A Simulation-Based Approach for Quantifying the Impact of Interactive Label Correction for Machine Learning

Yixuan Wang, *Graduate Student Member, IEEE*, Jieqiong Zhao, *Member, IEEE*, Jiayi Hong, *Member, IEEE*, Ronald G. Askin, and Ross Maciejewski, *Senior Member, IEEE*

***Abstract*—Recent years have witnessed growing interest in understanding the sensitivity of machine learning to training data characteristics. While researchers have claimed the benefits of activities such as a human-in-the-loop approach of interactive label correction for improving model performance, there have been limited studies to quantitatively probe the relationship between the cost of label correction and the associated benefit in model performance. We employ a simulation-based approach to explore the efficacy of label correction under diverse task conditions, namely different datasets, noise properties, and machine learning algorithms. We measure the impact of label correction on model performance under the best-case scenario assumption: perfect correction (perfect human and visual systems), serving as an upper-bound estimation of the benefits derived from visual interactive label correction. The simulation results reveal a trade-off between the label correction effort expended and model performance improvement. Notably, task conditions play a crucial role in shaping the trade-off. Based on the simulation results, we develop a set of recommendations to help practitioners determine conditions under which interactive label correction is an effective mechanism for improving model performance.**

***Index Terms*—Interactive label correction, label noise, machine learning, simulation, visual analytics.**

I. Introduction

THE success of machine learning is known to highly depend on the quantity and quality of the training data [1], [2], [3]; however, the process of labeled data acquisition is time-consuming and cumbersome. To solve this, researchers investigated collective approaches such as crowdsourcing [4] and web crawling [5]. Yet, even these methods can introduce label noise into the data, i.e., data instances are assigned to the wrong labels. Thus, visual interactive labeling, an **I**nteractive **M**achine **L**earning (IML) approach, has been proposed to achieve better data quality and improve model performance. IML is an iterative process that integrates a human and a machine learning model to solve complex real-world problems that could not easily be solved by a human or a machine alone [6]. Visual analytics research has proposed a variety of frameworks to facilitate the IML process [7], [8], [9], [10], where visualizations were utilized to support humans in inspecting suspicious data, and further labeling or relabeling data [11], [12], [13], [14] or cleaning mislabeled data [15], [16].

Such visual IML tools have claimed the complementary benefits between humans and machines to improve machine learning [12], [15], [17], [18]; however, the relationship between the costs of human intervention and the benefits of model improvements in an IML process have not been well studied in both Visual Analytics and human-computer interaction research fields. In this cross-cutting paper, we explore the relationship between the benefits and costs of involving humans in an interactive label correction process. We investigate whether the benefits (e.g., increased accuracy after interactive relabeling) consistently outweigh the costs (e.g., human labor cost during interactive relabeling). The benefits have an extensive scope, including the potential for humans to enhance the model quality, gain new insights, establish appropriate trust in the algorithm, and construct an accurate mental model of the underlying algorithm. Though some benefits can have far-reaching effects, they are still broad open research questions, e.g., quantifying human insight, mental models, and trust. To appropriately scope our analysis, this work focuses on the potential of improving model quality through interactive label correction. Following the prior research [19], we quantify human efforts on label correction as human cost in an interactive label correction process. Based on the unified workflow for visual interactive labeling (VIAL) proposed in previous work [11], we simulate different levels of human effort in correcting mislabels on training data and quantify benefits by estimating the model performance enhancement on testing data after the label correction. **Our simulation focuses on the best-case scenario where annotators will always accurately inspect and revise mislabeled instances with perfect visualization, which serves as an *upper-bound* for estimating the potential performance gains in an interactive label correction process. We call these annotators *perfect-agents*.** Then, we explore a set of environmental factors that form

Received 16 February 2024; revised 30 August 2024; accepted 12 September 2024. Date of publication 26 September 2024; date of current version 1 August 2025. This work was supported in part by the U.S. Department of Homeland Security under Grant 17STQAC00001-08-02, and in part by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under Grant 1939725. Recommended for acceptance by R. Metoyer. *(Corresponding author: Ross Maciejewski.)*

Yixuan Wang, Jiayi Hong, Ronald G. Askin, and Ross Maciejewski are with the Arizona State University, Tempe, AZ 85287 USA (e-mail: ywan1290@asu.edu; jhong76@asu.edu; Ron.Askin@asu.edu; rmacieje@asu.edu).

Jieqiong Zhao is with the Augusta University, Augusta, GA 30912 USA (e-mail: jiezhao@augusta.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3468352, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2024.3468352

different task conditions for classification, covering the *Data* and *Learning Model* blocks in the VIAL workflow, and study their influences on the relationship between the costs and benefits of the simulated human-assisted label correction (covering *User* block in VIAL workflow) concerning model performance. We validate the simulation result on five real-world noisy datasets (i.e., five noisy label sets of CIFAR-10 data) and develop a set of recommendations for interactive label correction. We found that, in our simulated upper-bound scenarios, there exists an optimal stopping point of label correction indicating the most cost-effective case. It infers that, in real scenarios, it may be unnecessary to invest excessive efforts in label correction because the benefits gained can be marginal as the costs increase, such as the cost of developing optimal visualizations and correcting all mislabels. Our contributions include:

- A simulation-based method to explore the relationship between the cost spent on relabeling and the benefits (i.e., cost-benefit trade-off) of interactive label correction.
- Evaluating simulation results to identify the task conditions that most benefit from interactive label correction.
- Providing recommendations to assist practitioners in revisiting visualization design for the interactive label correction process.

II. Related Work

In this section, we review the general IML pipeline, interactive labeling approaches, and typical methods for evaluating IML systems.

A. Interactive Machine Learning Pipeline

The IML process is an interactive paradigm that uses human inputs and verification feedback to build and iteratively refine a machine learning model [9], [20]. Fails and Olsen [21] were the first to propose the term IML when they implemented the train-feedback-correct cycle to enable model developers to revise training data (mark correct pixels on images) to reduce the errors made by a classifier. Since then, researchers have worked to define a general IML pipeline that is used extensively in the visual analytics community. While a variety of IML pipelines have been proposed (e.g., [8], [9], [10]), the common features focus on four components: data preprocessing, feature engineering, model building and selection, and performance evaluation. In these IML pipelines, humans can interact with any component to observe the updates in subsequent connected components. Furthermore, the former component's output becomes the latter component's input. The typical data preprocessing component supports data identification, extraction, cleaning, and transformation to ensure data quality [22], [23], [24]. The feature engineering component often includes feature generation and selection to determine the most representative set of features, often visualized using scatter plots and parallel coordinates [25], [26]. The model building and selection component focuses on training various models and interactively exploring their performance to determine an appropriate model choice. The performance evaluation component is designed to help analysts understand under what instances the model may underperform and may suggest mechanisms for improving training. Though recent advanced AI techniques have enabled automation in processing data, features, and model selection, IML is still an efficient approach for injecting human knowledge to achieve a well-performed and customized machine learning model.

B. Interactive Labeling

Learning from a sufficient amount of high-quality data is pivotal for the success of machine learning techniques, and the data preprocessing component has generally been recognized as the most time-consuming with respect to the human effort required to clean the data [24]. As such, interactive labeling and relabeling have been major directions in the visual analytics (VA) community. These tasks require visual inspections of salient instances that influence underlying models. In these interactive visual labeling systems, effective representations of large-scale instances in 2D space [6], [27] and intuitive interactions with visual interfaces help human annotators make wise choices of instances for labeling, which can accelerate error finding and correction. Labeling tasks and their strategies are also well-studied in the Machine Learning (ML) community. Techniques such as active learning [11], [13], [14], [28] and propagation algorithms [15], [29] are applied to reduce human labor costs by automatically recommending candidate instances to be labeled. Bernard et al. [30], [31] formalized the instance selection strategies of data labeling from both the VA and ML perspectives.

In this paper, we focus on investigating a specific labeling task, label correction, that impacts the model training and calibration, where humans can visually identify mislabeled instances and discover problems within the training data. Humans can interactively review and correct labeled instances through visualization to steer the model towards expected results [15], [16], [32], [33]. Researchers in the ML classification domain revealed that the dataset characteristics significantly affect ML classification performance and the choice of proper ML algorithms [34], [35], [36], [37], [38], [39]. Especially when encountering noisy training data, different choices on ML models can lead to distinct tolerance to training label noise, i.e., model robustness, as well as the model classification performance. Prior work has developed techniques for automatic label correction [40], [41], or overcoming label noise [42], [43]; however, little work has investigated the costs and benefits of human-centered label correction on different types of datasets. Human-centered label correction enables humans to selectively choose suspicious data for label cleaning based on their knowledge, thereby optimizing the model quality enhancement. Also, the instance selection strategies can greatly influence the efficiency of label correction on model performance improvement. For instance, previous work proved the efficiency of applying confidence-based [44], [45], [46], committee-based [47], [48], [49], [50], and active label correction [51], [52] methods in detecting label noise and augmenting classification performance. While we recognize that a variety of strategies can reduce the cost of interactive labeling, this paper aims to explore the overall bounds of the problem. To achieve this, our model focuses on random instance selection,

using it as a baseline to bound our space within a naive selection strategy.

C. Interactive Machine Learning Systems Evaluation

The quality of machine learning models can be assessed through conventional metrics such as accuracy, precision, recall, F_1 score [53], etc., while visual analytics frameworks are typically evaluated by usability metrics, such as how quickly and accurately a given task can be resolved [54]. For instance, Zhang et al. [19], [55] simulated different strategies of labeling (e.g., fully manual, active learning, fully automatic) to estimate the cost of interactions in IML systems, where they used the number of interactions as the estimation of cost. Particularly, Zhang et al. [19] evaluated the number of interactions needed to achieve the desired performance in active learning scenarios. Their results indicate that not every interaction can achieve an equivalent amount of performance enhancement. Other studies tracked the user performance in the label correction process. For instance, Xiang et al. [15] and Bauerle et al. [16] recorded the number of noisy labels revised or corrected by users, and Bernard et al. [12] examined two user strategies (labeling a single or multiple instances at once) for assessing the labeling efficiency. These works partially revealed the impact of human effort on the system performance gains (e.g., classification accuracy); however, they did not quantify the impact of the human intervention or explore how environmental factors can influence the relationship between human efforts and their corresponding benefits in the context of interactive label correction.

Other human-centered evaluations are done through user studies under well-defined conditions with groups of real users, for example, studies have been conducted to measure how much time IML systems have saved while solving practical problems (digitizing data created using paper and pens [55]), how appropriate trust is established in human-machine teaming [56], [57], and the perceived interpretability of a machine learning model [58]. However, evaluation through empirical human-subject studies has intrinsic drawbacks. For example, they may be applicable only to particular domains, hard to reproduce, or expensive to conduct [59]. To overcome such issues, model-based evaluation approaches have been developed for estimating interactive system usability by modeling users' procedure of completing tasks in the system, such as GOMS techniques [60], [61], [62]. Additionally, a simulation-based evaluation framework *analytic quality* (AQ) was proposed for measuring the system performance as well as varying user insights over time with different simulated human behaviors [63]. Zhang et al. [64] applied a simulation approach to evaluate the interfaces for data labeling systems and explored the effect of simulated factors on user operation time cost, such as interface layout, application scenario, and default label accuracy. Their work assumed that users conduct *error-free* labeling sequentially, and the machine operations time is negligible. Our work also adopts a simulation-based approach under the *perfect-agent* assumption, same as the *error-free* assumption [61], [64], yet we mainly investigate the trade-off between the costs in revising noisy labels and benefits in machine learning model performance in interactive label correction scenarios across a large parameter space, including data characteristics, initial label quality, and machine learning algorithm. We aim to offer guidance to aid practitioners in pinpointing the task conditions that yield the greatest benefits from interactive label correction.

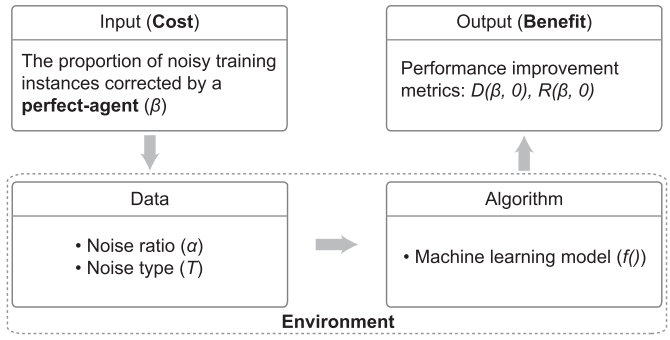

Fig. 1. The simulated process for interactive label correction. The cost refers to the expense of correcting labels, while the benefit of such correction is evaluated based on the output of the algorithm. For a given dataset, the simulation environment can be influenced by various factors: noise ratio, noise type, and machine learning model.

III. Simulation Approach

To better understand the impact of interactive label correction, evaluate the potential benefits of human intervention, and investigate the necessity of chasing perfect correction (perfect human and visual systems), we simulated humans consequently revising mislabels in the training data for various classification tasks with the *best-case scenario* assumptions. We conducted the simulation under various task conditions, across a spectrum of factors, including label noise and machine learning models, and on various types of datasets (Fig. 1). Our work explores two key questions:

- Q1. What is the relationship between the amount of effort expended in interactive label correction (costs) and the improvement in classification performance (benefits)?
- Q2. Under what task conditions are the costs of interactive label correction considered worthwhile given the benefits?

In this section, we introduce our simulation methodology, including how we created different simulation environments (i.e., task conditions), emulated the costs of interactive label correction, and evaluated the corresponding benefits for system performance upon label correction. We assume *best-case scenario* to simplify the simulation and explore the *upper-bound* of potential benefits from interactive label correction through visual analytics solutions. Our assumptions include:

- We assume *perfect-agents*, where annotators will utilize a visual analytics process that will allow them to always identify and inspect some proportion of the mislabeled training instances and accurately revise their labels to the associated ground truth labels ($correctness = 100\%$).

- We assume that the visual interface is entirely effective in revealing mislabels, aligning with the *perfect-agents* assumption.
- We assume *random selection* for the strategy of selecting mislabeled training instances for label correction, serving as our initial step to explore the overall bounds of the interactive label correction problem. Although a well-formed visual analytics framework could increase the efficiency of the label correction process (e.g., suggest revising the mislabeled instances that are most influential on classification performance enhancement), our upper-bound analysis allows us to determine if the application of visual analytics will be worth the return on investment.
- We assume a *single-iteration intervention* on cleaning mislabeled training instances, serving as a simplified implementation for the general interactive label correction process.

A. Simulation Environment

We simulated the labeling environments by varying two main parts in VIAL workflow [11]: labeled datasets and machine learning models. The datasets can be varied based on two label noise properties, namely noise ratio and noise type. With all the generated factors, we investigated the relationships between human intervention in label correction and model performance improvements. To obtain generalized results, our investigation explored one image dataset and one text dataset, sampling multiple subsets of data from each based on different classes (discussed in Section IV). Given a classification problem with a dataset and its associated label set $\mathcal{L}$, we assumed that each given data was labeled correctly initially. In order to quantitatively measure the performance gain from interactive relabeling training data, we purposely corrupt the training data to derive a contaminated training set while keeping the test set always clean. Inspired by studies on label noise in the ML community [65], [66], [67] where they tuned algorithms with various types of noise in a wide range of noise ratios, our work defined different noise conditions by diverse noise type (T) and noise ratio (α). Noise type is a function that controls the distribution of label assignments for training instances during the corruption process, and noise ratio is the proportion of mislabeled training instances among all training instances. Thus, for a given clean training data, we can create numerous corrupted training sets that vary based on the different levels of T and α. We then trained a classifier $f()$ on a contaminated training set, resulting in a classifier $f_0()$ and its accuracy acc_0 on a set of test data withheld from training. This allows us to quantify a baseline model classification performance with the contaminated training set, where no relabeling has occurred.

B. Cost Simulation

Using the contaminated training sets, we can examine the impact of humans in the interactive label correction process (i.e., *User* in VIAL workflow). To provide an appropriate framework for evaluating the potential effects of interactive label correction, we parameterized the cost of label correction in this simulation. We quantified the effort on interactive label correction by measuring the proportion of mislabeled instances corrected by a *perfect-agent*, among all noisy instances in the training set (β)(as shown in Fig. 2). We varied the levels of β in the range of (0%, 100%] with a step size of 10% to simulate a human subsequently correcting mislabeled instances in a training set, where $\beta = 0\%$ shows the baseline case without any label correction. The variable β serves as an *decision variable* of the simulation, which is directly tied to the cost in the simulated interactive label correction process. We use the proportion β instead of the absolute number of corrected instances to allow direct comparison across different dataset sizes.

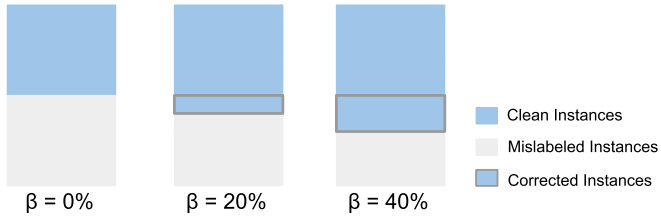

Fig. 2. An example of varying cost on label correction β (i.e., the proportion of noisy training labels corrected by a *perfect-agent*) in interactive label correction given a noise ratio ($\alpha = 0.5$). In this scenario, a corrupted training set exists where half of the training instances are mislabeled (left). A *perfect-agent* then randomly revise 20% (middle) and 40% (right) of noisy training instances and correct their labels to the ground truth.

C. Benefit Evaluation

After simulating a *perfect-agent* to correct β percent of noisy labels in a training set, we train a classifier $f_\beta()$ on the modified training data. With the classifier, we can further evaluate it on a testing set, where all testing instances are assumed to have correct labels. Our work focuses on the classification performance improvement achieved by label correction under various conditions, regarding the classification accuracy improvement (on the test set) as the primary measure of the benefits of interactive label correction. Thus, we let acc_0 denote a baseline classification accuracy without relabelling, and acc_β denote the accuracy after relabeling β percent of noisy training instances. Further, we use acc_β and the baseline accuracy acc_0 to form two evaluation metrics to assess the potential benefits of model classification accuracy after engaging effort on label clean: Improvement in Accuracy (1), and the Proportion of Reduction in Inaccuracy (2):

$$D(\beta, 0) = acc_\beta - acc_0 \tag{1}$$

$$R(\beta, 0) = \frac{acc_\beta - acc_0}{1 - acc_0} \tag{2}$$

We note that in a general visual interactive label correction process, there is a feedback loop between humans and the machine learning model. Humans subsequently clean labels in the training set, and the machine learning model is iteratively retrained on the updated training set. In this loop, training data and their labels are visualized in a graphical interface to assist humans in identifying mislabeled instances. As such, the overall cost and benefit of the human-assisted label correction would

TABLE I
THE DETAILED INTERACTIVE LABEL CORRECTION SIMULATION ENVIRONMENT SETUPS FOR TWO DATASETS: FASHIONMNIST AND AGNEWS-10PCT, INCLUDING ALL LEVELS OF MACHINE LEARNING MODEL ($f()$), NOISE TYPE (T), LABEL SET SIZE ($|\mathcal{L}|$), AND ITS CORRESPONDING CLASS LABEL SET ($\mathcal{L}$)

Dataset	Machine Learning Model ($f()$)	Noise Type (T)	Label Set Size ($\|\mathcal{L}\|$)	Dataset Size		Label Set ($\mathcal{L}$)
				Train	Test	
FashionMNIST	CNN, Random Forest, Decision Tree, LinearSVC, Logistic Regression	NCAR, NAR, NNAR	10	60000	10000	(0,1,2,3,4,5,6,7,8,9)
		NCAR, NNAR	2	12000	2000	(2, 4), (2, 6), (4, 6), (0, 6), (1, 9), (4, 7), (4, 5), (0, 7)
AGNews-10pct	BiLSTM, Multinomial Naive Bayes, SGD Classifier	NCAR, NAR, NNAR	4	12000	2000	(0,1,2,3)
		NCAR, NNAR	2	6000	1000	(0,1), (0,2), (0,3), (1,2), (1,3), (2,3)

depend on the usability of a specific visual analytics system. For example, an imperfect visual interface can misreport a correct label, which would then be inspected by an annotator. We reserve the evaluation of the visual inspection interface and imperfect label correction for future work.

IV. SIMULATION ENVIRONMENT DESIGN

We simulated *perfect-agent* to clean training mislabels through an ideal visual analytics interface under various task conditions. Different task conditions were formed by varying levels of three environmental factors hypothesized to impact the performance of a classifier in the presence of noise. These factors belong to two categories: data and algorithm (Fig. 1). The effects of environmental factors are explored by a full factorial design through the simulation. In order to investigate the generalization of the cost-benefit trade-off in visual interactive label correction, we perform the simulation on two datasets with various task conditions. We used high-performance computing nodes that are equipped with Intel Xeon E5-2680 v4 CPUs and NVIDIA A100 Tensor Core GPUs to implement the simulation in image recognition tasks with FashionMNIST dataset [68], and text classification tasks with sampled AGNews dataset [69].

A. Data

1) Dataset: To obtain a generalized quantification for the impact of interactive label correction, our simulation was implemented on two different types of datasets: one multi-class image dataset—FashionMNIST, and one natural language dataset, which is sampled from a large-scale multi-class text classification benchmark dataset—AGNews dataset. These two datasets are widely used in model robustness and label noise research [41], [65], [70], [71], [72]. The FashionMNIST dataset is a 10-class image classification dataset including pictures of fashion products (e.g., dresses and sandals). It contains 60,000 training instances and 10,000 test instances. Instances in the training and test sets are evenly distributed across ten classes. The entire AGNews dataset consists of 120,000 news articles in a training set and 7,600 news articles in a test set; instances in each set are evenly distributed across four topics: world, sports, business, and science. Due to the computational limit, we used a sampled AGNews dataset in this study by randomly sampling 10% to create training instances from each of the four classes in the AGNews dataset. Since a test set was expected to contain enough representative data instances, 500 test instances (around 26%) were randomly sampled from each class. In summary, the sampled AGNews dataset contains 12,000 training instances and 2,000 test instances. We denoted the sampled AGNews dataset as AGNews-10pct. Additionally, to better generalize the effect of label correction on datasets with varying complexity, we also implemented the simulation with binary datasets that were formed by randomly sampling two classes from the original label set $\mathcal{L}$. Specifically, eight binary subsets were sampled from the FashionMNIST dataset, and six binary subsets were sampled from the AGNews-10pct dataset. Including the original multi-class FashionMNIST and AGNews-10pct dataset, there are 16 different datasets in total for our simulation (Table I).

2) Noise-Related Factors: Since the different amounts and types of label noise have been found to have distinct impacts on classifier performance, we are interested in the influence of different label noise attributes on the relationship between costs and benefits of interactive label correction, including the *noise ratio* (α), which is the percentage of mislabeled instances in a training set, and the *noise type* (T), which defines the distribution of label errors in the training set. A statistical taxonomy of label noise [73] identifies three noise types. We simulate these noise types using noise models described by Algan and Ulusoy [65]:

Noisy Completely at Random (NCAR) Model: The NCAR model assumes that label errors occur independently of features and ground truth labels. To simulate NCAR, we generate *uniform* label noise by randomly sampling the α proportion of data instances in each class and then assigning incorrect labels to the sampled instances with equal probabilities of all other classes.

Noisy at Random (NAR) Model: The NAR model assumes that label errors occur independently of features but depend on ground truth labels. For example, instances in one class are more likely to be mislabeled as similar classes. To generate *class-dependent* noise, we first trained a neural network on the clean training set. Then, we used the classification confusion matrix on the test set as the noise transition matrix to identify the most similar class for each class. Finally, we diluted an α proportion of training instances that were randomly sampled from each class based on the confusion matrix such that similar classes are more likely to be mislabeled as each other.

Noisy Not at Random (NNAR) Model: The NNAR model assumes that label errors occur depending both on features and ground truth labels. For example, instances from different classes with similar features are more likely to be mislabeled as one another. In the simulation, we applied a label corruption method [74] to generate *localized* label noise by using a KNN model on the training data and then corrupting neighbors in a

class by assigning instances in the cluster to the same wrong labels, with the noise ratio.

As discussed in Section III, we simulate interactive label correction using synthetic label noise, starting with a dataset assumed to be clean initially. This means that each data instance was labeled with its ground truth. Then, we introduce label noise with a certain noise ratio α by artificially mislabeling the α proportion of training instances with a chosen noise type (T), while keeping the test set clean. Since synthesizing NAR and NCAR noise in the binary simulation may generate similar noisy training sets, we only corrupted the binary training sets with NCAR and NNAR noise types. Inspired by Algan and Ulusoy [65], we corrupted the training set with 13 levels of noise ratio (α) ranging from 0.05 to 0.65 with a step size of 0.05 for each noise type (T). In the simulation, we ensure that no new instances are added to the dataset upon corruption and that all noisy labels exist in the set of ground truth labels $\mathcal{L}$ such that each noisy instance can be relabeled with a known class in $\mathcal{L}$. Operations such as identifying instances of unknown classes or removing outliers are not considered in this simulation.

B. Algorithm

Finally, we explore how distinct classifiers influence results. Different machine learning algorithms $f()$ can have drastically different performances in the same classification task. Adopting an appropriate machine learning model is crucial for achieving good classification performance. In interactive label correction, the classification performance improvement is not only affected by the noise of a training set after label correction but also the robustness of an algorithm under the label noise conditions. As such, we investigated how the label correction process improves the model quality of different machine learning models by using several basic and popular classifiers for each dataset (Table I). Since CNN is a popular and commonly used model for image classification [75], we build a CNN model with five hidden layers for the FashionMNIST dataset. We also examined four other benchmark classifiers for the FashionMNIST dataset as explored in Xiao et al. [68]: Random Forest (RF), Decision Tree (DT), LinearSVC, and Logistic Regression (LR). For the AGNews-10pct dataset, we trained three other representative classifiers: Bidirectional LSTM (BiLSTM) [76], Multinomial Naive Bayes (MNB) [77], and SGD Classifier [78], which are well studied in the text classification domain with the original AGNews dataset. We specifically tuned these models with the AGNews-10pct dataset, the sampled 10% of AGNews training data. The details of model parameters and the code for model training on our simulated datasets are included in the supplemental materials.[1] Since models with similar architecture may exhibit similar robustness attributes [79], our work starts with simple yet fundamental ML models derived from different architectures (e.g., tree-based, neural-network-based) for both datasets. This serves as our initial exploration into the cost-benefit relationship of interactive label correction. Our goal is to benchmark the impact of interactive label correction on enhancing model quality and to investigate its generalizability across diverse ML model architectures. In future simulations, advanced ML techniques for image and text classification will be incorporated, such as VGG for the fashionMNIST dataset and BERT for the AGNews dataset.

[1][Online]. Available: https://github.com/VADERASU/cost-benefit-interactive-label-correction

V. Results

In this section, we present the analysis of the simulation results and highlight the key findings observed during the simulation. Table II provides a summary of the simulated factors and the evaluation metrics used in our simulated interactive label correction process.

A. Statistical Analysis

As noted in Section III, we simulated varying cost levels for interactive label correction by controlling the proportion of mislabeled instances corrected by a *perfect-agent* (β) and measured the corresponding benefits in terms of two metrics: $D(\beta, 0)$ and $R(\beta, 0)$. In this work, we employed one-way and two-way ANOVA (analysis of variance) models with these two evaluation metrics as response variables. Since we aimed to examine both single-factor effects and associated interaction effects of simulated factors on the classification performance improvement during the interactive label correction process, we employed Type III Tests of Fixed Effects (F-tests) to analyze the simulation result. Specifically, to answer **Q1**, we studied the relationship between the effort expended in interactive label correction (i.e., costs) and the associated model performance improvement (i.e., benefits) by examining if the same level of cost consistently leads to the same level of benefits. We used the difference between two consecutive β values as response values for each evaluation metric, denoting the benefits of each 10% label correction. For instance, $R(20\%, 0) - R(10\%, 0)$ denotes the benefits of further correcting another 10% mislabels after 10% label noise has been corrected by a *perfect-agent*, considering the metric $R(\beta, 0)$. With regard to **Q2**, we constructed generalized linear models (JMP version 17) using a full factorial combination of three environmental factors (noise ratio, noise type, and machine learning model) and the variable β with $D(\beta, 0)$ and $R(\beta, 0)$ as response variables. We built binary and multi-class classification separately for both FashionMNIST and AGNews-10pct with respect to the two response variables, resulting in $2 \times 2 \times 2 = 8$ statistical models. Using these models, we identified the significant influence of the environmental factors and their interaction effects on the cost-benefit relationship of interactive label correction (Appendix Table A.1). In the subsequent analysis, we discuss the simulation results for the FashionMNIST and AGNews-10pct datasets, and illustrate the varying trends in cost and benefit using average raw data values (associated standard deviation values are attached in supplemental materials). Due to the different levels of noise types simulated in binary and multi-class classification simulations, we report their results separately for each dataset. As listed in Table I, for the simulation of FashionMNIST image classification, we explored 13 noise ratios, five models, and ten values of β. Multi-class simulation

TABLE II
KEY TERMINOLOGIES AND NOTATIONS USED IN THE SIMULATION

Notation	Description
β	The proportion of mislabeled instances corrected by a *perfect-agent* among all noisy instances in the training data, measuring the cost of human effort in simulated interactive label correction process. In the simulation, $\beta = \{0\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$
$D(\beta, 0)$	Improvement in classification accuracy of a machine learning model, measuring the benefit of simulated interactive label correction process.
$R(\beta, 0)$	The proportion of reduction in the inaccuracy of a machine learning model, measuring the benefit of simulated interactive label correction process.
Z	Average benefit-cost ratio, measuring the average benefit ($D(\beta, 0)$ or $R(\beta, 0)$) gained from a single cost (β) unit in the simulated label correction process.
$f()$	Machine learning classifiers that categorize data into predefined classes based on learned patterns in the training data. The simulation includes eight different machine learning models: CNN RF DT LR LinearSVC BiLSTM MNB SGD
α	The noise ratio, controlling the amount of label noise in the training data. In the simulation, $\alpha = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65\}$
T	The noise type, controlling the distribution of label error in the training data. The simulation includes three different noise types: NCAR NAR NNAR
Binary Classification	A type of machine learning task where the goal is to categorize data into one of two distinct classes.
Multi-class Classification	A type of machine learning task where the goal is to categorize data into one of several possible classes.

For further details, please refer to sections III and IV.

trials were simulated on one 10-classes dataset (i.e., the original FashionMNIST dataset) with three noise types, resulting in $13 \times 5 \times 10 \times 1 \times 3 = 1950$ conditions. Binary simulation trials were executed on eight 2-classes datasets with two noise types (NCAR and NNAR), resulting in $13 \times 5 \times 10 \times 8 \times 2 = 10400$ conditions. For the AGNews-10pct text classification, we explored 13 noise ratios, three models, and ten values of β. Multi-class simulation trials were simulated on one 4-class dataset (i.e., the original AGNews-10pct dataset) with three noise types, resulting in $13 \times 3 \times 10 \times 1 \times 3 = 1170$ conditions. Binary simulation trials were executed on six 2-classes datasets with two noise types (NCAR and NNAR), resulting in $13 \times 3 \times 10 \times 6 \times 2 = 4680$ conditions. We performed ten repeated measures of each condition by randomly shuffling the training data, generating $(1950 + 10400) \times 10 = 123500$ trials for the FashionMNIST dataset and $(1170 + 4680) \times 10 = 58500$ trials for the AGNews-10pct dataset. Among them, 760 trials of the FashionMNIST dataset were removed where $acc_0 = 100\%$, as label cleaning was not expected to increase the classification accuracy when the initial accuracy was already at 100%. Thus, 122740 FashionMNIST trials and 58500 AGNews-10pct trials in total were used for analysis.

B. Impact of Interactive Label Correction

Our results show that simulated interactive label correction improves the machine learning model classification performance in 97.4% of FashionMNIST multi-class trials, 95.4% of FashionMNIST binary trials, 95.8% AGNews-10pct multi-class trials, and 94.0% AGNews-10pct binary trials. However, in the other trials, the simulated *perfect-agent* caused a decrease in the classification accuracy on test data, with averages of $D(\beta, 0) = -0.2\%$, $R(\beta, 0) = -1.2\%$ in FashionMNIST multi-class simulation, $D(\beta, 0) = -0.5\%$, $R(\beta, 0) = -26.3\%$ in FashionMNIST binary simulation, $D(\beta, 0) = -0.7\%$, $R(\beta, 0) = -3.3\%$ in AGNews-10pct multi-class simulation, and $D(\beta, 0) = -0.4\%$, $R(\beta, 0) = -7.9\%$ in AGNews-10pct binary simulation. The accuracy deterioration mostly occurred in a trial with a small noise ratio (e.g., $\alpha \leq 0.1$), where the model baseline accuracy (acc_0) was relatively high. It may be due to overfitting (a training set is overly clean) or a model itself being too sensitive to outliers. These examples suggest that **there are risks for correcting labels, and interactive label correction could reduce model accuracy when the noise ratio is small.**

To understand the impact of interactive label correction, we varied the β values and examined the trend of average performance gain as β increases. The value of β quantifies the effort expended by a *perfect-agent* on revising mislabels, and we explored ten values of $\beta = \{10\%, 20\%, \ldots, 100\%\}$ in our simulation. As shown in Fig. 3, the values of $D(\beta, 0)$ and $R(\beta, 0)$ grow as the value of β increases in both the binary and multi-class classification simulations. This is expected since larger β values indicate more mislabeled training labels were corrected for a given dataset. However, as β increases, the growth of both benefit metrics gradually flattens out, which implies a decreasing trend of marginal performance gains (Appendix Table A.2). In other words, each additional unit increase in β yields a smaller average performance gain than the previous unit. This pattern becomes more evident as the noise ratio increases. To examine the decreasing trend, we further built a one-way ANOVA model using β as the independent variable, with marginal performance gains (i.e., $D(\beta_1, 0) - D(\beta_2, 0)$; $R(\beta_1, 0) - R(\beta_2, 0)$) as the response variables. Regarding ten values of β, the particular $D(\beta_1, 0) - D(\beta_2, 0)$ values used in the statistical test are two consecutive β values such as $D(20\%, 0) - D(10\%, 0)$, $D(30\%, 0) - D(20\%, 0)$, for exploring the effect of each further 10% label correction. Type III Tests of Fixed Effects were conducted to compare $D(\beta_1, 0) - D(\beta_2, 0)$ and $R(\beta_1, 0) - R(\beta_2, 0)$ for all trials, confirming the significant difference of nine consecutive pairs of β values (Appendix Table A.3). Moreover, Tukey's HSD tests were applied to obtain the pairwise comparison. The results show that, in general, as β increases, the associated $D(\beta_1, 0) - D(\beta_2, 0)$ values (i.e., marginal performance gains) are significantly decreasing.

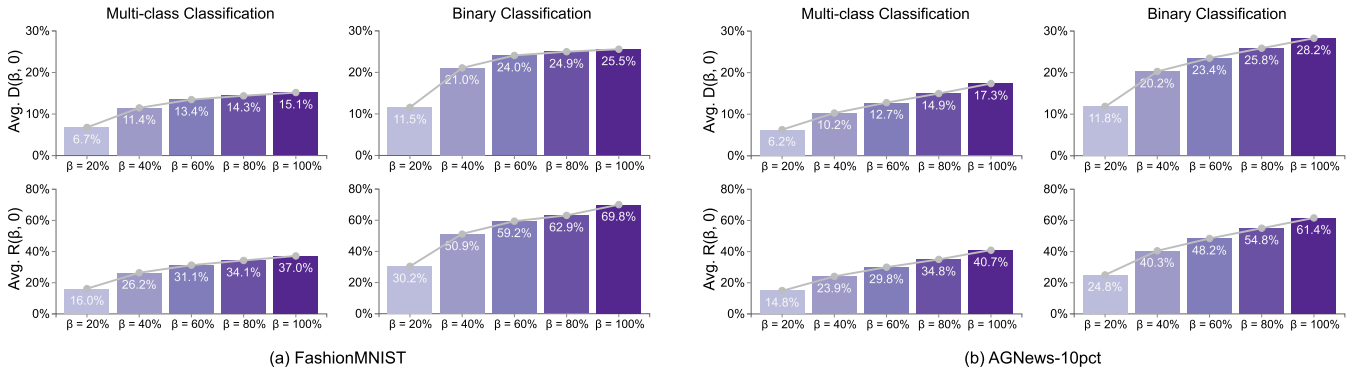

Fig. 3. Average classification performance improvement (y-axis) with increased cost for label correction (x-axis) for the (a) FashionMNIST dataset and (b) AGNews-10pct dataset. Performance improvement increases asymptotically with β in simulations for both binary and multi-class classifications.

Overall, the amount of interactive label correction does not have a linear positive relationship with the improvement of a classifier. The simulation of perfect-agent interactive label correction on both the FashionMNIST and AGNews-10pct datasets exhibits a decreasing effect on the marginal absolute ($D(\beta, 0)$) and relative ($R(\beta, 0)$) potential performance gains with every additional 10% label noise correction. This trend also implies that there might exist an optimal β value (i.e., cost on label correction) that can achieve the expected net benefit (i.e., total benefit after subtracting all cost) before fully cleaning the label noise in the training data. It might be due to the size of the label noise decreases with more label errors being corrected, and a classifier can usually mitigate the impact of noisy instances by smoothing or ignoring a small number of outliers due to its robustness.

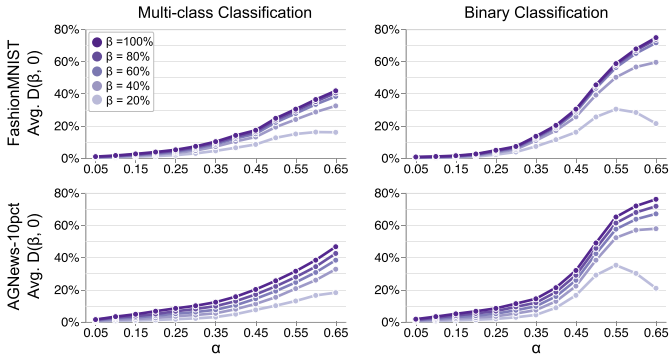

Fig. 4. Average classification performance improvement ($D(\beta, 0)$) with an increased value of noise ratio (α) and different levels of β (shades of purple) for label correction on the FashionMNIST and AGNews-10pct datasets. All β values exhibit increasing performance improvements with higher α, but for lower β values, this improvement is notably limited.

C. Impact of Environmental Factor

When inspecting the intersection of significant terms in the full factorial design analysis (Appendix A Table A.1), we identified environmental factors that interplay with the cost of label correction (β). Type III Tests of Fixed Effects proved statistical significance in both evaluation metrics for the interaction effects between environmental factors in both FashionMNIST and AGNews-10pct simulations (Appendix A Table A.4– A.11). Our analysis results show that the interaction effect of $\alpha \times \beta$ is the most influential term for the average performance improvement, as it directly determines the amount of mislabels modified by annotators for a given dataset. Moreover, we discovered interplays between the β and the interaction terms of noise ratio and noise type ($\alpha \times T$), as well as between noise ratio and model ($\alpha \times f()$). Their interaction effects had a more substantial impact on enhancing classification performance compared to just the T or $f()$ alone. These statistically significant terms confirmed that environmental factors considerably influence the cost-benefit relationship of interactive label correction. Further details about these impacts are described later in this subsection.

Noise Ratio: To understand how noise ratio affects the relationship between label correction and derived performance gains, we inspected the change in two benefit metrics across diverse β values with varying noise ratios (α). As illustrated in Fig. 4 and Appendix Figure C.1, **the average classification performance gains, $D(\beta, 0)$ and $R(\beta, 0)$, generally increase as the noise ratio rises.** The trends observed for $\beta = \{40\%, 60\%, 80\%, 100\%\}$ are more or less similar: the slopes become steeper in the middle range and then flatten out for both binary and multi-class classification. Yet, with $\beta = 20\%$, the performance gain grows more slowly than $\beta = \{40\%, 60\%, 80\%, 100\%\}$. Further, both $D(\beta, 0)$ and $R(\beta, 0)$ flatten out in multi-class classification and there are even significant drops at the end for binary classification. **The pattern reveals that a smaller degree of human effort (e.g., $\beta = 20\%$ or 40%) under large noise ratios may not lead to sufficient classification performance improvement**, especially in binary classification, because the remaining large proportion of label noise in the training set could still severely affect the performance. Conversely, correcting at least 50% of label noise is more likely to attain relatively satisfying performance gains.

We also analyzed two significant interaction effects, noise type $\times$ noise ratio ($T \times \alpha$) and machine learning model $\times$ noise ratio ($f() \times \alpha$), and explored how they interplay with the cost of revising labels (β). Specifically, we compare the trends associated with noise type (T) or machine learning models ($f()$) under diverse noise ratios and discuss the impact of the

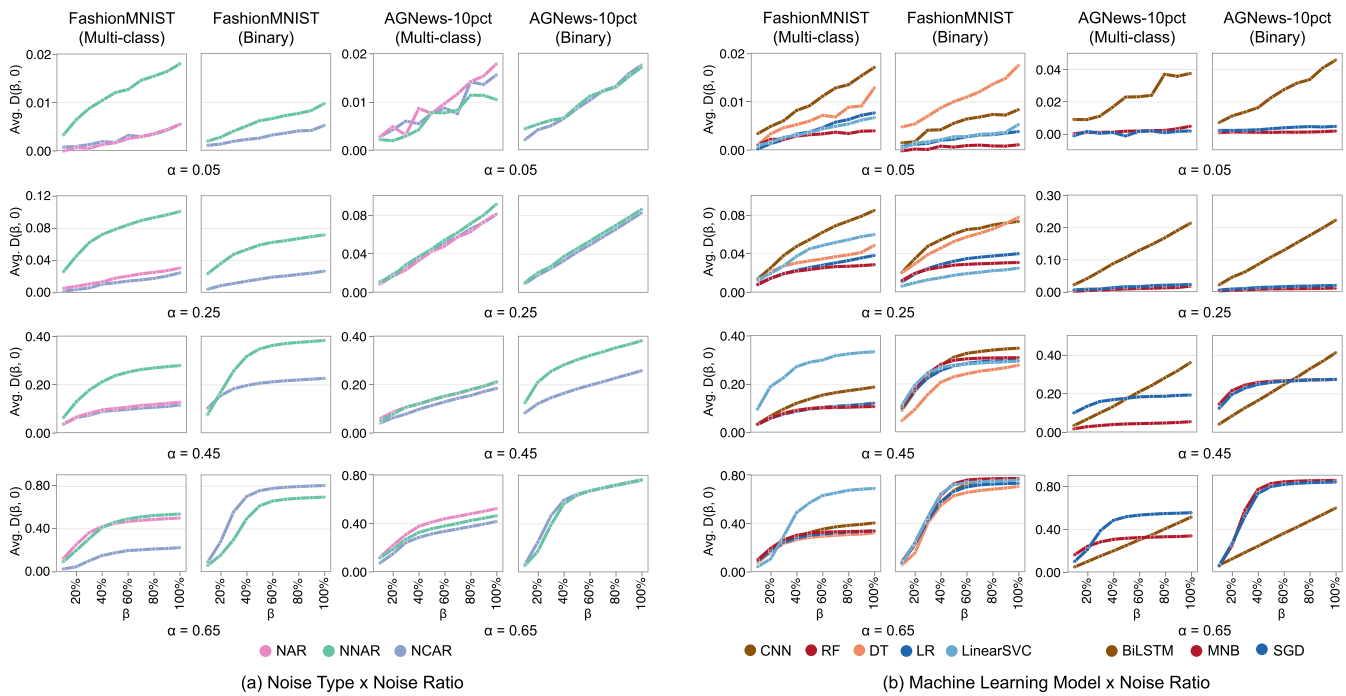

Fig. 5. Small multiple plots illustrating two interaction effects: (a) Noise Type $\times$ Noise Ratio and (b) Machine Learning Model $\times$ Noise Ratio. Each chart shows the relationship between cost (β values) on the x-axis and derived benefit (average $D(\beta, 0)$) on the y-axis. Charts are organized by noise ratios $\alpha = \{0.05, 0.25, 0.45, 0.65\}$ (rows) and classification settings (columns), showcasing how interaction effects affect the cost-benefit relationship.

interaction effect on the relationship between costs and benefits of interactive label correction. Fig. 5 presents the average improvements in classifier performance for different T and $f()$ in our simulation, with regard to the metric $D(\beta, 0)$. The patterns observed for the metric $R(\beta, 0)$ are similar and can be found in Appendix Figure C.2.

Noise Type $\times$ Noise Ratio: As shown in Fig. 5(a), we observed linear increasing trends as β increases for both evaluation metrics under the small noise ratio in both binary and multi-class simulations for FashionMNIST dataset, regardless of the noise type. Nonetheless, the magnitude of these increases is minimal, almost negligible. As the noise ratio increases, the growth in benefits starts to flatten out, especially when correcting more than 40% of label noise in the training data, revealing a diminishing effect of further correcting label noise across noise types. When comparing the difference in benefits gained from interactive label correction among noise types, for multi-class classification, cleaning NCAR and NAR label noise yields similar benefits on performance enhancement, yet correcting NNAR mislabels can achieve a much greater benefit than NCAR and NAR when $\alpha \leq 0.45$ for the average $D(\beta, 0)$ and $\alpha \leq 0.3$ for the average $R(\beta, 0)$. This is likely because models are less robust to the NNAR noise type [65], compared to the other two noise types. Interestingly, as α grows beyond these thresholds, the benefits of revising NAR noise reach close to and even can exceed the benefits of revising NNAR noise when a certain proportion of training mislabels have been revised. This pattern indicates that it becomes more worthwhile to clean NAR noise when more than half of the training instances are mislabeled. It may be because mislabeling over 50% training instances to their most similar classes, i.e., NAR label noise, might alter the decision boundaries of the classifier the most seriously compared to the other two noise types. On the other hand, revising NCAR noise consistently generates the least average performance gains in multi-class simulation. Similarly, in binary simulation, revising NNAR label noise was found to yield significantly greater performance gains than NCAR when $\alpha < 0.5$. Yet, the small amount of effort in label correction for NNAR noise becomes less beneficial in improving classifier performance compared to NCAR noise type when the noise ratio is getting larger. Along with the noise ratio rising to 0.65, we observed that revising NCAR noise is more beneficial for augmenting classification performance than NNAR across most levels of β for both evaluation metrics. It may be because mislabeling a larger portion of uniformly distributed data in a binary classification dataset makes it harder to create an accurate decision boundary compared to mislabeling only localized data. Unlike in multi-class simulation, where NCAR noise consistently results in the lowest average performance gain, in binary simulation, correcting NCAR noise can, with an increasing noise ratio, potentially surpass the performance gain achieved by correcting NNAR noise.

Binary and multi-class simulation results for AGNews-10pct data show a similar benefit growth trend among all noise types. Specifically, when $\alpha \leq 0.35$, the performance gains (both

$D(\beta,0)$, $R(\beta,0)$) rise steadily with increased label correction cost. However, for larger noise ratios, the increase in performance gains started to slightly flatten out as β increases. In the multi-class simulation, when $\alpha \leq 0.3$, the performance improvement obtained from cleaning distinct types of noise is almost indistinguishable. However, revising NNAR and NAR noise becomes more beneficial as the noise ratio grows over 0.3, then yields significantly greater performance gains than modifying NCAR noise when α reaches 0.45. By the time the noise ratio increases to 0.65, we observed that label correction on NAR noise generally attains the greatest performance gains, especially for $D(\beta,0)$. Likewise, results of binary simulation also indicate that revising NCAR and NNAR noise can yield a similar level of performance gains under $\alpha \leq 0.25$ for $D(\beta,0)$ and $\alpha \leq 0.15$ for $R(\beta,0)$. When the noise ratio becomes larger, modifying NNAR noise shows more potential for achieving performance enhancement than NCAR noise. Surprisingly, by around $\alpha = 0.6$, the impact of revising NCAR noise again reaches close to, and then exceeds, that of NNAR. **Overall, in both FashionMNIST and AGNews-10pct simulations, the same level of cost on label correction can result in different levels of performance gains under distinct noise conditions, in terms of noise ratio (α) and noise type (T). The two noise-related factors interplay on the performance gains.** In general, cleaning NNAR label noise is more effective in improving model quality than cleaning the other two noise types (NAR, NCAR) when less than half of training data are mislabeled. Yet, when more label noise exists, the noise types that are more likely to obscure boundaries between classes (NAR for multi-class, NCAR for binary) are more harmful to the model performance. Thus, revising these mislabels becomes more crucial for enhancing model performance.

Machine Learning Model $\times$ Noise Ratio: In the FashionMNIST simulation, we explored the effect of label correction across five models: CNN, Random Forest (RF), Decision Tree (DT), LinearSVC, and Logistic Regression (LR). As Fig. 5(b) shows, the FashionMNIST multi-class simulation results show that CNN achieved the greatest improvement in performance when $\alpha \leq 0.25$. We also observed that the enhancement of the LinearSVC model classification performance is close to that of CNN around $\alpha = 0.3$, and LinearSVC benefitted the most from label corrections when $\alpha \in [0.35, 0.55]$ and $\alpha \in [0.4, 0.5]$ for $D(\beta,0)$ and $R(\beta,0)$ correspondingly. When the noise ratio becomes larger than these ranges, the impact of a relatively small cost (e.g., $\beta < 50\%$) on label correction for the LinearSVC model decreases. The patterns of RF, DT, and LR models are more or less similar, where revising label errors normally contributes to comparably less performance improvement for these classifiers. Conversely, the binary simulation results on FashionMNIST data show a different pattern that the interactive label correction has the most remarkable impact on the DT model when α is relatively small. As the noise ratio rises, the effect of revising mislabels on performance gains for other models reaches close to and even exceeds the DT model. When $\alpha \geq 0.45$, the DT model attains the least performance gain with regard to both evaluation metrics, while the other four models (CNN, RF, LR, LinearSVC) yield a similar level of impact. Surprisingly, we notice that when the noise ratio is small ($\alpha = 0.05$), correcting no more than 20% of the mislabeled data for the CNN model, and no more than 90% of the mislabeled with the LinearSVC model provides limited benefits and may even undermine the model classification performance with regard to the average $R(\beta,0)$.

In the AGNews-10pct simulation, we adopted three other models: BiLSTM, MNB, and SGD. In multi-class simulation, we noticed that the same effort for label cleaning typically yields greater benefits to the BiLSTM model when $\alpha \leq 0.3$, while label correction results in a smaller magnitude of performance enhancement (compared to BiLSTM) but similar gains for the MNB and SGD classifiers. As the noise ratio increases beyond the threshold, the benefits of cleaning noisy labels for the SGD classifier keep growing and eventually resulting in the greatest performance gains, yet label correction still generates the least benefits for the MNB model when $\alpha \leq 0.45$. Upon exploring different cost levels of label correction, we noticed an interesting pattern under relatively large noise ratios, particularly when $\alpha > 0.5$: cleaning a small proportion of mislabels can yield the greatest average performance gain for the MNB model compared to other models, but the least for the BiLSTM, while the benefits of label correction for the BiLSTM model surge as β increases, and eventually exceed the MNB model. The binary simulation of AGNews-10pct data reveals a similar pattern for all three models, except that label correction always attains a similar level of benefits for the MNB and SGD classifiers across all levels of β.

In summary, even though the patterns of varying models of FashionMNIST and AGNews-10pct dataset are not directly comparable due to their use of entirely distinct sets of machine learning models, **both simulation results suggest that the same extent of label correction yields different magnitude of benefits (i.e., model performance enhancement) depending on the type of machine learning models ($f()$). Those effects vary along with the noise ratio (α).** Notably, we observed the diminishing effect of marginal performance gains of further correcting label noise for most machine learning models in both FashionMNIST and AGNews-10pct datasets, except for the BiLSTM model, where a steadily increasing linear effect on benefits of performance gains as β increases was observed.

D. When to Apply Interactive Label Correction

To drill down into which task conditions benefit the most from interactive relabeling, we further explore how three environmental factors (noise ratio, noise type, and machine learning model) jointly affect the relationship between cost and corresponding benefit. Inspired by previous works [80], [81], we derived a benefit-cost ratio to evaluate which task conditions benefit the

most from interactive label correction. We used our simulated cost (β) and one of the benefit measures ($D(\beta, 0)$ or $R(\beta, 0)$) to calculate the average benefit-cost ratio (Z) across ten repeated measures for each environmental condition:

$$Z = \frac{\sum_{i=1}^{10} Benefit_i}{10 \times \beta} \tag{3}$$

This ratio describes the average benefit gained from a single cost unit and can reveal the trade-off between the cost and benefit, for a given task condition in our simulation. Fig. 6 shows the average benefit-cost ratio in all simulated conditions, with regard to the benefit metric $D(\beta, 0)$ (Appendix Figure C.3 for $R(\beta, 0)$). Across all conditions, there is a common pattern where the effect of cleaning label noise generally diminishes as more label errors are corrected, while the diminishing effect is more pronounced under larger noise ratios. Moreover, we found that it is less worthwhile to correct label noise under small noise ratios (e.g., $\alpha \leq 0.2$), compared to larger noise ratios. However, when there are more noisy labels in a training set, more label corrections (larger β value) are needed to achieve the optimal benefit-cost ratio. For example, modifying 10% of label noise can generally achieve the greatest benefit-cost ratio when $\alpha = 0.45$, while cleaning just 10% of label errors is not sufficient to attain the best benefit-cost ratio at $\alpha = 0.65$. These patterns indicate that a sufficient amount of clean data can aid in building a well-performing classifier, even if some label noise remains in the training data. However, the benefit-cost ratio exhibits slightly different trends across different noise types, models, and datasets (FashionMNIST vs. AGNews-10pct, multi-class vs. binary). For instance, in the FashionMNIST multi-class simulation, revising NCAR noise yields much greater benefits in improving model quality for the LinearSVC model than other models; in the simulation on AGNews-10pct data, label correction under small noise ratios is more beneficial for the BiLSTM model compared to others. Additionally, the average performance gain for the BiLSTM model shows a linear increase as β grows.

Overall, the findings are consistent with the discussion in Section V-B and V-C. **This phenomenon also indicates that when designing visual analytics systems, we would also need to consider the costs in the design process. Though perfect visual analytics tools can assist people in detecting all wrongly assigned labels effectively, the costs for developing such tools can be too high and may not be worthwhile considering the benefits gained from the model performance gain. A reasonable, well-formed visual analytics system may provide the most cost-effective solutions for label correction tasks. Also, in the correction process, it is crucial to raise people's awareness about the delicate balance between human effort (cost of correcting labels and developing visual analytic interface) and potential gains (benefits of model performance improvement).** For instance, when annotators encounter varying environmental conditions during relabeling tasks, the visual interface should empower them to thoroughly examine the cost-benefit relationship under different scenarios (e.g., given noise ratio, noise type, and machine learning model). This allows for informed decision-making and facilitates comparative analysis.

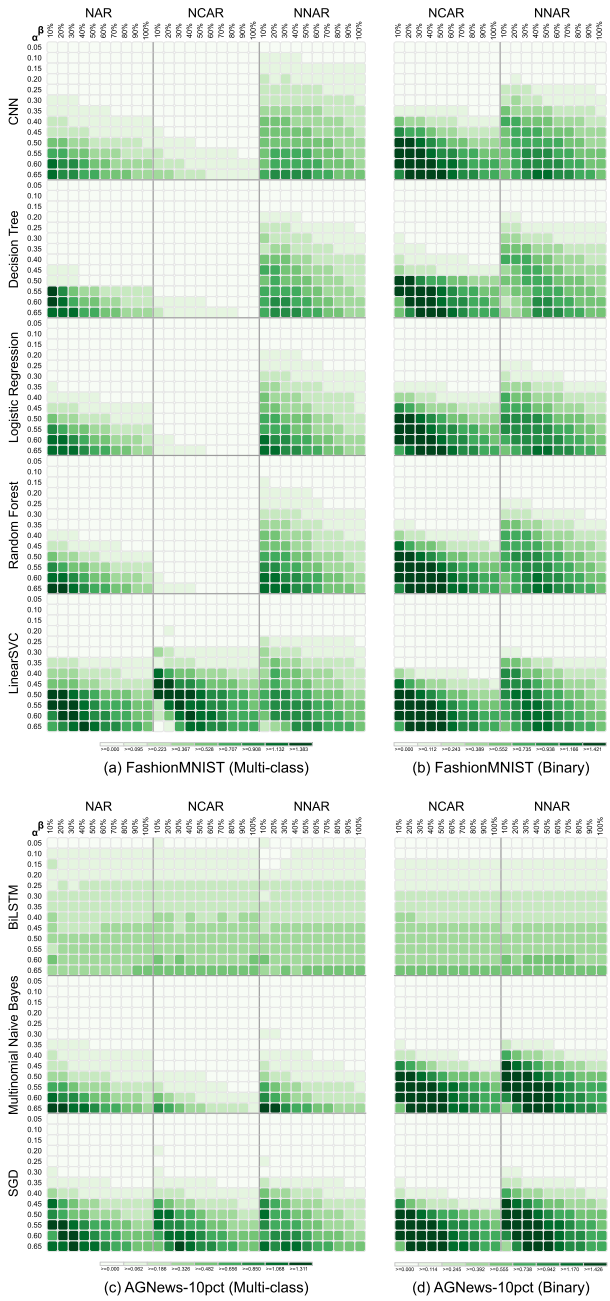

Fig. 6. Comparison of average benefit-cost ratios (Z), defined as the average performance gain of $D(\beta, 0)$ achieved by per unit of cost on label correction, among all conditions consisting of diverse levels of noise types (x-axis), and noise ratios and machine learning models (y-axis) in FashionMNIST and AGNews-10pct simulations. The color scale encodes the value of the benefit-cost ratio after a logarithmic transformation, where darker shades of green denote higher Z values, denoting conditions that are more worthy of interactive label correction.

VI. Real-World Label Noise Validation

We validated the results of the simulation of interactive label correction on real-world noisy datasets—CIFAR-10N datasets,[2] which include CIFAR-10 train images with five label sets that contain different levels of human-annotated noisy labels,

[2][Online]. Available: http://www.noisylabels.com

namely, CIFAR-10N Aggregate ($\alpha = 0.09$), CIFAR-10N Random1 ($\alpha = 0.17$), CIFAR-10N Random2 ($\alpha = 0.18$), CIFAR-10N Random3 ($\alpha = 0.18$), and CIFAR-10N Worst ($\alpha = 0.4$). Labels in various sets are chosen using distinct mechanisms from three annotators; thus these five noisy datasets vary in the quantity and the structure of label noise. Similar to the simulation of FashionMNIST and AGNews-10pct datasets, we adopted the same ten levels of human cost on cleaning training mislabels for each CIFAR-10N dataset, i.e., ten β values, and implemented ten repeated measures for each β value. Besides, each CIFAR-10N dataset was trained using a 22-layer CNN model, which is well-performed on the original CIFAR-10.

In the examination of interactive label correction on real-world noisy datasets, we found that cleaning label noise generally improves the classification performance with respect to both established benefits metrics $D(\beta, 0)$ and $R(\beta, 0)$. However, a *perfect-agent* may offer limited improvement and even undermine the model performance when the noise ratio is small (e.g., in CIFAR-10N Aggregate and CIFAR-10N Random 1, 2, 3 datasets). Moreover, **there is also no linear positive relationship between the expended effort on label noise correction (β) and the performance gains considering both evaluation metrics, as well as the benefit-cost ratio (Appendix D): as more label noise is corrected, the potential benefits from further label corrections can be reduced.** We observed a general diminishing effect on the marginal performance improvement of revising the same amount of label noise, especially when a dataset with larger noise ratios was applied (e.g., CIFAR-10N Worst). In contrast, when the noise ratio is relatively small, the magnitude of the performance gains is negligible (e.g., CIFAR-10N Aggregate). Given the real-world label noise was produced by human annotators, replicating all the conditions used in the synthesized datasets for real-world scenarios becomes challenging. This complexity impedes the ability to conduct one-to-one comprehensive comparisons between synthetic noisy data simulations and real-world noisy data validations. However, the relationship between the cost and benefit of correcting mislabels on real-world noisy datasets demonstrated similarity with the FashionMNIST and AGNews-10pct datasets, though the pattern of cost-benefit association varied slightly when different datasets were applied (e.g., CIFAR-10N Random1 vs. CIFAR-10N Random2, CIFAR-10N datasets vs. FashionMNIST datasets).

VII. Discussion

Through simulating interactive label correction on synthetic noisy datasets and validating the simulation result on real-world noisy datasets, our work reveals consistent patterns in the cost-benefit relationship of interactive label correction. This provides practical implications for practitioners to perform a cost-effective label clean through the VIAL process.

A. Implications for Applying Interactive Label Correction

Our simulations show that a *perfect-agent* annotator can generally improve the classification performance. Unsurprisingly, the more label noise being corrected, the better the classification performance. However, the relationship between the cost of revising labels and the gained benefits on model quality is not linear, especially under large noise ratios: the marginal performance improvement by label correction can gradually diminish as the expended effort on label correction (β) increases, which indicates that **there is a trade-off between the cost of revising labels and the gained benefits on model quality**. Thus, it is important to consider the efficiency of large-scale human-assisted label correction since the benefits of label correction may not always compensate for the human labor cost of correcting labels. Also, a certain amount of label error correction can yield a significant improvement in classification accuracy, but an optimal stopping point should exist before a complete label inspection and correction. This implication might also be applicable for general labeling problems (i.e., labeling unlabeled training data), especially when involving semi-supervised learning or weak supervision techniques, where a model can be well-trained using a limited amount of labeled data.

The cost-benefit trade-off is exhibited in both synthetic data simulations (FashionMNIST, AGNews-10pct datasets) and real-world noisy data validations (CIFAR-10N datasets). The consistent findings across diverse data characteristics and noise structures suggest that the cost-benefit tradeoff observed during interactive label correction may be a universal phenomenon applicable to a broader range of datasets in different domains. However, the trend slightly varied when different datasets were applied, indicating the importance of considering the context (e.g., multi-class vs. binary data, image classification vs. text classification) of classification tasks.

The relationship between the cost and benefit of interactive relabeling varied along with the simulated environmental condition changes, i.e., noise ratio (α), noise type (T), and machine learning model ($f()$). Moreover, there are active interaction effects between simulated environmental conditions on the classification accuracy improvement (e.g., noise ratio $\times$ noise type, noise ratio $\times$ machine learning model). In our work, we assess the efficiency of human-assisted label correction by accuracy improvement, and it was directly calculated by the baseline accuracy acc_0 and the accuracy after correcting β proportion of label noise acc_β. As stated, the acc_0 decreases as the noise ratio grows. The accuracy deterioration is a non-linear decline and varies with the robustness of different machine learning models to distinct amounts (α) and distribution (T) of label noise. (Appendix B). A classifier may mitigate the attacks generated by those noisy labels and provide accurate predictions under relatively small noise ratios. Yet, when the noise ratio exceeds a certain threshold, a trained classifier may be overwhelmed by a large number of label errors and yield predictions even worse than random guesses. Different models have distinct thresholds of tolerance for noise, i.e., model robustness, due to their inner classification mechanism. Therefore, the classification performance improvement is not only affected by the label noise of a training set after label correction but also by the robustness of a classifier under different label noise conditions. This highlights the importance of considering task complexity (e.g., label noise, machine learning models, dataset context) when estimating the impact of cleaning label noise.

Our simulation under the *perfect-agent* assumption aims to estimate the *upper-bound* of the impact of human intervention in the interactive label correction process. We believe that the cost-benefit trade-off can also be applicable to scenarios where humans or visual interfaces are likely to be imperfect for interactive label correction. In these cases, achieving the same level of benefits, such as performance gains in our simulation, may require more human effort on label correction compared to our idealized *perfect-agent* simulation. To be detailed, when the noise ratio is greater than 0.45, we see the largest performance gains when revising a certain amount of mislabels, especially for the type of label noise that more severely distorts the decision boundary of a model, e.g., class- dependent noise (NAR) in multi-class simulation, and uniform noise (NCAR) in binary simulation. However, the real-world dataset that needs human-assisted label correction might only contain equal or less than 40% mislabeled instances [82]. From our simulation and real-world noisy data validation, we identified that **cleaning mislabels by a *perfect-agent* may only provide little enhancement and could potentially harm the model performance under small noise ratios**, especially when the noise ratio is less than 0.2. Depending on the criticality of the machine learner and the task itself, these gains may be important, and we recommend practitioners make decisions on how to intervene in label correction based on their own initial task conditions, expected costs, and associated benefits, as well as the risk of undermining the model performance.

There are slight differences between performance gains in terms of two evaluation metrics $D(\beta, 0)$ and $R(\beta, 0)$ under various conditions. Thus, the choice of deploying which evaluation metric and its success threshold depends on the ultimate goal of practitioners, and these choices will also determine the conditions worth human intervention for cleaning label noise. Indeed, one may wish to consider $1 - acc_0$ prior to undertaking label correction to understand the potential benefit from label correction.

B. Implications for Visual Interactive Labeling System

Although our work did not incorporate an actual visual labeling interface, the effectiveness of visualization becomes apparent through the way annotators select noisy labels for cleaning. Our simulation results suggest that the same level of label correction can result in different magnitudes of performance improvement depending on the task conditions, e.g., noise type and machine learning models. Hence, a well-designed visualization interface can assist annotators in promptly identifying a set of noisy data points, the modification of which could lead to a substantial model performance improvement [12]. An ideal visual interface is expected to foster annotators to detect and correct mislabels efficiently and reduce the cost of human effort on label correction. However, given the existence of an optimal stopping point for label correction, it is not always necessary for a labeling visual interface to be of optimal quality, especially when considering the costs associated with developing such an advanced visualization. This raises a strategic consideration for practitioners: allocating resources toward additional human labor for label correction or improving the visual interface should be guided by the potential for net profit increases. Lastly, a typical visual interactive label correction process usually consists of multiple iterations where annotators continuously update the training data and re-train the machine learning models. We recommend employing visualizations that allow annotators to monitor the progress of label correction and associated performance improvement. As such, annotators can be aware of the diminishing returns of further corrections and make informed decisions about when to cease the label correction, thus avoiding unnecessary costs.

C. Implications for General Interactive Machine Learning

Our research revealed a trade-off between the cost of correcting mislabels and the resulting benefits on model performance during interactive label correction, observed in both synthetic noisy datasets and real-world noisy datasets. This cost-benefit tradeoff likely arises from intrinsic limitations in human intervention, task complexity, and the probabilistic nature of machine learning models. Initially, human input can significantly improve model quality, especially when addressing the most impactful errors or refining key features. However, as the model becomes more robust and refined, the remaining tasks become increasingly nuanced and challenging for human intervention, leading to the diminishing marginal benefit of further human involvement. Since IML systems (e.g., interactive label correction, interactive feature selection, and interactive model tuning) share similar mechanisms governing human interaction and model improvement processes, the relationship between cost and benefit observed in our simulated interactive label correction process could potentially apply to other IML scenarios [83], [84], [85]. Future research could leverage simulation-based approaches to explore and compare the cost-benefit relationship across various IML tasks.

VIII. Conclusion and Future Work

In this paper, we examined the potential *upper-bound* impacts of interactive label correction on performance improvement by simulating a *perfect-agent* consecutively cleaning label noise in a given training data. We explored the relationship between the cost of label correction and the associated benefits of model classification performance gains on the test data. We implemented our simulation under different conditions created by varying levels of three environmental factors: noise ratio (α), noise type (T), machine learning model ($f()$). To generalize the findings, the simulation was conducted with two different types of datasets: one image dataset and one text dataset. We sampled several datasets from these two datasets, divided each sampled dataset into a training set and a testing set, diluted each training set based on the different setups of noise-related factors, applied distinct machine learning classifiers, and simulated interactive label correction with varying degrees of effort on label clean (β). Additionally, we confirmed simulation results using five real-world noisy datasets. We proposed two metrics, $D(\beta, 0)$ and $R(\beta, 0)$, and used their values to evaluate the model performance enhancement from the label correction. Although there

are minor differences between the analysis results of these distinct datasets and evaluation metrics, their commonality lies in that all environmental factors tested in the simulation influence the effect of interactive label correction on classification performance improvement. Furthermore, the impact of interactive label correction for different noise types and machine learning models varies across different noise ratios. Notably, in simulations for FashionMNIST and AGNews-10pct datasets, and the real-world label noise validations on CIFAR-10N datasets, we observed that as human effort increases (larger β), it has a more significant impact on enhancing the classification performance. However, the effect of cleaning label noise diminishes as more label errors are corrected, especially when the noise ratio is relatively large ($\alpha \geq 0.4$). This indicates there is a trade-off between the expense of label correction and the attained benefits of model classification performance. As mentioned above, all simulation factors and their active interactions, such as $\alpha \times T$ and $\alpha \times f()$, are influential to the effect of interactive label correction on classification performance improvement. The simulation results suggest that practitioners have to scrutinize how much effort is warranted to improve classification performance considering task complexity, including but not limited to the noise-related, dataset-related, algorithm-related factors, the ability of the human, as well as the performance measures of interactive machine learning system, and corresponding desired performance gains. Since the relationship between benefits and costs of interactive label correction vary under distinct conditions, our work provides a practical means to determine when to ask humans to intervene in interactive relabeling considering its cost-benefit trade-offs.

Limitations: While our simulation model provides a reasonable approach to exploring the possible bounds of improvement that can be expected from label revision, one major limitation of the work is that we did not account for all crucial factors in human-in-the-loop scenarios, e.g., human expertise and visual design. Additionally, in each trial, we only simulated interactive label correction on any one of two types of datasets (image or text) with one type of synthetic noise, whereas real-world datasets often contain mixed noise types. Furthermore, our simulation was initialized using balanced, clean data (i.e., the instances across diverse classes are evenly distributed) and applied uniform contamination to the training data for each class. This approach does not always align with real-world label noise scenarios. Data imbalance can significantly influence the ML model performance and exacerbate the adverse effects of label noise. Moreover, varying amounts and distributions of label noise (α, T) could result in distinct degrees of data imbalance, leading to different levels of deterioration on model performance [86], [87], [88], [89].

Our simulation studied a best-case scenario that assumes a *perfect-agent* is able to correct all inspected mislabels through the IML system while lacking the investigation on the expertise of an *imperfect-agent* human annotator, who might mistakenly and inconsistently flip instance labels, and the various difficulty levels of recognizing and correcting noisy instances for a human annotator. Besides, we did not incorporate advanced techniques (e.g., weak supervision), which may achieve strong performance with minimal precisely labeled data and could reach the optimal stopping point for label correction sooner than fundamental machine learning models tested in our simulations. Even in some cases, label correction may not be necessary at all if the system starts with sufficient clean data. Furthermore, we did not consider human strategies or apply active learning on instance selection for relabeling but only implemented random selection. In practice, a visual interface can help humans identify suspicious labels, further aiding humans in relabeling noisy instances.

We focused on the benefit of human interactive label correction in terms of classification performance improvement but did not fully explore the associated costs, such as human labor and consumed time, or the cost of different types of classification errors (e.g., false negative vs. true positive). Our work only used classification accuracy for quantifying the benefits with two class-balanced datasets, yet real-world datasets can occur with arbitrary class distributions, thus other metrics (e.g., precision, recall, F_1 score) may be more suitable for measuring performance improvements. Additionally, we only simulated one iteration of interactive label correction, while a practical interactive labeling system usually involves multiple incremental iterations for enhancing performance, and the relationship between benefits and costs of human interactive label revision may vary as the number of iterations increases. Finally, the simulation-based approach might not be applicable to massive datasets due to computational limitations.

Future Work: Future studies can extend the simulation methods and factors to relax *best-case scenario* assumptions, including exploring different relabeling accuracies of *imperfect-agent*, human perception when using visual interfaces for relabeling, and human strategies for label inspection and correction. This will allow the generalization of the cost-benefit relationship for more general and practical usage scenarios. Moreover, other algorithm-related factors (e.g., the baseline model performance, active learning approach for maximizing the expected model performance enhancement, and label propagation algorithms to reduce the expense of label correction) and advanced algorithms, like BERT for text classification and ResNet for image recognition, should be included in future simulations. We also intend to extend the interactive label correction simulation to more types of datasets with various benefit and cost measures, and include more label correction and model retraining cycles for generalizing the rule of cost and benefit trade-offs through a complete interactive label correction process. Additionally, modeling the cost and benefit of interactive label correction, using optimization on a parameterized model, and developing threshold regions for determining the optimal level of label correction for various parameter regions will also be our future work. Except for the interactive label correction process, the impact of human intervention in other IML approaches (e.g., feature selection) has not been explored yet, which is still an open question.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

REFERENCES

[1] S. Liu, C. Chen, Y. Lu, F. Ouyang, and B. Wang, "An interactive method to improve crowdsourced annotations," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 235–245, Jan. 2019.

[2] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5552–5560.

[3] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artif. Intell. Rev.*, vol. 33, no. 4, pp. 275–306, 2010.

[4] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 139–147.

[5] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2691–2699.

[6] L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *J. Vis.*, vol. 22, no. 2, pp. 401–417, Apr. 2019.

[7] O. M. Aodha, V. Stathopoulos, G. J. Brostow, M. Terry, M. Girolami, and K. E. Jones, "Putting the scientist in the loop - Accelerating scientific progress with interactive machine learning," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 9–17.

[8] D. Sacha et al., "Human-centered machine learning through interactive visualization: Review and open challenges," in *Proc. Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn.*, 2016, pp. 641–646.

[9] J. J. Dudley and P. O. Kristensson, "A review of user interface design for interactive machine learning," *ACM Trans. Interactive Intell. Syst.*, vol. 8, no. 2, pp. 1–37, Jun. 2018.

[10] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski, "The state-of-the-art in predictive visual analytics," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 539–562, Jun. 2017.

[11] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner, "VIAL: A unified process for visual interactive labeling," *Vis. Comput.*, vol. 34, no. 9, pp. 1189–1207, Sep. 2018.

[12] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing visual-interactive labeling with active learning: An experimental study," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 298–308, Jan. 2018.

[13] M. Chegini, J. Bernard, P. Berger, A. Sourin, K. Andrews, and T. Schreck, "Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning," *Vis. Inform.*, vol. 3, no. 1, pp. 9–17, Mar. 2019.

[14] M. Chegini et al., "Interactive visual labelling versus active learning: An experimental comparison," *Front. Inf. Technol. Electron. Eng.*, vol. 21, no. 4, pp. 524–535, 2020.

[15] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu, "Interactive correction of mislabeled training data," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2019, pp. 57–68.

[16] A. Bäuerle, H. Neumann, and T. Ropinski, "Classifier-guided visual correction of noisy labels for image classification tasks," *Comput. Graph. Forum*, vol. 39, no. 3, pp. 195–205, Jul. 2020.

[17] M. Khayat, M. Karimzadeh, J. Zhao, and D. S. Ebert, "VASSL: A visual analytics toolkit for social spambot labeling," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 874–883, Jan. 2020.

[18] L. S. Snyder, Y.-S. Lin, M. Karimzadeh, D. Goldwasser, and D. S. Ebert, "Interactive learning for identifying relevant tweets to support real-time situational awareness," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 558–568, Jan. 2020.

[19] Y. Zhang, B. Coecke, and M. Chen, "On the cost of interactions in interactive visual machine learning," in *Proc. IEEE VIS Workshop Eval. Interactive Vis. Mach. Learn. Syst.*, 2019, Art. no. 5.

[20] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Mag.*, vol. 35, no. 4, pp. 105–120, Dec. 2014.

[21] J. A. Fails and D. R. Olsen, "Interactive machine learning," in *Proc. Int. Conf. Intell. User Interfaces*, 2003, pp. 39–45.

[22] C. Bors, T. Gschwandtner, and S. Miksch, "Capturing and visualizing provenance from data wrangling," *IEEE Comput. Graph. Appl.*, vol. 39, no. 6, pp. 61–75, Nov./Dec. 2019.

[23] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2334–2346.

[24] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2011, pp. 3363–3372.

[25] J. Krause, A. Perer, and E. Bertini, "INFUSE: Interactive feature selection for predictive modeling of high dimensional data," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1614–1623, Dec. 2014.

[26] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions - A dual visual analysis model for high-dimensional data," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2591–2599, Dec. 2011.

[27] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Comput. Graph. Forum*, vol. 31, pp. 1335–1344, Jun. 2012.

[28] M. Nadj, M. Knaeble, M. X. Li, and A. Maedche, "Power to the oracle? Design principles for interactive labeling systems in machine learning," *Künstliche Intelligenz*, vol. 34, no. 2, pp. 131–142, Jun. 2020.

[29] X. Zhang, X. Zhu, and S. Wright, "Training set debugging using trusted items," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 549:1–549:8.

[30] J. Bernard, M. Zeppelzauer, M. Lehmann, M. Müller, and M. Sedlmair, "Towards user-centered active learning algorithms," *Comput. Graph. Forum*, vol. 37, no. 3, pp. 121–132, Jul. 2018.

[31] J. Bernard, M. Hutter, M. Sedlmair, M. Zeppelzauer, and T. Munzner, "A taxonomy of property measures to unify active learning and human-centered approaches to data labeling," *ACM Trans. Interactive Intell. Syst.*, vol. 11, no. 3–4, pp. 20:1–20:42, Sep. 2021.

[32] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2012, pp. 83–92.

[33] A. Endert, P. Fiaux, and C. North, "Semantic interaction for visual text analytics," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 473–482.

[34] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Appl. Soft Comput.*, vol. 52, pp. 109–119, Mar. 2017.

[35] Q. Song, G. Wang, and C. Wang, "Automatic recommendation of classification algorithms based on data set characteristics," *Pattern Recognit.*, vol. 45, no. 7, pp. 2672–2689, Jul. 2012.

[36] C. Chen and M.-L. Shyu, "Clustering-based binary-class classification for imbalanced data sets," in *Proc. IEEE Int. Conf. Inf. Reuse Integration*, 2011, pp. 384–389.

[37] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Appl. Soft Comput.*, vol. 6, no. 2, pp. 119–138, Jan. 2006.

[38] O. Kwon and J. M. Sim, "Effects of data set features on the performances of classification algorithms," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1847–1857, Apr. 2013.

[39] M. Y. Kiang, "A comparative assessment of classification methods," *Decis. Support Syst.*, vol. 35, no. 4, pp. 441–454, Jul. 2003.

[40] B. Nicholson, J. Zhang, V. S. Sheng, and Z. Wang, "Label noise correction methods," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics*, 2015, Art. no. 9.

[41] G. Zheng, A. H. Awadallah, and S. Dumais, "Meta label correction for noisy label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11 053–11 061.

[42] T. Kaneko, Y. Ushiku, and T. Harada, "Label-noise robust generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2467–2476.

[43] A. Drory, S. Avidan, and R. Giryes, "How do neural networks overcome label noise?," in *Proc. Workshop Track Int. Conf. Learn. Representations*, 2018, Art. no. 3.

[44] X. Zeng and T. R. Martinez, "An algorithm for correcting mislabeled data," *Intell. Data Anal.*, vol. 5, no. 6, pp. 491–502, Jan. 2001.

[45] X. Zeng and T. Martinez, "A noise filtering method using neural networks," in *Proc. IEEE Int. Workshop Soft Comput. Techn. Instrum. Meas. Related Appl.*, 2003, pp. 26–31.

[46] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *J. Intell. Inf. Syst.*, vol. 22, no. 1, pp. 89–109, 2004.

[47] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *J. Artif. Intell. Res.*, vol. 11, pp. 131–167, Aug. 1999.

[48] S. Verbaeten and A. Van Assche, "Ensemble methods for noise elimination in classification problems," in *Multiple Classifier Systems*. Berlin, Germany: Springer, Jun. 2003, pp. 317–325.

[49] X. Zhu, X. Wu, and Q. Chen, "Eliminating class noise in large datasets," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2003, pp. 920–927.

[50] S. Venkataraman, D. Metaxas, D. Fradkin, C. Kulikowski, and I. Muchnik, "Distinguishing mislabeled data from correctly labeled data in classifier design," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, 2004, pp. 668–672.
[51] U. Rebbapragada, C. E. Brodley, D. Sulla-Menashe, and M. A. Friedl, "Active label correction," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 1080–1085.
[52] J. Kremer, F. Sha, and C. Igel, "Robust active label correction," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2018, pp. 308–316.
[53] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw Hill, 1997.
[54] S. Lee, S.-H. Kim, and B. C. Kwon, "VLAT: Development of a visualization literacy assessment test," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 551–560, Jan. 2017.
[55] Y. Zhang, B. Coecke, and M. Chen, "MI3: Machine-initiated intelligent interaction for interactive classification and data reconstruction," *ACM Trans. Interactive Intell. Syst.*, vol. 11, no. 3–4, pp. 18:1–18:34, Dec. 2021.
[56] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?," in *Proc. Int. Conf. Intell. User Interfaces*, 2020, pp. 189–201.
[57] F. Sperrle et al., "A survey of human-centered evaluations in human-centered machine learning," *Comput. Graph. Forum*, vol. 40, no. 3, pp. 543–568, Jun. 2021.
[58] S. Mohseni, J. E. Block, and E. Ragan, "Quantitative evaluation of machine learning explanations: A human-grounded benchmark," in *Proc. Int. Conf. Intell. User Interfaces*, 2021, pp. 22–31.
[59] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-AI decision making: A survey of empirical studies," Dec. 2021, *arXiv:2112.11471*.
[60] S. K. Card, T. P. Moran, and A. Newell, "The keystroke-level model for user performance time with interactive systems," *Commun. ACM*, vol. 23, no. 7, pp. 396–410, Jul. 1980.
[61] B. E. John and D. E. Kieras, "Using GOMS for user interface design and evaluation: Which technique?," *ACM Trans. Comput.-Hum. Interaction*, vol. 3, no. 4, pp. 287–319, Dec. 1996.
[62] A. Ramkumar et al., "Using GOMS and NASA-TLX to evaluate human–computer interaction process in interactive segmentation," *Int. J. Hum.–Comput. Interaction*, vol. 33, no. 2, pp. 123–134, 2017.
[63] J. Zahálka, S. Rudinac, and M. Worring, "Analytic quality: Evaluation of performance and insight in multimedia collection analysis," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 231–240.
[64] Y. Zhang, M. Tennekes, T. de Jong, L. Curier, B. Coecke, and M. Chen, "Simulation-based optimization of user interfaces for quality-assuring machine learning model predictions," *ACM Trans. Interactive Intell. Syst.*, vol. 14, no. 1, pp. 1:1–1:32, Mar. 2024.
[65] G. Algan and I. Ulusoy, "Label noise types and their effects on deep learning," Mar. 2020, *arXiv: 2003.10471*.
[66] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," May 2017, *arXiv: 1705.10694*.
[67] Z. Zhang, H. Zhang, S. O. Arik, H. Lee, and T. Pfister, "Distilling effective supervision from severe label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9294–9303.
[68] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," Aug. 2017, *arXiv: 1708.07747*.
[69] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 649–657.
[70] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.
[71] Y. Xu, P. Cao, Y. Kong, and Y. Wang, *L_DMI: A Novel Information-Theoretic Loss Function for Training Deep Nets Robust to Label Noise*. Red Hook, NY, USA: Curran Associates, Dec. 2019, pp. 559:1–559:12.
[72] I. Jindal, D. Pressel, B. Lester, and M. Nokleby, "An effective label noise model for Dnn text classification," Mar. 2019, *arXiv: 1903.07507*.
[73] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[74] D. I. Inouye, P. Ravikumar, P. Das, and A. Dutta, "Hyperparameter selection under localized label noise via corrupt validation," in *Proc. Workshop Neural Inf. Process. Syst.*, 2017, Art. no. 6.
[75] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *Proc. Int. Conf. Res. Comput. Intell. Commun. Netw.*, 2018, pp. 122–129.
[76] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting LSTM networks for semi-supervised text classification via mixed objective function," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6940–6948.
[77] B. Liu, Y. Zhou, and W. Sun, "Character-level text classification via convolutional neural network and gated recurrent unit," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 8, pp. 1939–1949, Mar. 2020.
[78] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. Sang Choi, "Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model," *Comput. Intell.*, vol. 37, no. 1, pp. 409–434, Feb. 2021.
[79] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy? – A comprehensive study on the robustness of 18 deep image classification models," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Heidelberg, Springer, 2018, pp. 631–648.
[80] M. Chen and A. Golan, "What may visualization processes optimize?," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 12, pp. 2619–2632, Dec., 2016.
[81] M. Chen, "A short introduction to information-theoretic cost-benefit analysis," Mar. 2021, *arXiv:2103.15113*.
[82] X. Liang, X. Liu, and L. Yao, "Review–A survey of learning from noisy labels," *ECS Sensors Plus*, vol. 1, no. 2, pp. 021 401:1–021 401:8, Jun. 2022.
[83] R. Kong et al., "Getting the most from eye-tracking: User-interaction based reading region estimation dataset and models," in *Proc. Symp. Eye Tracking Res. Appl.*, 2023, pp. 10:1–10:7.
[84] Á. A. Cabrera et al., "Zeno: An interactive framework for behavioral evaluation of machine learning," in *Proc. ACM Conf. Hum. Factors Comput. Syst.*, 2023, pp. 419:1–419:14.
[85] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh, "Explaining machine learning models with interactive natural language conversations using TalkToModel," *Nat. Mach. Intell.*, vol. 5, no. 8, pp. 873–883, Aug. 2023.
[86] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data Knowl. Eng.*, vol. 68, no. 12, pp. 1513–1542, Dec. 2009.
[87] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, 1st ed. Berlin, Germany: Springer, Nov. 2018.
[88] M. Koziarski, M. Woźniak, and B. Krawczyk, "Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise," *Knowl.-Based Syst.*, vol. 204, pp. 106 223:1–106 223:16, Sep. 2020.
[89] R. K. Kennedy, J. M. Johnson, and T. M. Khoshgoftaar, "The effects of class label noise on highly-imbalanced big data," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, 2021, pp. 1427–1433.

Yixuan Wang (Graduate Student Member, IEEE) received the bachelor's degree in automation from Central South University, China, in 2017 and the master's degree in computer systems engineering from Northeastern University, MA, in 2020. She is currently working toward the PhD degree with the School of Computing and Augmented Intelligence, Arizona State University. Her research interests include visual analytics, machine learning, human-computer interaction, and human-AI teaming.

Jieqiong Zhao (Member, IEEE) received the MS degree in computer science from Tufts University in 2013 and the PhD degree in electrical and computer engineering from Purdue University in 2020. She is an assistant professor with the School of Computer and Cyber Sciences, Augusta University. She was a postdoctoral research associate with the School of Computing and Augmented Intelligence, Arizona State University. Her broad research interests include visual analytics, information visualization, human-computer interaction, and applied AI and machine learning.

Jiayi Hong (Member, IEEE) received the bachelor's degree from Zhejiang University, China, in 2018, the MS degree in computer science from the University of Bristol, U.K., in 2019, and the PhD degree from the Université Paris Saclay, Inria, France, in 2023. She is a postdoctoral fellow with Arizona State University in the School of Computing and Augmented Intelligence. Her research interests include interactive visualization techniques, visualizations for machine learning, and human-computer interaction.

Ronald G. Askin received the BS degree in industrial engineering from Lehigh University, the MS degree in operations research from the Georgia Institute of Technology, and the PhD degree in industrial and systems engineering from the Georgia Institute of Technology. He is an emeritus professor of Industrial Engineering with the School of Computing and Augmented Intelligence and executive director of the Center for Accelerating Operational Efficiency (CAOE), a DHS Center of Excellence, Arizona State University. He is a fellow of the Institute of Industrial and Systems Engineers (IISE), a fellow of the Institute for Operations Research and Management Sciences (INFORMS), and former editor-in-chief of IIE Transactions. He has received several awards for his textbooks and research, including a National Science Foundation (NSF) Presidential Young Investigator Award and the Shingo Prize for Excellence in Manufacturing Research. In 2017, he was honored with IISE's Albert G. Holzman Distinguished Educator Award. His primary research interests are the application of operations research to the design and operation of production systems.

Ross Maciejewski (Senior Member, IEEE) is a professor and director with the School of Computing and Augmented Intelligence at Arizona State University and Director of the Center for Accelerating Operational Efficiency (CAOE) – a Department of Homeland Security Center of Excellence. His primary research interests are in the areas of geographical visualization and visual analytics focusing on homeland security, public health, dietary analysis, social media, criminal incident reports, and the food-energy-water nexus. He has served on the organizing committees for the IEEE Conference on Visual Analytics Science and Technology and the IEEE/VGTC EuroVis Conference, and he currently serves as the co-chair of the Visualization Executive Committee (VEC).