr
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
Orsay, France

Jules Françoise
jules.francoise@lisn.upsaclay.fr
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
Orsay, France

Michèle Gouiffès
michele.gouiffes@lisn.upsaclay.fr
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
Orsay, France

Baptiste Caramiaux
baptiste.caramiaux@sorbonne-universite.fr
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR
Paris, France

ABSTRACT

While human-centered approaches to machine learning explore various human roles within the interaction loop, the notion of Interactive Machine Teaching (IMT) emerged with a focus on leveraging the teaching skills of humans as a teacher to build machine learning systems. However, most systems and studies are devoted to single users. In this article, we study collaborative interactive machine teaching in the context of image classification to analyze how people can structure the teaching process collectively and to understand their experience. Our contributions are threefold. First, we developed a web application called TeachTOK that enables groups of users to curate data and train a model together incrementally. Second, we conducted a study in which ten participants were divided into three teams that competed to build an image classifier in nine days. Qualitative results of participants' discussions in focus groups reveal the emergence of collaboration patterns in the machine teaching task, how collaboration helps revise teaching strategies and participants' reflections on their interaction with the TeachTOK application. From these findings we provide implications for the design of more interactive, collaborative and participatory machine learning-based systems.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Empirical studies in collaborative and social computing.*

KEYWORDS

Interactive Machine Learning, Machine Teaching, Collaborative Interaction, User Study

IUI '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645204

ACM Reference Format:
Behnoosh Mohammadzadeh, Jules Françoise, Michèle Gouiffès, and Baptiste Caramiaux. 2024. Studying Collaborative Interactive Machine Teaching in Image Classification. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3640543.3645204

1 INTRODUCTION

Machine Learning (ML) has seen tremendous development and success in various problems and applications in recent years. As a result, it has become one of the fundamental building blocks of the design of interactive applications facing diverse users, from laypeople to domain experts. However, while the deployment of ML technologies on a large scale affects individuals and populations, sometimes with biases leading to harmful consequences, public scrutiny remains limited. In most cases, end users still have little or no control over the training data used to build the model, the evaluation of its performance, and the correction of errors. Involving a wide range of stakeholders in ML models' training and evaluation process could improve their performance, transparency, and fairness. Allowing user groups to create their own ML models to align them with their specific needs and values would benefit many communities of practice and knowledge. However, making ML models accessible to a broad group of users remains challenging because the development process usually requires technical expertise, from providing a dataset to training and testing the model.

Research in Human-Computer Interaction (HCI) has studied how to engage end users in an interactive design process for ML models. This line of research, known as Interactive Machine Learning (IML), is at the intersection of HCI and ML [1]. It investigates ways to involve end users, usually ML novices, in different steps of machine learning model development, from creating and labeling datasets to more advanced levels of control over features, model architecture, quality assessment, and debugging [9]. Through this process, end users can iteratively define and convey concepts based on their domain knowledge to produce customized models that are more accurate and transparent for their specific application domain [21, 22]. As part of this endeavor, recent research has advocated leveraging the teaching ability of humans to convey concepts to learning machines more efficiently [31]. This approach, called

Machine Teaching [41], focuses on the human teacher rather than the learner (ML model). This perspective shift is powerful in exploring how novices in ML convey concepts to learning algorithms and extract insights for designing systems of more accessible and democratic technologies [33].

To the best of our knowledge, most studies in machine teaching have focused on involving individual users in teaching tasks, whether they are novices or domain experts. However, human-centered research has recently encouraged researchers and civil society to move beyond individual opinion to center the values of the broader public into the development and deployment of ML systems [3]. Recent research on "Participatory AI" aims to lead to systems of community empowerment to acknowledge that communities have knowledge, expertise, and interests that are essential to strengthen justice and prosperity [3]. Emerging work calls for greater involvement of affected communities to deliberate on the expectations, concepts, and requirements around ML systems for their communities. The necessity of considering group-level insights and preferences has a long history back to public deliberation or deliberative democracy [13, 45].

Nevertheless, knowledge is scarce regarding the potential and challenges of collective interaction with the learning process in an interactive machine teaching workflow that elicits human knowledge more than deliberation on data annotation. In this paper, we propose to study *collaborative* machine teaching, where a group of users, new to ML, teams up to teach a supervised ML model performing image classification. Our contributions are threefold. First, we designed and developed a web-based collaborative, interactive machine teaching application for image classification. It enables a group of users to build a model collectively through dataset curation, interactive model training, performance inspection, and real-time communication. Second, we conducted a user study where three teams competed for nine days to teach a robust dance-style classifier. The results describe the emerging collaboration in machine teaching tasks and insightful reflections on participants' interaction with TeachTOK. Finally, we build upon these findings to derive a set of implications for the design of future collaborative MT systems.

The paper is structured as follows. The next section reviews previous work on interactive machine learning, machine teaching, and collaborative approaches to machine learning. Then, we present TeachTOK, a web-based application for collaborative machine teaching of an image classifier. In Section 4, we present the study methodology, followed by the results in Section 5. Arising from these results, we discuss our findings and propose three implications for design, presented in Section 6.

2 RELATED WORK

This section provides some background in Interactive Machine Learning and Machine Teaching. We then present recent work on collaborative approaches in interaction with Machine Learning systems, such as Participatory AI and Crowdsourcing.

2.1 Background in Interactive Machine Learning and Machine Teaching

Research in Interactive Machine Learning (IML) proposes involving end users in the training process of ML systems. IML is "an interaction paradigm in which a user or user group iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review" [9]. While conventional ML workflows do not involve end users in the learning phase, IML focuses on flexible workflows where users can potentially participate in activities such as feature selection, hyperparameter and model selection, model steering through the curation of training samples, and quality assessment [9]. Among all activities, model steering requires the most user effort and has received significant attention in IML research. While engaged in the steering task, the user seeks to train the model by providing knowledge in an iterative loop [1, 9]. Wall et al. [44] noted that people can actively contribute knowledge in three main ways in IML: through sampling, in which the learning system sees new examples; labeling, which provides subject-domain expertise as a source of truth; and featuring, which enables the selection of properties to enhance the learning system's internal concept representation.

Interactive machine learning has been investigated in diverse tasks such as image segmentation [10], web image search [14], human-robot interaction [5], or text classification debugging [27]. It is particularly interesting in creative domains where personalization is key and limited data is available [12, 15, 17]. It has also proved useful for ML professionals in model development [21]. Research in IML has focused on the design and evaluation of interaction techniques and novel workflows, given rise to several systems and tools [2, 16, 20].

IML has triggered a shift of perspective: instead of considering fixed training data as representative of a problem that ML can "model", it acknowledges the power of data as a way to steer models in certain directions to address the needs of the task. Consequently, Simard et al. [41] introduced *Machine Teaching* (MT), where the focus is on the human "teacher" who elaborates teaching strategies to convey domain knowledge to a learning algorithm as a "learner". They proposed a research agenda focusing on developing languages and technologies that enable a single user with domain knowledge and teaching experience to teach without requiring ML experts and engineers. They formalize principles for the design of teaching languages, such as feature completeness that provides teachers with all that is needed to teach the model efficiently; the availability of a rich and diverse sampling set; and the distribution robustness that allows the teacher to explore and label data freely.

Ramos et al. [31] further proposed the notion of Interactive Machine Teaching (IMT), where the focus is on leveraging the teaching skills of humans and implicit and explicit forms of their knowledge (labels, features, rules, etc.) in the design of IML systems. The authors emphasize that IMT is distinct from IML by focusing on model-building and the specific role of the human-in-the-loop as a teacher. According to Ramos et al., a person acts as a teacher when engaging in three main activities: *Planning*, where teachers identify diverse, challenging examples to teach, reflect on their strategy, and adjust their approach as they assess the evolution of concepts.; *Explaining*, where they provide the necessary knowledge to the learning algorithm, such as labeling data for classification; and *Reviewing*, to evaluate the confusion, debug errors, and correct labels to gain a comprehensive understanding of the model performance. The authors emphasized a shared teaching language between the teacher and the model to be adopted and used to facilitate the three teaching activities.

Involving stakeholders with various domains of expertise requires a better understanding of how novices in ML can contribute as teachers of algorithmic systems. Wall et al. [44] explored Machine Teaching with MT- novices vs. a group of researchers experienced in MT principles. First, they extracted teaching patterns from users with MT expertise as a set of guidelines including different phases they passed through the teaching process, or patterns regarding when and how to evaluate the learner's knowledge. They concluded that novices without special ML and MT expertise trained models that were not far behind the ones built by MT experts. They observed that novices who received basic MT guidance put less effort and mental demand into performing the task. In the same line of work, Sanchez et al. [33] recently conducted online MT sessions with novices in ML to understand how they intuitively teach a deep neural network on image classification tasks. Their results revealed that participants gained insights into how the model works and what features it takes into account. In addition, they observed participants employed diverse teaching strategies in terms of training size, variability, and sequencing. In a follow-up study, they observed that novices who trained a model with a diverse large dataset understood the ML uncertainty (in particular Deep Neural Network uncertainty) better, leading them to predict the outcome and improve the classifier training [34].

In the context of IMT for image classification, Hong et al. [23] studied how participants experience and reflect on training a robust image recognition model using images taken with their mobile phones. Their results indicate that participants struggled between favoring consistency by providing identical teaching examples or incorporating diversity and edge cases. They analyzed the type of variability induced in the data, showing that participants used diverse teaching examples from humans' perception of diversity independent of size, viewpoint, or illumination that infer variability in object recognition tasks. To further study human teacher interaction with the ML model, Zhou and Yatani [46] showed that innovative interaction techniques for IMT, such as using deictic gestures, can significantly improve the time required to create a model able to recognize objects in a visual scene.

In IML and IMT research, significant effort has been dedicated to integrating various stakeholders in designing and assessing algorithmic systems. Yet, most studies have focused on single-user interactions, and knowledge remains scarce about how the teaching process can be structured collectively.

2.2 Collaboration between users and stakeholders in Machine Learning

Artificial Intelligence (AI) has an increasing effect on people's lives. Consequently, there is growing interest in involving groups of people (users or stakeholders) in creating ML-based systems. This approach is sometimes referred to as "Participatory AI" [3].

On the one hand, previous work has looked at ways to involve a group of people in creating data-driven algorithms to assist collective decision-making. Lee et al. [28] proposed WeBuildAI, a participatory framework allowing each individual to build computational models reflecting individual perspectives, and then work collectively to aggregate model decisions. They used the framework to assess equity and efficiency trade-offs in a matching algorithm for on-demand food donation transportation. Their results indicated that designers and policymakers could use this framework to inform algorithm design or as an auditing or evaluation metric to assess the algorithm's effects from diverse stakeholders' perspectives.

On the other hand, previous work has looked at ways to facilitate deliberation in AI-assisted collective decision-making. Zhang et al. [45] recently framed a prototyping tool as a method to involve a group of users in deliberation with ML models to make fairer organizational decision-making. Through four different stages of data exploration, feature selection, model training, and model evaluation, they observed that ML models could structure deliberation and discussion over abstract beliefs among participants. They emphasized that the goal of their study was not to improve the performance and quality of the ML model. They leveraged ML as just a boundary object [42] that serves as a "frame of reference that [was] used during deliberations by participants to convey their own rationale and understand other people's reasoning".

These recent studies have shown a recent effort to facilitate collaboration in creating ML models to serve collective interests and values. In these works, however, people have not been involved in data curation and its impact on model behavior.

Collective data curation, such as crowdsourcing practices, has been discussed in recent works to increase data quality and model quality [40], but collaboration between people to build ML models is generally limited. For instance, Kellenberger et al. [26] developed an Annotation Interface for Data-driven Ecology (AIDE), an open-source web platform that integrates users and an active learning model [38] into a feedback loop, where user-provided annotations are employed to re-train the model. Although AIDE supports multiple users, there is no collaboration between them. Similarly, Heimerl et al. [19] proposed NOVA, a system that incorporates multiple human annotators into a semi-supervised active learning model to guide users in inspecting and correcting machine-generated labels and, therefore, accelerates the data annotation procedure. In Nova, multiple users were monitoring data annotation, but there was no collaboration among them. Chang et al. [7] presented Revolt, a crowdsourcing labeling system for ML that enables groups of workers to label data collaboratively. However, in this case, users have the opportunity to collaborate to solve confusion on labeling the data, but users do not have the opportunity to act upon the model itself beyond data annotation. Finally, above data annotation, Ferrario et al. [11] presented ALEEDSA, an Interactive Machine Learning application that supports novices in designing, interpreting, and evaluating ML models with Augmented Reality. In this work, each user has a personal workspace. At the same time, ML engineers and domain experts can co-develop a shared workspace to share insights on personal ML model outcomes and investigate the results. However, users can not collaborate to train the model together.

2.3 Summary

Our review of the related works underlines various approaches to integrate a group of end users and communities in building machine learning models: participatory AI, and collective data curation practices. On the one hand, participatory AI is focused on raising the awareness and authority of end-users in ML models to overcome inequalities and uneven power dynamics.

On the other hand, collective data curation practices involve limited agency over the design and training of the models. However, the importance of deliberation, communication, and collaboration in ML data curation has been emphasized by many scholars in recent years [25, 35]. Data production is acknowledged as a collective and interpretive task [18] that should be supported by annotation tools to foster dynamic collaboration and explore its influence on ML models. Machine Teaching and Interactive Machine Learning incorporate users more tightly into the entire process of building and assessing the ML model's performance and have shown ways to build interactive ML-based systems where users could curate training datasets. Still, existing approaches consider almost exclusively individual users. Building upon previous studies that investigated how non-ML-experts develop teaching strategies to create an image recognizer [23, 31, 33], we propose to study collaborative interactive machine teaching, an approach to machine teaching involving a group of users who coordinate, cooperate, deliberate, and discuss to build an image recognizer together.

3 A COLLABORATIVE APPROACH TO IMT IN IMAGE CLASSIFICATION

This paper presents an exploratory study of collaboration in IMT, focusing on image classification. Our goal is to understand what collaboration strategies emerge when a group of users share a common teaching task and object. This section describes the overall scenario of collaborative teaching and our motivations for studying image classification. Then, we present the design and implementation of a system for collaborative machine teaching called TeachTOK, adapted to the use case of image classification.

3.1 Overview of a Collaborative Interactive Machine Teaching Process

In Interactive Machine Teaching, people use a variety of teaching skills to convey domain knowledge to an algorithmic system to produce a model [31]. Ramos et al. emphasizes the diversity of potential roles that users and stakeholders can take in IMT according to their domain and level of expertise, including demonstrating, annotating, curating data, selecting features and models, etc. Our review of the related work shows that even people with limited ML and MT experience can effectively teach ML systems, often by developing their own teaching strategies to make the model robust (e.g. through consistency or variability in the training examples).

Collaborative IMT (CIMT) aims to build models that embody knowledge shared by a community of stakeholders. In collaborative IMT, a group of users addresses a teaching task using a shared teaching medium. Based on related work in collaborative problem solving by Roschelle and Teasley, teaching collaboratively means that teachers "must have ways to introduce knowledge, monitor exchanges for evidence of divergent meanings, and repair any divergences identified to allow meaningful conversations about the problem" [32]. As IMT is usually structured around Planning, Explaining and Reviewing activities, collaborative IMT should encourage the creation of shared knowledge in each activity, using discussion and deliberation. Therefore, we propose that a CIMT process should have the following properties:

- *Shared model*: with a goal of model-building, stakeholders must share access to a common model at any time. This is particularly important for the reviewing activity to be collaborative, so that shared interpretations of the model's performance or behavior can emerge.
- *Shared Data*: Data production, annotation and curation are often at the core of the teaching process, and stakeholders involved in these tasks should share access to a common dataset at any time.
- *Diverse Interfaces*: while the model and data are shared by all stakeholders, visualizations and instruments to operate on these objects should be adapted to the expertise, role or task of the users. Different stakeholders might have different permissions over these objects, and particular interfaces should be developed to support individual contributions.
- *Rich Communication Media*: discussion and deliberation is essential to successful collaboration. Rich communication interfaces must be provided to mediate interaction around the various activities.

Additionally, the implementation of collaborative IMT can vary along several dimensions according to the specifics of the task and context. In particular, stakeholders' involvement over time and space can vary: stakeholders can either be *colocated* or *distributed* in space, and collaborative teaching can take place either *synchronously*, when a collaboration interface depends on users contributing at the same time, or *asynchronously*, if the teaching process happens iteratively over time.

3.2 TeachTOK: Workflow and User Interface

We designed TeachTOK to study how a group of novices in ML and MT would collaborate to solve a machine teaching task. Our main motivation was to observe the emergence of shared teaching strategies and approaches to collaboration. We chose an image classification task because visual data is easier to understand and distinguish, with common grounds for participants from different cultures and languages. Providing teaching examples with different variability is accessible to novices in ML, as shown by previous works [23, 33]. Furthermore, image classification methods are well established, with numerous pre-trained models available online that facilitate the rapid fine-tuning of classifiers.

Our main requirements for the design of this application were that (1) people should be able to train the model using their own images; (2) they should be able to assess performance in real-time and inspect errors; (3) the teaching should be collective, meaning that people are organized in a group where they share a common dataset and model, and they should be able to communicate with each other; (4) the collaboration process should be asynchronous and distributed, meaning that people can contribute to the task from anywhere at different moments.

TeachTOK, a Collaborative IMT application, is designed to meet the previous requirements. The application's workflow is illustrated in Figure 1. In TeachTOK, a group of users constitutes a collective dataset of images to train an image classifier. Users receive feedback on the classifier's performance, and they can inspect errors. To contribute, users can upload images to the platform to build a personal dataset. The latter dataset is used to retrain the classifier

locally, with both the collective data and the new contributions, updating performance measures to let the user assess the effect of their contributions. They can then share their new data publicly with the group. The data is added to the collective dataset and synchronized with all group members.

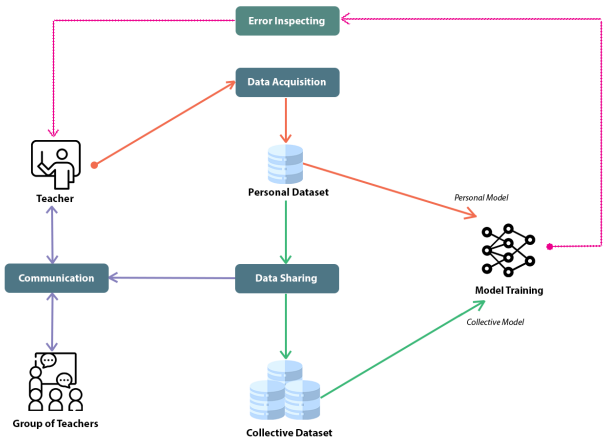

Figure 1: TeachTOK application workflow. The workflow draws upon interactive machine teaching systems and extends the concept to collaborative, interactive machine teaching.

3.2.1 Basic Teaching Mechanisms. The main teaching page of the application is depicted in Figure 2. When an image is uploaded through drag and drop, a prediction is made with the current version of the model. The user receives instant feedback on the classification results with a label and the confidence of each class. If the user decides to use the images for teaching, they can click the button below the drag-and-drop area and indicate a label through a popup window. The training example will be recorded to their personal dataset. At any point, the user can share their personal dataset with the rest of the group (see Figure 7, left). This is done through a separate page that visualizes their own data. When data is shared, it will be made available in real-time to all teammates and will be taken into account when each user retrains the model.

3.2.2 Performance Measures and Inspection Tools. The main performance measure is global accuracy, displayed at all times in the status bar at the top of the page (see, for instance, Figure 2, top right corner). It is measured using 3-fold cross-validation over the total dataset, including shared and personal data (see technical details below). This approach was motivated by the design choice that we needed to provide the user with feedback on the accuracy of the classifier without having a predefined test set.

A dedicated page enables users to inspect errors in more detail, as illustrated in Figure 3. The page includes an interactive confusion matrix to help users understand what categories are confused with one another. Clicking on a cell of the confusion matrix will display the associated images to the top right. Selecting an image will display a larger version and the related class confidences on the bottom row. The model is retrained whenever the training data is modified by the user or one of their teammates, and all performance measures are automatically updated.

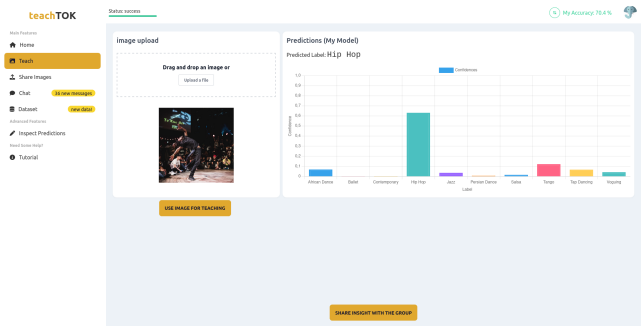

Figure 2: TeachTOK's page dedicated to teaching enables users to upload images, visualize the classification results instantly, and optionally add the image to the training data, if relevant.

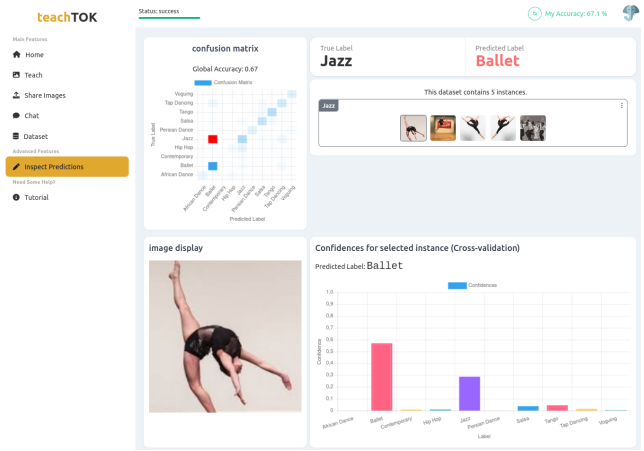

Figure 3: TeachTOK enables users to inspect errors using an interactive confusion matrix (top left). Clicking on the confusion matrix displays the corresponding data (top right). Clicking on an image displays its associated prediction with the latest model, in particular through a bar chart of class confidences (bottom right).

3.2.3 Communication Tools. The main communication medium in TeachTOK is a chat on a dedicated page, illustrated in Figure 4. There are three ways to post content to the chat. First, users can send messages on the chat page. Second, whenever a user shares a set of images with the group, they can specify a comment, and a summary of the contributions will be posted to the chat (categories, number of images, image thumbnails, accuracy). Third, users can share "Insights" from the teaching page, which consist of an image and its associated prediction and are meant to convey specific discoveries and results.

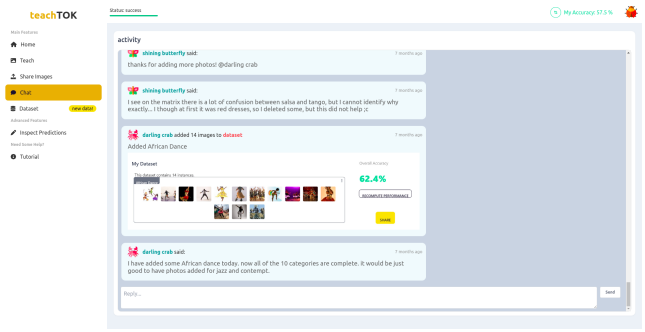

Figure 4: TeachTOK integrates a chat enabling real-time communication among the group of teachers.

3.2.4 Performance Dashboard. The classifier is retrained on startup using the collective dataset and the user's personal data. Users land on a page summarizing information about the system's current performance, as shown in Figure 7 (right) in the Appendix. In particular, line charts display the team's accuracy and number of classes over time, a scoreboard provides a ranking of the current group compared to others.

3.3 Technical Details

3.3.1 Machine Learning Pipeline. Deep Neural Networks (DNNs) are well suited for building interactions involving rich and complex input data for image classification tasks. To enable users to train their own DNN-based classifiers from small datasets, we use transfer learning [37], as it has been proposed in other IML systems [6, 29, 33]. Specifically, we use embeddings from a pre-trained MobileNet [24] model as features, and classification is done using a Multilayer Perceptron (MLP).

Whenever changes occur in the training data, the classifier is trained from all data available for a particular user, and performance metrics are updated. We use three-fold cross-validation for performance evaluation so that a global accuracy measure considers the entire dataset. There are three iterations. In each iteration, the classifier is trained on two-thirds of the dataset and tested on the remaining images, from which predictions are computed and stored. The displayed accuracy is the average of the three computed accuracies. At the end of the process, three MLPs have been trained on three different subsets of the dataset. Real-time predictions make use of these three MLPs, which form a model *ensemble*. Each MLP provides a set of class confidences for a given image, which are averaged to estimate the likeliest label according to the three models.

3.3.2 Implementation. TeachTOK is a web application developed using Marcelle[1], a JavaScript toolkit dedicated to the design of interactive machine learning systems [16]. Following Marcelle's architecture, it uses a Node.js[2] server associated with a MongoDB database for data storage and synchronization, and a web client written in TypeScript. The server relies on the Feathers.js[3] framework for authentication and data storage. Feathers brings real-time updates to all connected clients, which facilitates the synchronization of data and messages within a group of teachers. For IML applications, Marcelle relies on components that can be displayed in a web application and which can be composed into reactive pipelines.

We used standard components to build the core of the application and created custom components for the needs of TeachTOK, in particular, a "Chat" component and several custom visualizations for the dashboard. To prototype our application and run a pilot study, we developed the first version of TeachTOK as a Marcelle *Dashboard*. Eventually, to refine the user experience, we redesigned the app using the SvelteKit framework, and the relevant Marcelle components were displayed in this web app. TeachTOK is accessible as open-source software to ensure reproducibility.[4]

4 USER STUDY

We conducted an online user study with novices in ML to explore the potential and characteristics of collective interactions for machine teaching, regarding teaching strategies, communication and collaboration. Participants were assigned to a team and invited to teach an image classification model as described in section 3 and compete to reach the highest accuracy.

4.1 Participants

We recruited 12 participants with limited or no knowledge of machine learning. Participants were recruited through mailing lists of both our university and professional associations. Applicants were invited to a pre-study interview to explain the procedure, ensure their interest and assess their knowledge of ML. Among the 12 participants, 5 are female and 7 are male, aged between 22 and 29 (mean=25, std=1.95). We asked all participants for a self-assessment of their knowledge of machine learning. 9 participants claimed to be novices, 2 participants claimed limited theoretical knowledge of machine learning, but nothing in the technical part, and 1 stated medium knowledge of both. Participants were assigned to one of the three teams: A, B or C. We refer participants in our results with a code of the form P*XY* where X is the team (A, B or C) and Y is the participant number in the team.

4.2 Setup

We used the video-conferencing tool Zoom to conduct the pre-study interview, during which we explained the modalities and objectives of the study and gave a brief introduction to the TeachTOK application. To start the study, we shared with participants a link to a demographic form created by Google Forms. Then, participants were authorized to carry out the task on their own computer, using TeachTOK in their browser. The application was made available through a URL we communicated to the participants. The application was connected to a Node.js web server and a MongoDB database, hosted at University of Paris-Saclay, to collect data such as the images uploaded by participants. Moreover, a video tutorial

[1] https://marcelle.dev/
[2] https://nodejs.org/en/
[3] https://feathersjs.com/
[4] https://github.com/marcellejs/teachtok

explaining how to use TeachTOK was made available in the application, from which participants could get help whenever they encountered a problem. At the end of the task, participants were invited to take part in a videoconference focus group interview, the audio of which was recorded.

4.3 Procedure

Participants were randomly assigned to three teams of four people to start the experiment. Each participant was provided a link to the TeachTOK application, their associated group, and login details, including an email, a password, and a unique username. Participants were given nine days to teach an image classifier to recognize ten predefined dance styles: Ballet, Contemporary, Voguing, African Dance, Persian Dance, Tango, Salsa, Tap dancing, Hip-Hop, and Jazz. We chose Dance as context because of the expressive nature of dance gestures and postures in a visual representation. We also defined the categories instead of giving users the freedom to choose them by themselves because we wanted to be able to compare strategies between the groups, and therefore require them to work on the same task.

Teams were competing to reach the highest accuracy for a model trained on all ten categories. Each team was initially provided with the same default dataset and the associated model, which included two categories with three instances each. During the experiment, we sent participants daily digest emails informing them of their group performance (displayed on the interface, as explained in the previous section) and competitors' performance (displayed in the dashboard) and encouraging them to collaborate with their teammates.

The study ran for nine days. At the end of the study, participants were invited to participate in a focus group, lasting one hour, to analyze 1) the experience of each participant as an individual and a member of a team; 2) how they perceive the collaborative workflow of the TeachTOK application; 3) individual or collaborative teaching strategies they used; 4) their insights on the data curation; and, 5) their understandings of the ML model.

4.4 Data Collection

Demographic data was collected during the pre-study interview. Data about the participants' behavior was collected during the experiment through the web application. In particular, we collected images uploaded by participants, image labels, comments, and chat messages published by the participants in the web application. Furthermore, models trained from the participants' data, along with the features of the MobileNet network, were also collected. Finally, post-study focus groups were audio recorded.

4.5 Data Analysis

We conducted both a quantitative and qualitative analysis. In this section, we detail the measures and methods used in these analyses. Ten participants were kept for the analysis. Two participants were excluded from the analysis as they did not contribute to the teaching process or the focus group.

4.5.1 Quantitative analysis. The following measures were used to analyze how the participants in each team achieved the final accuracy:

- *Performance evolution over time.* We analyzed each trained model performance evolution over the nine days of the user study based on 1) the accuracy over time implemented in the TeachTOK application and 2) the evolution of the number of classes imported to the model by participants over time.
- *Frequency of different actions*, i.e., how many times each participant performs a specific action during the study, including uploading photos for teaching tasks and observing the model's prediction, inspecting confusion, and inspecting prediction for an individual instance in the confusion matrix.

4.5.2 Qualitative Analysis of the Focus Groups. We applied reflexive Thematic Analysis (TA) [4] to analyze the transcribed audio recordings of the focus group for each team. For each team, two researchers first coded each meaningful verbalization, describing the participant's actions or thoughts. Then, the researchers reviewed and harmonized the codes for each of the transcripts. Themes were constructed from these codes through an iterative process involving three researchers, where candidate themes were proposed, discussed, and reviewed collectively. The analysis resulted in four main themes:

- Two themes relate to **Emerging Collaboration In Machine Teaching**: *"All teams collaborated on the Planning stage."*, and *"A Collective Teaching strategy Emerged through individual values and explanation strategies."*.
- Two results regarding **Interacting with TeachTOK: Opportunities for Reflection and Discussion**: *"Participants reflected on and discussed the teaching process with TeachTOK."*, *"Reflection on the need for collaboration on all teaching activities."*.

The study was conducted in English, as well as the transcripts and the analysis. We gathered the codes, discussed their alignments, and categorized 193 quotes from the 10 participants over the 4 themes mentioned above.

5 RESULTS

In this section, we report the results of the user study. We start with an analysis of the global participation and evolution of the contest, through quantitative data analysis and selected quotes. Next, we present the results of the qualitative analysis of the feedback during the focus groups, from which we report four themes grouped into two main sections concerning 1) emerging collaboration in machine teaching and 2) opportunities for reflection and discussion through interacting with TeachTOK.

5.1 Participation and Outcomes of the Contest

TeachTOK integrated a measure of the accuracy of each team's model on their own data (see Section 3). At the end of the study, the results were:

(1) Team B reached 62.4% accuracy with a total of 197 teaching examples.
(2) Team C reached 58.7% accuracy with a total of 218 teaching examples.
(3) Team A reached 37.3% accuracy with a total of 101 teaching examples.

Figure 5 illustrates how each team addressed the task through the evolution in accuracy and number of classes over time. We observed different strategies for completing the task in each team: Team A focused on five categories during the study while reviewing the errors and evaluating the performance. However, one participant individually rushed to reach ten classes with few examples on the last day without any performance evaluation "*I didn't have time to do the usual corrections on the last day*" [PA1].

On the contrary, Team B (the winner) gradually reached ten categories while maintaining a high accuracy overall. In comparison, Team C completed nine categories on the first day. The participants of this team collaborated to review errors and increase the accuracy during the study. They completed the last category in the last few days with a decrease in accuracy.

Furthermore, focus group discussions highlighted that participants described the experience as fun: "*It was fun to see to what extent I understood what's going on and how I could communicate my ideas to other people who might not necessarily see things the way I see them* " [PB1]. In addition, some participants believed they developed their knowledge about dance styles? "*I actually got familiar with some of the dances I didn't know about like Voguing. It was fun I must say*" [PA2]. And they also gained experimental knowledge about ML: "*I didn't know about ML before, I was familiar with the name and I knew very basic ideas about it. Now I have some kinda experimental knowledge*" [PB2].

5.2 Emerging Collaboration In Machine Teaching

We found two themes in the qualitative analysis of the focus group regarding how collaboration emerged in the machine teaching task, and how it helped participants revise a collective teaching strategy.

5.2.1 All teams collaborated on the Planning stage. As mentioned in section 2, the IMT process can be decomposed into three stages: planning, explaining, and reviewing [31]. We found that all teams managed to collaborate on the planning activity of the teaching task, using different strategies to coordinate the work among team members.

By communicating through the chatbox, members of team B assigned each category to a member to distribute the task: *"we decided to split the dataset among each other, it was, I think, one of the first discussions we had in the chat."* [PB1]. In team A, the teaching task was distributed by assigning different roles to each team member. One member was focused on Explaining activity by curating a dataset. In contrast, the other was focused on Reviewing activity by evaluating the provided dataset, the errors and the model performance. Concerning this kind of task coordination, members considered their parts as complementary roles as one member claimed, *"I was trying to upload as many pictures as I can, and he was giving me directions about the pictures because I am not familiar with the dances nor ML or AI. So he actually helped me to do this."* [PA2], while the other member said, *"If no one offered images in the first place, I was unable to do anything..."* [PA1]. Finally, team C's members claimed they were coordinating each other implicitly: *"It wasn't explicit like chatting or saying you do this category, I do this one, splitting the task between each other was implicit."* [PC3]. They coordinated the task by observing each other's activity without direct communication. However, collaboration emerged gradually through these observations, *"I saw for ballet we have enough accuracy, and it was good, but on the other hand, for Persian dance or Voguing, there was not so much data shared, so I was trying to work on that part."* [PC1].

Eventually, regardless of the different approaches to coordinating the task as a group, we observed that all teams began to communicate between team members to share feedback on the curation of the dataset, such as identifying additional criteria for selecting teaching examples and incorporating these criteria into teaching strategies. In other words, all teams collaborated on the Planning activity through the study.

Finding 1: Regardless of the different approaches for coordinating the teaching task, all teams proceeded to collaborate on the Planning activity.

5.2.2 A collective teaching strategy emerged through individual values and explanation strategies. Participants were asked to build a classifier within a limited timeframe, with an emphasis on accuracy as the main performance measure. Participants became aware of their different strategies for selecting teaching examples by observing contributions *"not from the discussion but from the kind of images uploaded I think we had different ways of selecting photos. "* [PB1]. Focus groups highlighted important values beyond performance expressed by a single accuracy measure. Some participants valued "unambiguous" images to facilitate the discrimination of categories; for instance, a participant mentioned *"I felt that in some dances, there are really slight differences between them and the images were looking very similar. I was thinking I have to provide very good quality photos that are very clear to understand."* [PB3]. On the contrary, PB1 was consciously building a representation for a dance that she wanted to be diverse: *"I was trying to represent diversity and messiness, so that even if it is a messy screenshot or something, the model would be able to just recognize because I thought since it is a movement it is not always beautiful and staged."* [PB1]. Going deeper into the participants' insights on diversity, we found that it was intuitively correlated with identifying biases in how dances are represented. For instance, PC2 commented that *"normally, for the Ballet, we can see that it is based on the clothes, and I thought perhaps if I added images of famous dancers it could not eventually recognize, I went for practices with normal clothes. Then I added the men because I thought perhaps if I just added women's dancing, it would not work well. So I just tried to cover all kinds of movements, genders, and all we can see in Ballet."*.

Consequently, we observed how participants gradually reflected on each others' strategies and tried to revise a collective teaching strategy through collaboration and communication. First, some participants noted that collaboration helped them identify biases: *"if someone provides you with inputs… you actually can find some of your own biases that you weren't aware of."* [PA1]. Furthermore, we observed how they deliberated over their teaching strategies by communicating with other team members. For instance, a participant from team A claimed: *"When I saw his comments on images, mostly on Salsa and Tango, I tried to use images with different colors. If I wanted to upload pictures about Salsa or Tango, I tried to go further and ignore red dresses or colorful dresses. I even tried to upload some B&W pictures."* [PA2]. Moreover, collaboration helped them

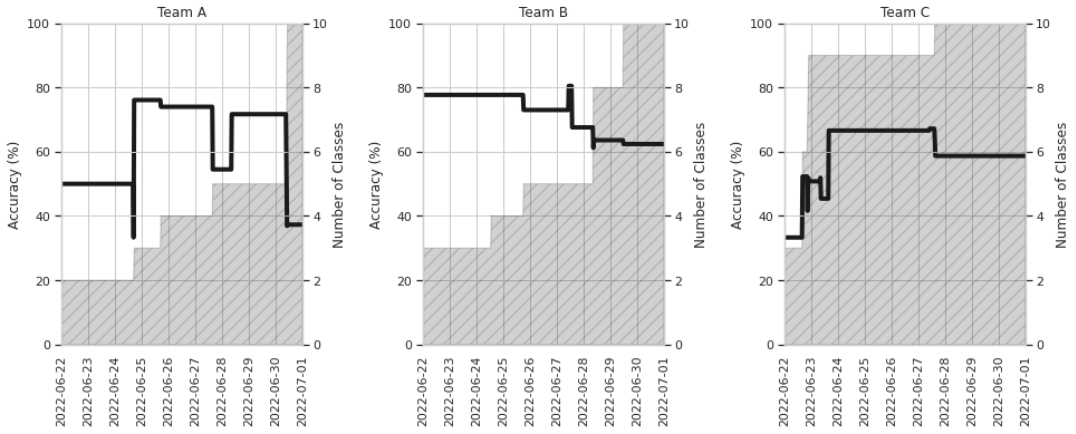

Figure 5: The evolution of the model accuracy (solid black line) and number of classes (grey dashed area) throughout the user study.

integrate potential objectives into their teaching strategies: *"My own strategy was mine, what I didn't have in my mind and someone mentioned that to me was that I didn't look specifically about the diversity and in the few last days what I did was to expand the diversity."* [PB2].

Finding 2: Collaboration helped participants become aware of potential teaching objectives and integrate them to revise a collective teaching strategy.

5.3 Interacting with TeachTOK: Opportunities for Reflection and Discussion

We found two themes regarding the interaction with TeachTOK, the collaborative interactive machine teaching system that we developed. In particular, focus groups highlighted that TeachTOK enabled participants to reflect on the teaching process and gain ML experience, but with limitations regarding collaboration and communication.

5.3.1 Participants reflected on and discussed the teaching process with TeachTOK. The interactive nature of the teaching process was instrumental in helping participants reflect on the teaching activity. Figure 6 reports the count of different interactions of individual members of each team with TeachTOK, including Inspecting Confusions, Inspecting Predictions and Uploading teaching examples. We observe that participants used TeachTOK's features in different ways. Some participants heavily relied on the global accuracy metric and on the real-time feedback provided when uploading an image: "*for me it was mainly the accuracy so if I saw that the accuracy was improving it was good okay it's correct or if it was decreasing I had to find more photos for the dataset. I didn't think of more of that*" [PB3]. On the contrary, other participants integrated insights from the confusion matrix to plan their image curation strategies. This interaction between reviewing and planning included different levels of granularity. For example, PC1 used it for prioritizing categories: *"for me it was mostly the matrix, I was looking at the matrix and I chose which categories needed to work on."* Other participants used the confusion matrix to evaluate the model performance and review its errors: *"most of the observations I made I tried to get them in what was visible from the website, that is to say the confusion matrix which was particularly useful just seeing what images were wrongly labeled could given even to a layman a very clear indication of what was the problem."* [PA1].

In addition, we observed that the confusion matrix and instance inspection led participants to reflect on visual attributes shared among categories. A participant noted that *"I started investigating the confusion matrix to see what is going on, which photos of my categories are confused with other categories to be able to identify potential similarities in the photos."* [PB2]. Another participant added: *"I always thought of Tango as a dance that needs two people that women wear red dresses, but when I was uploading pictures, we have that in Salsa too or maybe in Hip-Hop even African dances that they use that same red maybe"*, [PA2]. Accordingly, the confusion matrix and real-time predictions helped participants avoid biasing the model by identifying edge cases: *"Sometimes I would decide not to upload this photo to the dataset because I thought the model already works very well for this photo so why would I reinforce the same thing and I would prefer to upload the images that had kind of borderline confidence."* [PB1]. TeachTOK's design seems to have facilitated individual reflection on machine teaching, through a tight loop integrating reviewing, planning and explaining. Additionally, participants used the chatbox to deliberate with other team members over the task. In particular, participants in teams A and B exchanged comments and messages in the chat, asking others for feedback or actions: "*couple of messages were about "hey I just uploaded a new dataset, take a look if you have time to give feedback"*" [PA2]. And then also saying that "*oh that category is abandoned, can someone do something for that category?"*" [PB1]. They also used communication tools for sharing strategies and advice, as the same participant believed *"it would be helpful for someone and also I could get feedback in case I forgot to think about something. "* [PB1]. Additionally, all teams used the chatbox to reflect on their teaching strategies by observing their teammates' contributions, including insights or dataset updates in the chat, to coordinate their planning for providing additional teaching examples. For instance,

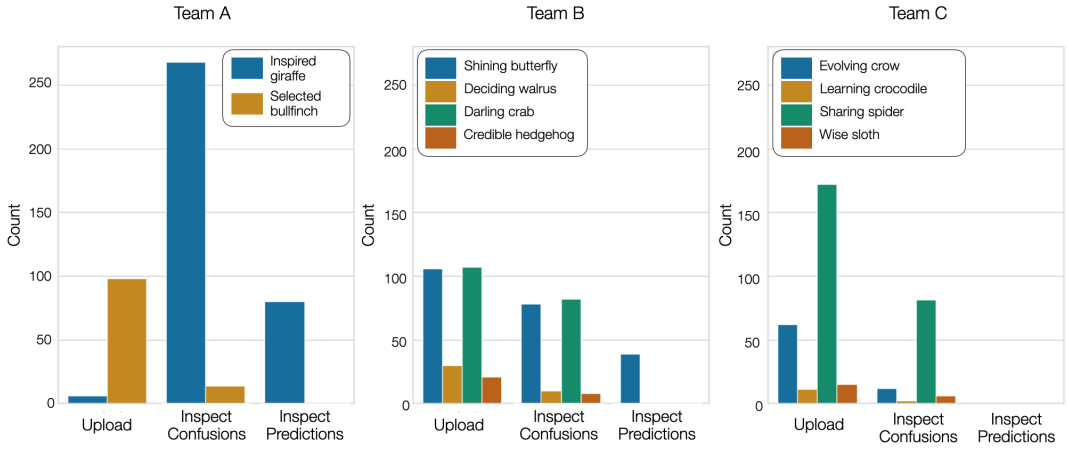

Figure 6: The frequency of use of various actions across teams. Actions are: "Upload", which means a participant uploaded a teaching example and saw model prediction; "Inspect-confusion", meaning that a participant clicked in the confusion matrix; and "Inspect-prediction", meaning that a participant looked at the prediction for a given image of the dataset.

PC2 stated: *"I was kinda looking at what others are doing so not to add extra."* . We found that participants were following the chat, not necessarily for discussion, but to compare their approach to teaching task with others: *"It was more like to see what everyone did, what was different from my method."* [PB4].

Finding 3: Participants reflected on the teaching task through individual interaction with the TeachTOK application. They used communication tools to deliberate with other team members over the task.

5.3.2 Reflection on the need for collaboration on all teaching activities. Our focus group discussion brought up the limitations participants encountered during the task. We observed that these limitations were due to the need for more communication and collaboration on all aspects of the teaching task.

The thematic analysis first highlighted that participants were required to set an explicit common goal at the beginning of the task *"What I understood is that at first, it's important to decide on the goal, whether we want the algorithm to be more varied, better on different types, or we want it focused and more accurate in a few categories."* [PB4]. This requirement is correlated with "collaboration" definition by Roschelle and Teasley[32] that collaboration involves symmetrical structure including a collective goal, a shared conception of the problem.

Moreover, participants reflected on their collaboration strategies and the necessity to collaborate on various aspects of the teaching task. For instance, Participants of Team B understood the importance of data reviewing. Two participants realized that they should have focused not only on the image selection task but also on inspecting what others did. *"For example, each of us could have two categories to upload photos and two categories to review the photos. So this way, everyone could easily upload photos for the categories, and we would make sure that someone else would also look over the uploaded photos, and if they are missing something, the other reviewers could mention it."* [PB2]. PB1 also said: *"... there would be one person who selects the photos also check the photos of other categories. So each category have several opinions, at least."* . On the other hand, the participant in team A, with the role of reviewing the model errors, asserted that he should have collaborated with the other teammate on providing the teaching examples *"I personally would try to be more consistent, to say to try to do more, this strategy to just wait for others to do something is not very sustainable in the long term. If I wanted to begin again I would provide images as well and not rely specifically on teammates to work for me."* [PA1]. To establish collaboration, participants generally asserted the necessity of more discussions over the teaching task. This includes discussion on image selection *"I wish maybe that for the photos I had uploaded, I had received more feedback from others."* [PB1], and discussion about teaching strategies *"There wasn't any discussion about whether or not we should include this specific image."* [PA1], *"I kinda did my mental checklist of how I chose the photos to train and I wish I could get more insights about how others did that."* [PB1]. Eventually, a participant regretted a lack of interactive communication: *"There was no interactive part in the discussion, it was just doing something, justifying and reporting what we have done basically. I have done this specific thing for this specific reason. Kind of report to someone."* [PA1], however, he believed they used the utmost of what TeachTOK communication tools provided them *"it was very consistent with the tools that the web application provided us."* [PA1].

Finding 4: Participants required more discussion and effective collaboration on all teaching activities (Planning, Explaining, and Reviewing) to converge a collective knowledge of the model and the teaching task.

6 DISCUSSION

We have analyzed how a group of ML novices can collaborate in teaching an image classifier to recognize dance styles. Qualitative analysis of focus groups revealed despite various organizational strategies, including different task coordination approaches, all teams collaborated on the teaching task. It brought insights into participants' teaching values and strategies, such as choosing a

diverse set of examples and how they collaborate to integrate them into a collective teaching strategy. By analyzing teachers' interaction with TeachTOK, we also showed participants collaborated to communicate their reflections on the teaching task, while they required more discussion on other teaching activities. Although our sample of 12 participants was small, our qualitative results provided valuable insights. In this section, we first build on these results to propose three implications for designing collaborative interactions with machine learning systems. We then discuss our findings in more detail.

6.1 Implications for the design of collaborative IMT

Grounded on our user study's findings, the section proposes to list three implications for the design of Collaborative Interactive Machine Teaching system.

6.1.1 Let participants negotiate teaching criteria and choose performance metrics. Although our user study focused on accuracy as the main objective in realizing the task, we showed that participants used additional criteria in their strategy to choose their teaching examples, such as diversity in the representation of each class considered (Finding 2). Criteria such as diversity related to values that were important to some participants regardless of their impact on the task outcomes. Additionally, we observed that participants used diverse techniques to assess the model's quality beyond accuracy, the only quantitative performance measure provided in TeachTOK. Participants inspected datasets, confusions, and predictions to understand where the model failed, making hypotheses about the causes. We suggest letting participants negotiate the success criteria and performance measures before and during the teaching process, for instance by providing additional performance metrics. For instance, users need to understand the dataset's diversity level from the model's perspective. The availability of a set of metrics would enable discussion between participants and improve the collective teaching process.

6.1.2 Let participants negotiate roles collectively. TeachTOK was designed with various features available to all participants and without predefined roles in order to observe the emergence of collaboration patterns. We observed that some teams distributed tasks and roles, illustrating that organizing collaborative activities requires agreements on the division of labour and roles to facilitate group work [36, 43]. Allowing participants to negotiate their roles collectively can be valuable in coordinating a collaborative task. Participants can define their roles based on their expertise and interests while remaining open to reassigning roles as the teaching task evolves "*I have mostly been busy with what I thought were the funniest part of the experiment*" [PA1]. Interactive machine teaching platforms should give flexibility in the definition of roles. This involves managing access control over various aspects of the ML pipeline (data, annotation, features, models, hyperparameters, etc.) and providing appropriate user interfaces over the pipeline according to the role.

6.1.3 Facilitate discussion in the machine teaching loop. We found that participants required more discussion and effective collaboration on the teaching task (Finding 4). By improving communication tools, we can foster collaboration in all teaching stages (Planning, Explaining, Reviewing). Participants should be able to effectively share insights on various aspects of the model. For instance, communication tools should facilitate discussion at the Reviewing stage to converge towards a collective understanding of the sources of errors and properties the model is sensitive to. Moreover, collaboration on the Explaining activity could be improved by designing interactions that support discussion and negotiation within the data curation process, for example by including validation and reviews of data and annotations by other stakeholders.

6.2 Collaboration or Cooperation

We found that teams A and B began the task by collaborating during the first cycle of the Planning activity. In particular, Teams A and B organized their work by dividing responsibilities among team members, either by categories or roles. Following each team's dynamic, we observed that in team A, one member was responsible for the Teaching activity and the other for the Reviewing activity as an individual work. In team B, each member individually iterated all teaching activities for specific categories. In contrast, Team C did not divide the task and relied on individual contributions. Teams therefore used different strategies involving either collaboration or cooperation. Indeed, cooperation is characterized by the division of labor among participants, with each person responsible for a specific aspect of problem-solving [32]. This division of labor was implemented by teams A and B, at least at the beginning of the experiment. We then observed participants in all three teams exchange insights and discuss to converge towards a collective understanding of the model. This form of communication made them aware of potential criteria identified by other team members and helped them integrate these criteria into a collective teaching strategy. This indicated a shift from cooperation to collaboration in the planning activity, which aligns with Roschelle and Teasley [32]'s definition of collaboration as an ongoing effort to establish and maintain a shared problem-solving framework. Correspondingly, all three teams did the Planning activity by collaborating on identifying potential criteria, reflecting on their strategies and adjusting their approach. This suggests that forms of cooperation have been established in advance of forms of collaboration. An interesting line of research would be to study the development and dependencies between these forms of "working together" in the context of collaborative interactive ML.

Although all groups continued collaborating on the Planning activity, we observed that participants mostly performed the Explaining and Reviewing tasks individually in all teams. This prompts us to consider what form of collaboration might emerge in the context of these activities. The participants' feedback during focus group interviews provided limitations regarding the objectives set out in this study. For instance, participants often lacked discussions about their data curation strategies, image selection criteria, and the quality of knowledge transmitted to the model. This suggests a potential collaborative approach in the Explaining activity, where participants could collectively discuss the model's required knowledge, image selection strategies, and feedback on labeling. Such collaboration might enhance participants' domain expertise over time, opening up opportunities to evaluate knowledge evolution

within collaborative efforts. On the other hand, participants also demonstrated the need for a shared understanding of the errors. Achieving this collective agreement requires group discussions and collaboration within the Reviewing activity, exchanging insights on the confusion matrix and performance indicators.

6.3 Collective Interactions to Support Fair and Participatory AI

In our study, we explored the concept of collaborative, interactive machine learning to understand better how a group of people (mainly ML novices) can work together to improve an image classifier. In doing so, we place ourselves in a broader context of recent efforts in participatory AI, which emphasizes the construction of predictive algorithms by the community on which that algorithm can impact [3]. As a matter of fact, two key papers in both IML and Participatory AI are entitled "Power to the People" [1, 3], which emphasizes their shared goal of bringing ML technology closer to users and stakeholders. Therefore, looking at CIMT through the lens of participatory AI can be instructive in several ways, and we elaborate on three of them.

First, bridging Participatory AI and Collaborative Interactive Machine Teaching can facilitate the design and development of interactive machine learning systems that are more effective and aligned with the community's perspective. Collaborative participation of community members could provide fruitful insights into the design of IML workflows, including community standards for the required system, their goals, and limitations. Even though participants in our study were not domain experts in the field of dance, we could already observe the way CIMT can be instrumental in triggering discussions and reflections, especially on input data that was curated by participants, contrary to much of the previous work in participatory AI [28, 45]. We encourage incorporating a participatory perspective that enables users to collaborate on the design choices of the task and of the IML workflow so that they align with community preferences. For example, the teaching stages (planning, reviewing, and explaining) can help define a framework for organizing people's participation in developing an ML model.

Second, CIMT has the potential to foster AI literacy, which is critical in participatory AI. Recent studies in Interactive Machine Teaching have highlighted how this approach can contribute to transmitting ML-related knowledge and help ML novices acquire experiential skills in ML and AI [33]. The introduction of greater interactivity in collaborative systems incorporating AI could encourage the development of these skills. Interactivity can be enabled on models, as most previous work on participatory AI has shown [45], but also on datasets, as shown in this article, which still needs to be studied. Allowing participants to manage datasets and explore their impact on the model seems fundamental to understanding the systems involved.

Third, CIMT has the potential to enhance the quality of ML models concerning fairness, inclusivity, and the mitigation of harmful behaviors such as discrimination and bias. In our study, we observed that participants demonstrated an understanding of the importance of diversity in training sets despite being novices in machine learning. They recognized that diversity in datasets can help overcome some biases in the model's predictions. As recent works addressed, incorporating end-users perspectives can help identify and fix fairness issues [30]. The collective efforts of users to perceive such issues and discussions to raise awareness among the community have also been documented [8, 39]. A promising research direction could involve communities in CIMT, encouraging collaboration in reviewing and correcting their machine learning models to mitigate potential harm and biases.

7 CONCLUSION AND FUTURE WORK

In this paper, we presented a study of collaborative and interactive teaching of an image classification system by novices. The study was made possible by the development of TeachTOK, a web-based application that allows users to upload images, teach a classification model, review the model's performance, communicate with team members, and share the model and data with the group. We conducted a user study with 10 participants divided into three teams. The goal was to reach the highest classification accuracy in recognizing ten dance styles from images. We observed that the teams adopted different strategies to coordinate tasks among team members. All teams collaborated to share insights on the model, convey their different criteria, and integrate collective teaching strategies. Additionally, we analyzed the opportunities and challenges of the TeachTOK application from the participants' perspective. We observed that participants required more discussion and collaboration to converge a collective understanding of the teaching task. Finally, arising from these findings, we proposed three implications for the design of collaborative interactions with machine learning systems.

By exploring the potential of collective interaction in the context of machine teaching, this study constitutes a step toward more accessible artificial intelligence for user groups and communities. Because of the richness of people's experiences and perspectives, collaborative approaches involving a broad spectrum of stakeholders can potentially improve practices around data curation and model development. Nevertheless, this approach raises several new questions and challenges. For instance, the impact of the level of diversity perceived by participants and their reflection on biases remains to be systematically studied. The CIMT approach from a participatory perspective can help to address these problems by enabling communities to act on their data and models.

ACKNOWLEDGMENTS

This research was supported by the ARCOL project (ANR-19-CE33-0001) – Interactive Reinforcement Co-Learning, and by the ELEMENT project (ANR-18-CE33-0002) - Enabling Learnability in Embodied Movement Interaction, from the French National Research Agency. We want to acknowledge and thank everyone involved in each stage of the research. We want to express our sincere gratitude to the anonymous reviewers for their constructive comments.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[2] Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. 2022. Symphony: Composing interactive interfaces for machine learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118.

[6] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382839

[7] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[8] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[9] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. https://doi.org/10.1145/3185517

[10] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[11] Andrea Ferrario, Raphael Weibel, and Stefan Feuerriegel. 2020. ALEEDSA: Augmented reality for interactive machine learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

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

[13] James S Fishkin. 2002. Deliberative democracy. *The Blackwell guide to social and political philosophy* (2002), 221–238.

[14] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*. 29–38.

[15] Jules Françoise and Frederic Bevilacqua. 2018. Motion-sound mapping through interaction: An approach to user-centered design of auditory feedback using machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–30.

[16] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 39–53.

[17] Marco Gillies. 2019. Understanding the role of interactive machine learning in movement interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 1 (2019), 1–34.

[18] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[19] Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. NOVA-a tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 109–115.

[20] Clarice Hilton, Nicola Plant, Carlos González Díaz, Phoenix Perry, Ruth Gibson, Bruno Martelli, Michael Zbyszynski, Rebecca Fiebrink, and Marco Gillies. 2021. InteractML: Making machine learning accessible for creative practitioners working with movement interaction in immersive media. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[21] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[22] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[23] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[25] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[26] Benjamin Kellenberger, Devis Tuia, and Dan Morris. 2020. AIDE: Accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution* 11, 12 (2020), 1716–1727.

[27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[28] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[29] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Trans. Interact. Intell. Syst.* 12, 3 (Sept. 2022), 1–30.

[31] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.

[32] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*. Springer, 69–97.

[33] Téo Sanchez, Baptiste Caramiaux, Jules Françoise, Frédéric Bevilacqua, and Wendy E Mackay. 2021. How do people train a machine? Strategies and (Mis) Understandings. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[34] Téo Sanchez, Baptiste Caramiaux, Pierre Thiel, and Wendy E Mackay. 2022. Deep Learning Uncertainty in Machine Teaching. In *27th International Conference on Intelligent User Interfaces*. 173–190.

[35] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[36] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously: Supporting articulation work. *Computer Supported Cooperative Work (CSCW)* 1 (1992), 7–40.

[37] Tyler Scott, Karl Ridgeway, and Michael C Mozer. 2018. Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning. *Advances in Neural Information Processing Systems* 31 (2018).

[38] Burr Settles. 2009. Active learning literature survey. (2009).

[39] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[40] Victor S Sheng and Jing Zhang. 2019. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9837–9843.

[41] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).

[42] Susan Leigh Star. 1989. The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In *Distributed artificial intelligence*. Elsevier, 37–54.

[43] Anselm Strauss. 1985. Work and the division of labor. *Sociological quarterly* 26, 1 (1985), 1–19.

[44] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using expert patterns in assisted interactive machine learning: A study in machine teaching. In *IFIP Conference on Human-Computer Interaction*. Springer, 578–599.

[45] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[46] Zhongyi Zhou and Koji Yatani. 2022. Gesture-aware Interactive Machine Teaching with In-situ Object Annotations. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

A ADDITIONAL FIGURES

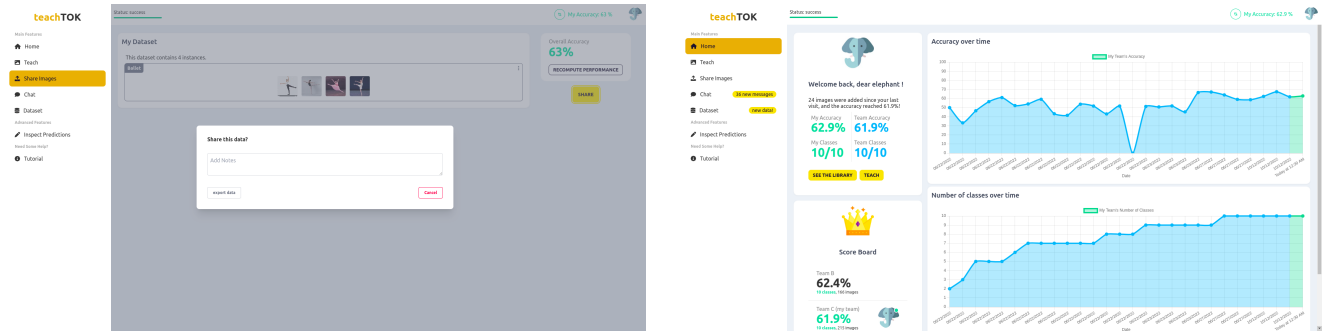

Figure 7: Additional Screenshots of TeachTOK. *Left:* TeachTOK's sharing page. Users can upload a set of multiple images to teach the model. They can assess improvements in accuracy before sharing the data publicly with the group. *Right:* TeachTOK's home page is a dashboard summarizing the progress of the teaching task. On the right, it includes charts summarizing the evolution of the accuracy and number of classes over time. On the left, it summarizes the current model performance in comparison with other teams.