IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Active Learning Based Federated Learning for Waste and Natural Disaster Image Classification

LULWA AHMED[1], KASHIF AHMAD[1], NAINA SAID[2], BASHEER QOLOMANY[3], JUNAID QADIR[4] (Senior Member, IEEE), ALA AL-FUQAHA[1] (Senior Member, IEEE)

[1]Information and Computing Technologies (ICT) Division, College of Science and Engineering (CSE), Hamad Bin Khalifa University, Doha, Qatar
[2]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan
[3]Department of Cyber Systems, College of Business & Technology, University of Nebraska at Kearney, Kearney, NE 68849
[4]Department of Electrical Engineering, Information Technology University, Lahore, Pakistan

Corresponding author: Kashif Ahmad (e-mail: kahmad@hbku.edu.qa).

The statements made herein are solely the responsibility of the authors. The publication of this article was funded by the Qatar National Library.

ABSTRACT

The feasibility of Federated Learning (FL) is highly dependent on the training and inference capabilities of local models, which are subject to the availability of meaningful and annotated data. The availability of such data is in turn contingent on the tedious and time-consuming annotation job that typically requires the manual analysis of training samples. Active Learning (AL) provides an alternative solution allowing a Machine Learning (ML) model to automatically choose and label the data from which it learns without involving manual inspection of each training sample. In this work, we explore how FL can benefit from unlabelled data available at each participating client using AL. To this aim, we propose an AL-based FL framework by employing and evaluating several AL methods in two different application domains. Through an extensive experimentation setup, we show that AL is equally useful in federated and centralized learning by achieving comparable results with manually labeled data using fewer samples without involving human annotators in collecting training data. We also demonstrated that the proposed method is dataset/application independent by evaluating the proposed method in two interesting applications, namely natural disaster analysis and waste classification, having different properties and challenges. Promising results are obtained on both applications resulting in comparable results against the best-case scenario where each sample is manually analyzed and annotated (Baseline 1), and improvement of 3.1% and 4% with best methods respectively over the training sets with irrelevant images on natural disaster and waste classification datasets (Baseline 2).

INDEX TERMS

Federated Learning; Deep Learning; Active Learning; CNNs; LSTM; Natural Disasters; Waste Classification

I. INTRODUCTION

Federated Learning (FL) is a machine learning (ML) technique that enables collaborative training of an ML model across multiple decentralized edge devices without sharing their data. In recent years, FL has been widely explored in different privacy-sensitive application domains. Apart from several other factors, the feasibility of FL in an application is also constrained by the availability of quality annotated data at each participating client to properly train local models. The process of data collection and annotation is one of the main bottlenecks especially in supervised ML where human annotators are generally used to annotate training data for an ML model [1]. To this aim, usually, a large population is involved in a crowd-sourcing activity to manually analyze and annotate data. The process involves two key challenges. Firstly, each sample is needed to be carefully analyzed, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

is a tedious and time-consuming job. Secondly, the process does not guarantee the selection of quality samples, which are more meaningful for a model, having an impact on the model's performance. Active Learning (AL), a learning strategy that allows a learning algorithm to interactively query an information source to pick and label new training samples, provides a potential solution to these challenges. On the one side, it allows ML algorithms to choose the data from which it learns, and eases the annotation process on the other hand by automatically labeling training samples from a large pool of unlabelled samples via a model trained on a very small manually annotated dataset.

The existing literature on FL assumes the availability of pre-defined fixed manually labeled training set at each client. However, in many cases, each client may have a large variety of unlabelled data, which could be utilized in training the local models resulting in an ultimate improvement in the performance of the global model.

In this work, we aim to explore how unlabelled data at a client could be exploited in an FL environment by proposing a novel AL-based FL framework to utilize the unlabelled data available at each client in building a global model collaboratively without sharing data in a multi-stake environment. To this aim, we employ and evaluate multiple AL methods with different sampling and disagreement strategies. More specifically, two pool based methods, namely (i) uncertainty sampling and (ii) query by committee, are analyzed with three different sampling and disagreement strategies, respectively.

In the current implementation, as a first step to avoid complexities in terms of communication and biases of learners at each client, we keep the AL task offline where the communication with the server starts once the data at each client is annotated. On the one hand, the offline AL reduces the communication rounds in FL by avoiding sample selection at each communication round. On the other hand, choosing and annotating training samples during FL (i.e., at each communication rounds) may introduce complexities in the convergence of the global model due to biases of the samples chosen at each communication round as the global model will be used as a learner in that case. In such a case, there would be two main challenges. Firstly, if we keep the number of samples to be picked at each iteration very high, there will be higher variations in the performance of the global model, and secondly, keeping it low will increase the number of communication rounds. Besides these challenges, another concern related to the so-called online active learning-based federated learning is that it would require a higher number of manually annotated initial training sets (i.e., seed) as we would need to split it among all the clients to train the learner.

The proposed method is evaluated in two interesting applications—namely, (i) natural disaster analysis in social media images, and (ii) waste classification—in which there is little annotated data but an abundance of unlabelled data. Besides the novelty in the methodology, the work also explores a different aspect of the applications compared to the existing literature on the applications including our previous contributions (i.e., natural disaster analysis [2]–[4] and waste classification [5]) as detailed in Sections II, and it is expected to provide a baseline for future work in the domain.

The main contributions of the work can be summarized as:

- We explore the possibility of automatically labeling training samples in FL via a novel framework for building a global model by utilizing the unlabeled data available at each local device.
- We evaluate the performance of two pool based AL methods with six different sampling and disagreement strategies in both federated and centralized learning, where a model is trained by uploading data from all participating parties to a server on the cloud, in two different applications.
- We show that AL is equally beneficial in both federated and centralized learning by achieving comparable results without involving manual annotation.
- We also show that the performance of FL could be significantly affected in the case of the unavailability of sufficient training samples at each client to train the local models, however, AL could be useful in such cases to obtain relevant samples without involving human annotators.

The rest of the paper is organized as follows. Section II describes the existing literature on AL, FL and both applications. Section III provides a detailed description of the proposed methodology. Section IV provides the details of the datasets, experimental setup, conducted experiments and results. Section V lists the lessons learned from the experiments and finally Section VI provides some concluding remarks and future research directions.

II. RELATED WORK

In this section, we provide a survey of the existing literature on the different aspects of the work including AL, FL, and the two applications, namely (i) natural disaster analysis, and (ii) waste classification, used for the evaluation of the proposed method.

A. ACTIVE LEARNING

In literature, AL has been widely exploited for image, text, videos, and multimedia retrieval in different application domains [3], [6]–[9]. For instance, in our previous work [10], an AL learning-based technique has been employed for social events recognition in personal photo-collections where an SVMs classifier is used as a learner to identify and annotate relevant pictures in photo collections. More recently, we also analyze the effectiveness of pool-based AL methods in the classification of disaster-related images [11]. Sun et al. [12] utilized AL for context-aware image annotation by exploiting the associated additional information available in the form of meta-data. In detail, four different features namely geo-location information, time stamps, users' tags, and camera tags are used in clustering to categorize images into different

labeled groups. In [13], AL has been employed in person re-identification through an AL framework namely early AL, which annotates pairs of images instead of an instance. As the name suggests, the framework is applied at the early stages of the experiments when no pre-labeled samples (i.e., reference point) are available for human annotation. Ngo et al. [14] proposed an AL scheme for content-based image retrieval where a ranking function exploiting SVMs scores along with another similarity measurement between the queered image and the images in a database. Yuan et al. [15] on the other hand propose a multi-criteria AL scheme to automatically annotate samples for training their CNN architecture for image classification.

AL has also been proved very effective in other challenging applications, such as the classification of hyperspectral images, where usually a limited number of training samples are available to train an ML model. For instance, Cao et al. [16] proposed a CNN-based AL framework for the classification of hyperspectral images by firstly training a CNN model on a smaller collection of annotated pixels, which is then used to annotate/pick potential pixels from the unlabelled pool. In [17], AL techniques are employed in a fusion framework combining spatial and spectral information for the classification of hyperspectral images. The AL scheme is mainly used to acquire the most relevant training samples for the framework.

The proven performance of AL techniques in such relevant and challenging applications provides a basis for our proposed solution, and evaluation in the two applications as detailed in Section IV-A.

B. FEDERATED LEARNING

Existing literature on FL mainly focuses on the challenges associated with the optimization of a global model with non-IID, unbalanced, and highly distributed data, and focuses on ensuring privacy and communication efficiency [18]–[21]. For instance, McMahan et al. [22] proposed a global optimization technique namely FedAvg to deal with the unbalanced and non-IID nature of data in an FL environment, where parameters of locally trained models are combined efficiently. To reduce the communication rounds, the framework selects a fraction of clients in each iteration instead of all participants. One of the main limitations of FedAvg is its inefficacy in dealing with heterogeneous data. To cope with heterogeneous data sources in an FL environment, Li et al. [23] proposed a modified version of FedAvg namely FedProx guaranteeing convergence in heterogeneous networks. To this aim, a proximal term has been added to the objective function of the model to deal with the heterogeneity associated with partial information. Smith et al. [18] proposed a multi-tasking based learning framework namely MOCHA to analyze how multi-tasking can cope with statistical challenges associated with FL. In contrast to the state-of-the-art solutions, instead of a single global model, multiple global models are trained one for each node.

A large portion of the literature also aims at the protection of model updates. In FL, data privacy has been categorized as global and local privacy [24]. The former aims at ensuring the privacy of the global model's parameters while the latter ensures that the local parameters are kept private. In [25], a Secure Multiparty Computation (SMC) protocol is developed to secure local model updates from the server where the server can only aggregate the local models. Differential privacy techniques have also been employed for the protection of model's updates in FL [26].

In the literature, FL has been deployed in several applications, such as sentiment analysis, monitoring, and tracking activities of mobile users, different tasks of autonomous vehicles, and healthcare [27], [28], where data is distributed at multiple devices.

To the best of our knowledge, the literature still lacks in solutions for utilizing unlabelled data available to clients. We believe this is one of the interesting directions to be explored, which may ultimately improve the performance of the global model.

C. NATURAL DISASTER ANALYSIS

Natural disaster analysis in images from different social media platforms is one of the interesting key applications that recently got the attention of the multimedia and signal processing community [3], [29]. Over the past few years, several interesting solutions covering different aspects of natural disaster analysis have been proposed. In [3], we provided a detailed survey of different solutions proposed for disaster analysis in images, videos, text, and remotely sensed data. In our previous work [4], we proposed a tool namely "Jord" to crawl, analyze, and filter disasters-related information obtained from social media and satellites. Similarly, Johnson et al. [30] analyzed Twitter data for the detection and classification of hurricane-related images. In [31], disaster-related images from social media are analyzed for damage estimation.

Disaster analysis in images has also been part of the benchmark competition namely MediaEval for three consecutive years since 2017, where a different aspect of natural disasters has been analyzed each year [32], [33]. In MediaEval-2017, 2018, and 2019, the competition focused on retrieval of flood-related content from social media, route passability analysis in a flooded region, and multi-modal flood-level estimation in news, respectively [32], [33]. The majority of the proposed solutions for these tasks rely on existing pre-trained models, such as AlexNet, GoogleNet, VggNet, and ResNet, which are either fine-tuned on the task-specific smaller datasets or used as feature descriptors [3]. For instance, in [34], existing pre-trained CNNs models are used for retrieval of disaster-related images. Similarly, in [35] multiple deep models pre-trained on Imagenet are used for the classification of flooded and non-flooded routs in social media images. We also contributed to the MediaEval challenge in our previous works [2], [36]. In [36], a CNN and Generative Adversarial Networks (GANs) based solution has been proposed for the detection of food-related events

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

in social media and satellite imagery. In [2], we proposed a deep architecture based framework for identification passable routs after floods in both social media and satellite imagery.

The existing work mostly focuses on a single type of natural disaster events. For instance, all the three tasks proposed in the MediaEval challenge are based on food events, only. To the best of our knowledge, the domain lacks in a large-scale benchmark dataset covering several types of disaster events, and this is one of the main motivations in the selection of the application for the proposed framework [3]. The unavailability of annotated and the abundance of unlabelled data available on different social media platforms make it a better choice for the evaluation of the proposed work. On one side, the proposed framework allows us to overcome the unavailability of training samples issues. On the other side, it will enable collaborative learning in a multi-party environment without sharing their data, leading to improved data privacy.

D. WASTE CLASSIFICATION

Waste classification is another interesting smart city application that has been widely explored in the literature. More recently, some interesting image-based solutions have also been introduced for waste classification [5], [37], [38]. For instance, Adede et al. [38] fine-tuned a pre-trained model namely ResNet on waste materials images. Vo et al. [39] also employed a pre-trained model namely ResNext for classification of waste into organic, inorganic, and medical waste. In [40], a detailed comparison of deep learning and traditional methods have been provided. In [41], a CNNs based framework namely "compostNet" is proposed for image-based classification of meal waste. Chu et al. [42] proposed a multilayer hybrid deep learning-based solution for waste materials classification and recycling. On the other hand, in our previous work [5], we employed several fusion methods for improved waste classification, where the fusion schemes are used to combine the capabilities of the different deep models in both early and late fusion.

The recent work in the literature shows the interest of the computer vision and ML community in the application. However, one of the key challenges in the domain is the unavailability of large-scale benchmark datasets. We believe a large collection of waste-related images could be easily obtained and the proposed framework could help in dealing with the data annotation, without involving manual annotation, as well as data privacy if multiple parties are involved in the learning process.

III. METHODOLOGY

There are three main components of the framework, namely (i) feature extraction, (ii) AL, and (iii) FL as illustrated in Figure 1, which provides the block diagram of the proposed methodology for our AL-based FL framework. The process starts with feature extraction from input images via an existing pre-trained deep model namely ResNet [43] (Section III-A). Subsequently in the AL phase, a classifier is trained on a smaller annotated training set also known as a seed. The classifier is also known as the learner is then used to annotate and pick unlabelled samples from a large-scale unlabelled pool of images via different sampling and dis-agreement strategies iteratively (Section III-B). The AL process continues until a sufficient number of training samples are obtained from the unlabelled pool of images. Finally, the training samples acquired through AL are used to train local ML models at participating clients, which are then aggregated to form a global model (Section III-C).

A. FEATURE EXTRACTION

Since the main focus of the work is to analyze how FL can benefit from unlabelled data available to each participating client using AL, thus, for feature extraction, we adopted rather a standard method without digging deeper in this aspect of the work. To this aim, we employed ResNet pre-trained on a large-scale ImageNet dataset [44] to extract object-level features from the input images. Our choice of the deep model for feature extraction is motivated by some recent works on the both applications [5], [34], [38], [45]. It is important to mention that the feature extraction part is independent of the AL and FL parts so it is expected that the choice of the model used for feature extraction will not have much impact on the overall analysis and insights of the AL-based FL. We used the ResNet configuration with 101 layers where features are extracted from the last fully connected layer without any fine-tuning and retraining.

B. ACQUISITION OF TRAINING SAMPLES VIA AL

The basic goal of the AL phase is to acquire and label training samples from the unlabelled pool of images available at local devices. To this aim, we relied on pool-based AL methods where a model, for example, "θ", is used to pick and annotate training samples from a pool of unlabelled samples namely $p = \{x_j\}_{j=1}^{n}$. To this aim, the model "θ" is initially built on a smaller manually annotated set namely "Seed". We mainly utilized two pool-based methods, namely (i) uncertainty sampling and (ii) query by committee. Uncertainty sampling methods allow a learner/model to judge the usefulness of a sample to be picked from a pool based on uncertainty (i.e., how much uncertain the learn is in assigning a label to the sample). On the other hand, query by committee relies on several hypotheses/learners in the selection of a sample, and the decision is made based on disagreement among the learners.

Both methods are evaluated under several sampling and disagreement strategies. The basic motivation for the evaluation of the methods under different sampling and disagreement strategies is to provide a detailed comparative analysis of the available strategies, which are expected to provide a base-line for future work in the domain. For the uncertainty sampling method three sampling strategies namely Least confidence, Margin Sampling, and Entropy Sampling. On the other hand, query by committee method is evaluated under three disagreement strategies, namely Vote Entropy,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

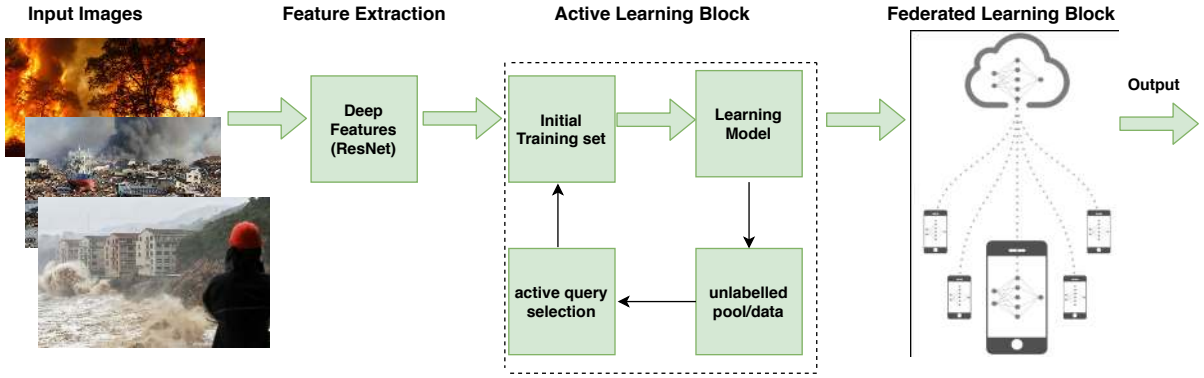


FIGURE 1: Block diagram of the proposed AL-based FL framework. The framework is mainly composed of two blocks, namely AL and FL where the output of the AL block is provided as input to the FL block.

Consensus Entropy, and Max Disagreement. The sampling and disagreement methods are described below.

- *Least Confidence*: This strategy picks the sample for which the learner/model is least confident, and can be computed using Eq. 1, where s represents the sample to be chosen, and $y^{'}$ is the most probable label.

$$U_{lc}(S) = \arg\max_{s} 1 - p_\theta(y^{'}|s) \tag{1}$$

- *Margin Sampling*: It aims to pick the sample/instance having the least difference between the probabilities of the two most probable classes as shown in Eq. 2. Here s is the sample to be predicted and $y_1^{'}, y_2^{'}$ are the two most probable labels.

$$U_{ms}(S) = p_\theta(y_1^{'}|s) - p_\theta(y_2^{'}|s) \tag{2}$$

- *Entropy Sampling*: The strategy selects the sample with the highest entropy as calculated by Eq. 3 where $P(y|x)$, U_{ES} and Y represent posterior probability, uncertainty measure and output, respectively.

$$U_{es}(x) = -\sum_{y \epsilon Y} P_\theta(y|x) \log_2 P_\theta(y|x) \tag{3}$$

- *Vote entropy*: It is a query by committee generalization of uncertainty sampling with entropy sampling relying on the distribution of the votes in sample selection, and can be computed using Eq. 4. Here y_i represents all possible labels, C represents the number of committee learners/classifiers while $V(y_i)$ shows the number of learners/classifier predicting label y_i.

$$QC_{ve}(S) = \arg\max_{s} -\sum_{i} \frac{V(y_i)}{C} \log \frac{V(y_i)}{C} \tag{4}$$

- *Consensus entropy*: Instead of vote distribution, this method firstly computes the consensus of learners/classifiers by averaging their class probabilities. The entropy of the consensus probability is then computed using Eq. 5, and an instance with the highest consensus entropy is selected.

$$QC_{ce}(S) = \frac{1}{C} \sum_{c=1}^{C} P_\theta(y_i) \tag{5}$$

- *Maximum disagreement*: The method computes each learner/classifier's disagreement with the consensus probabilities and chooses the sample having maximum disagreement for a learner.

In the AL part, as a first step, a learner is trained on the seed (a small manually labeled) which is then used to predict labels for the samples in an unlabelled pool of images and add them to the seed under a criterion defined in the underlying sampling and disagreement strategy. It is important to mention that it is an iterative process where the most relevant sample is fetched from the pool at each iteration. The process will keep fetching samples from the pool until a stopping criterion is met. The max number of iterations is a key parameter to be chosen as after certain iterations the relevancy of the selected sample will start decreasing and AL will force to add irrelevant samples in the training set at a certain point. To this aim different strategies could be used to fix the number of iterations. One of the possible solutions is to stop the process when the accuracy of the model reaches a stable point. Our stopping criteria are based on the max number of iterations, which is represented as "N" in the following Algorithm 1.

C. BUILDING THE GLOBAL MODEL IN FL ENVIRONMENT

The final component of our framework is based on an FL architecture, inspired by Federated Averaging (FedAvg) algorithm [22], to build a global model by combining the stochastic gradient descent (SGD) of the local models. Figure 2 describes the basic architecture of the FL algorithm, where parameters 'θ_t" of the global model are shared by the server among the participating clients, which in response train their local models on their data. After successful training, the parameters of the local models (e.g., "θ_t^k of the nth client) are shared with the server to update the parameters of the global

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

Algorithm 1 Acquisition/annotation of training samples via AL at each client

Require: Input images, learner (i.e., a classification algorithm) and sampling or dis-agreement strategies for query selection.

Ensure: Labels for samples from unlabelled pool of images.

Step 1: Division of data into seed (an initial small training set) and unlabelled pool of images.

for i=1:N **do**

Step 2: Train learner/model on seed.

Step 3: predicting labels for the samples in the unlabelled pool of samples and selection/queering samples to be added in seed.

end for

model (i.e., "$\theta_{(t+1)}$"), which repeats the process by sharing the updated model's parameters with the clients again. The process continues for a certain number of communication rounds.

In our case, a total of five clients participate in FL, while a Recurrent Neural Network (RNN) namely Long short-term memory (LSTM) has been used as the learning algorithm. It is important to mention that LSTM is trained on the deep features extracted with a pre-trained model namely ResNet as described earlier. The choice of LSTM is motivated by the fact that re-training a deep CNNs each time to update the model in FL requires heavy computation resources at the edge devices. Thus, the hybrid CNN-LSTM model will help to combine the capabilities of CNNs and RNNs for better image classification as reported in [46], [47]. Details of the LSTM model in terms of the number of layers, number of neurons in a layer, and other parameters are provided in Section IV-B.

For building the global model via aggregation of the local models' parameters, we adopted the optimization algorithms namely FedAvg [22]. Algorithm 2 provides the pseudo-code of FedAvg algorithm. The algorithm is divided into two parts. The first part shows the operations on the server-side while the second part depicts the operations made by each client. Here θ_t represents the parameters of the global model while θ^k K, B, E, η, n_k, and n show the kth local model's parameter, the total number of clients, mini Batch size, the total number of training iterations, learning rate, data size at client k, and the size of the whole data, respectively.

Moreover, the details of the FL parameters, such as the number of communication rounds, the total number of clients, and the number of clients contacted per iteration, are provided in Section IV-B.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

In this section, we provide the details of the datasets used in both applications are provided below.

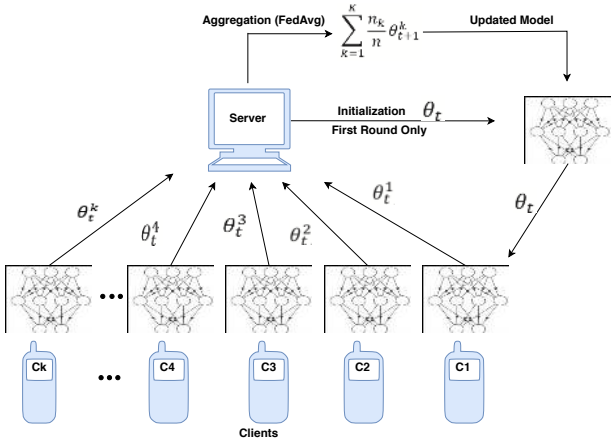


FIGURE 2: Block diagram of FL architecture [48]. The process starts with the initialization of the global parameters by the servers, which are then shared with k clients. The clients then update the parameters via training on local data and then send it back to the server, which repeats the process. Here "θ_t represents the parameters of the initial global model while K, n, n_k, θ_t^k, and $\theta_{(t+1)}$ represent the total number of clients, size of the whole data, the data size of the kth client, the parameters of the local model trained by the kth client, and the updated parameters of the global model at time $t+1$, respectively.

Algorithm 2 Building a federated model via FedAvg algorithm with a total number of K clients [22], [48].

Require: K, B, E, n_k, n, η, and n .

Ensure: θ^t global model's parameters.

Operations on the server side:

for each communication round t= 1, 2, 3 ... **do**

(i) Select a fraction of clients $m = C\text{x}K$ where $C \in (0, 1)$

(ii) Download θ_t to each client k

for each client $k \in$m **do**

(i) Wait Client k for synchronization

(ii) Compute $\theta_t = \sum_{k=1}^{m} \frac{n_k}{n} \theta^k$

end for

end for

Operations on the clients' side (suppose client at K):

$\theta^k = \theta_t$

for each iteration 1 to E **do**

for batch $b \in B$ **do**

$\theta^k = \theta^k - \eta \bigtriangledown L_k(\theta^k, b)$

end for

end for

return θ

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

1) Natural Disaster Analysis

For this application, we crawled images from social media platforms. In total, the dataset is composed of more than 7,000 natural disaster-related images from eight different disasters, namely cyclone, drought, earthquake, floods, landslide, thunderstorms, snowstorms, and wildfires. Figure 3 depicts some sample images from the dataset. We divided the dataset into training and test sets. The test set is composed of 2,540 images while the training set contains more than 5,000 images. The training set is further divided into a smaller manually annotated dataset also known as "Seed" and a larger pool of unlabelled images. Further details of the subsets of the training set are provided in Section IV-B.

2) Waste Classification Dataset

For waste classification, we used a benchmark dataset provided in [49]. The dataset is composed of a total of 2,527 images from six different waste categories, namely cardboard, glass, metal, paper, plastic, and trash. Figure 4 provides sample images from the dataset. To cover different challenges, such as rotation and illumination issues, the images are taken at different angles under different lighting conditions. Similarly to natural disaster analysis applications, the training set is further divided into "seed" and pool of unlabelled images where the labels of the images are ignored. Besides, irrelevant images are added to the unlabelled pool of images to challenge the learner in picking relevant samples for the training purposes. More details are provided in the next subsection.

B. EXPERIMENTAL SETUP

To show the effectiveness of the proposed AL-based FL framework, we conducted several experiments. On one side, we evaluate and compare the performances of the AL methods in the FL environment against two baselines, namely (i) *Baseline 1* (i.e., manually annotated training set) and (ii) *Baseline 2* (i.e., the one containing impurities, which we termed as a loosely labeled set). Since the work aims to evaluate the benefits of active learning in a federated learning environment thus, we believe, the two baselines seem more feasible options for comparisons instead of SoA in both domains. The first baseline shows the best-case scenario, where manually annotated training data is available, while the second scenario represents the worst case where a model is trained on a dataset containing a reasonable amount of irrelevant samples. To this aim, the waste classification dataset is synthesized by adding up-to 35 to 40% irrelevant images in the unlabelled pool of images. On the other hand, the natural disaster analysis application represents a more practical scenario where the second baseline is trained on a collection of images from social media with the corresponding tags/queries without manual inspection and removal of irrelevant images. However, for the manually annotated baseline, all the images are manually analyzed and annotated via crowd-sourcing.

TABLE 1: Salient parameters used during experimentation

Parameters	Values
Max. AL iterations in the Natural Disaster Use Case	2000
Max. AL iterations in the Waste Classification Use Case	1500
Total clients	5
Max. communication rounds	50
Number of clients contacted per round	5
Number of epochs	10
Batch size	10
Total number of LSTM layers	2
Number of neurons in the first layer of LSTM	100
Number of neurons in the second layer of LSTM	20

We also aim to show how the performance of AL methods vary when deployed in federated and centralized learning. In addition, we analyze how the performance of a global model is affected by clients with fewer samples. The experiment setup is kept unchanged for both AL and FL throughout the experiments. In the next subsections, we provide the details of the experimental setup specific to AL and FL. In Table 1, we summarize the parameter values used during the experimentation process.

1) Active Learning

The most important parameters to be set in the AL part are the number of images/samples in "Seed" (i.e., the initial training set for the learner) and the maximum number of iterations defining the stopping criteria for AL. In practice, the size of the seed depends on the availability of the manually annotated training samples for an application. Since the learner is trained on the seed so the number and quality of samples in the seed are very crucial for the performance of the learner [50]. However, acquiring more annotated samples for the initial training set requires human labor, thus, it shows a trade-off between the labor required for annotation and the performance, which is one of the main themes of AL. For a successful AL method, it is very crucial to obtain better results with a smaller seed. In our experiments, we started with a total of 160 and 120 samples (20 images from each class) in the seed for natural disaster and waste classification applications, respectively, which are then iteratively increased through the query selection schemes by adding the most relevant image at each iteration. Moreover, we used a total of 2000 and 1500 iterations for natural disaster analysis and waste classification, respectively. It is important to mention that the test and seed sets are manually annotated.

2) Federated Learning

Similar to the AL part, a fixed experimental setup has been used for FL throughout the experimentation. For instance, the dataset is divided into six parts where five of them cover the training set and are distributed among five clients in such a way that each client gets sufficient samples from each of the classes. The sixth part is composed of the test images only. It is important to mention that the division of the training set into clients is based on the fewer samples in the dataset. The LSTM, which is used as the learning model, is composed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

FIGURE 3: Sample images from the natural disaster dataset. The dataset is composed of eight different classes.

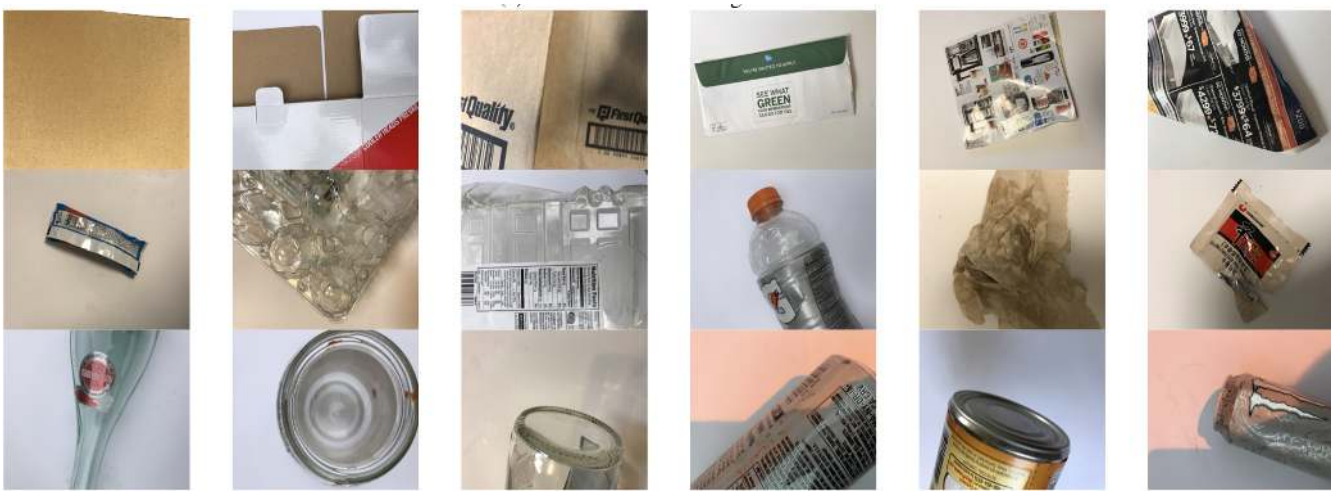


FIGURE 4: Sample images from the waste classification dataset. The dataset is composed of six different classes.

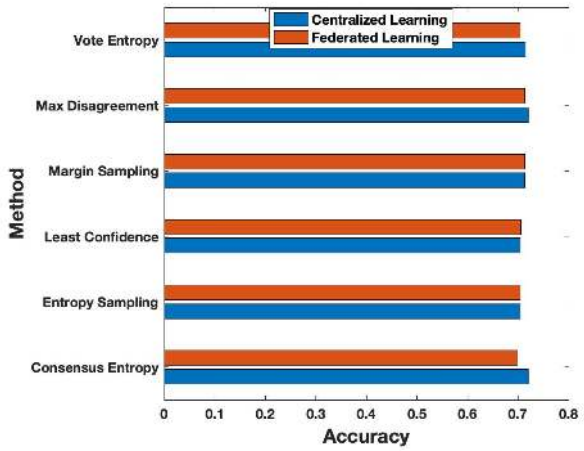
of two layers (containing 100 and 20 neurons, respectively), dropout, and a classification layer. The dropout layer, which randomly removes certain features by setting them to zero, is used to deal with the data over-fitting issue. Some other key parameters of the FL framework include the number of communication rounds and clients contacted per round, which is set with values of 50 and 5, respectively. Moreover, to analyze the impact of variation in the number of clients on the performance of the framework, we also experimented with various number of clients as detailed in Section IV-C.

C. EXPERIMENTAL RESULTS

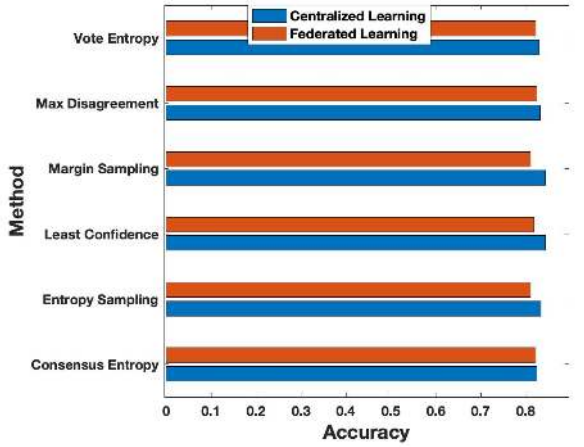
1) Natural Disaster Analysis in Social Media Images

Figure 5a provides the experimental results of the proposed framework under different sampling and disagreement strategies in both centralized and federated learning in terms of accuracy on the natural disaster analysis dataset. As can be seen, no significant differences have been observed in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access



(a) Natural Disaster Dataset.



(b) Waste Classification Dataset.

FIGURE 5: Comparison of the AL methods in Federated Learning (#clients = 5) and Centralized Learning environments under different sampling and disagreement strategies.

performance of the AL methods under different sampling and disagreement strategies after the maximum number of iterations. However, during experiments, we observed significant variations in the performances of the methods under these sampling and disagreement strategies at the initial 500 iterations. One of the possible reasons is that the performance stabilizes after a certain number of iterations for all of the methods. Generally query by committee methods are observed to achieve the highest accuracy a bit sooner (i.e., with fewer samples) compared to uncertainty method. Since one of the key motivations of the AL method is to obtain higher or comparable results with fewer samples than manually annotated datasets, thus on this basis, we can say query by committee method performed better than uncertainty method in terms of obtaining the maximum accuracy with fewer samples. As far as the performance of the AL methods in centralized and federated learning environments is concerned, interestingly the performance of the methods is comparable in most of the cases except query by committee with consensus entropy-based disagreement scheme where the performance is slightly reduced in FL compared to centralized learning.

We also evaluate the methods in terms of other metrics, namely precision, recall, and F-measure, which will help to evaluate the methods in a fair way by considering the imbalance classes of the dataset in generally and after deploying the AL methods in particular, where a higher number of samples may be obtained from the unlabelled pool for certain classes compared to the others. As can be seen in Table 2a, a mostly similar trend has been observed in the results also in terms of weighted precision, recall, and F-measure. In order to better describe the variation in the performance of the methods, we also provide the standard deviation of the variations in the performance of methods in FL and CL setups. The lower values of the standard deviation suggest a lower impact on the performance in FL with an added privacy.

2) Waste Classification

Figure 5b provides experimental results of the proposed framework on the waste classification dataset. One of the main objectives of the experiments on the waste classification dataset, which is slightly smaller in the number of images, is to analyze the impact of fewer samples on the performance in centralized and federated learning as AL results in a further reduction in the number of training samples. Thus, it is important to analyze the feasibility of the proposed framework on a smaller dataset. Similar to natural disaster analysis, comparable results are obtained with AL methods using fewer training samples without involving manual annotation in both federated and centralized learning. Similarly, there's no clear winner among the AL methods under different sampling and disagreement strategies, however, query by committee methods obtain the highest accuracy with fewer samples compared to the uncertainty methods.

In contrast to the natural disaster analysis use-case, in waste classification, a slight reduction can be observed in the performance of the AL methods when deployed in centralized and federated environments. One possible reason could be the lower number of samples in the waste classification dataset overall as well as in certain classes. For instance, trash, cardboard, and metal classes have fewer samples. In the FL the dataset is further divided into five subsets each allocated to a client to train a local model where the performance of the global model may be affected due to insufficient training for local models. In Table 2b, we provide the experimental results in terms of weighted precision, recall, and F-measure.

3) Trade-off analysis between accuracy and number of clients

Figure 6 provides the results of our second experiment, where we analyze the trade-off between the number of clients and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access

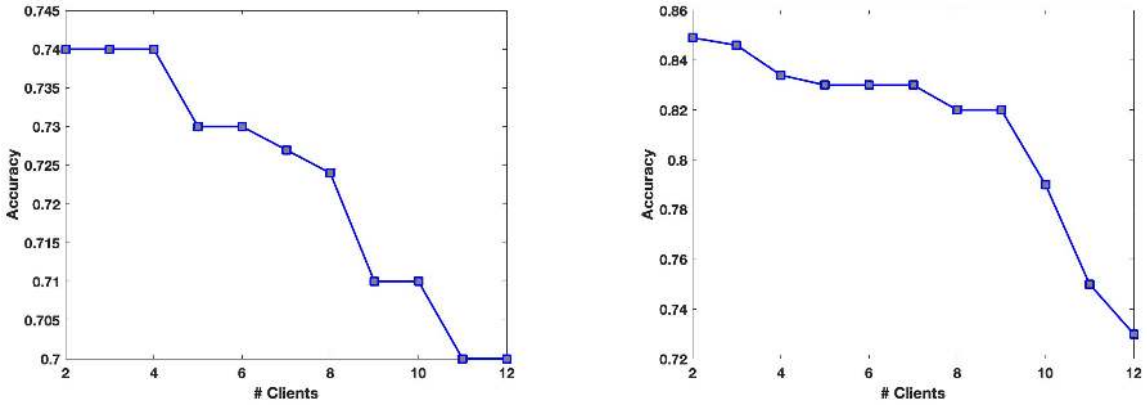
TABLE 2: Comparison of the AL methods in Federated Learning (#clients = 5) and Centralized Learning environments under different sampling and disagreement strategies in terms of weighted precision, recall, and F1-score.

Method	**Federated Learning (# clients = 5)**			**Centralized Learning**			**Standard Deviation (FL-CL)**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Least Confidence	0.722	0.706	0.714	0.723	0.704	0.713	0.0007	0.001	0.0007
Margin Sampling	0.728	0.714	0.720	0.735	0.714	0.724	0.004	0	0.002
Entropy Sampling	0.721	0.704	0.713	0.725	0.705	0.715	0	0	0
Vote Entropy	0.720	0.705	0.712	0.735	0.715	0.725	0.007	0.007	0.007
Consensus Entropy	0.714	0.695	0.706	0.741	0.720	0.730	0.021	0.021	0.021
Max Disagreement	0.726	0.713	0.717	0.741	0.720	0.730	0.014	0.007	0.014

(a) **Natural Disaster Image Dataset**

Method	**Federated Learning (# clients = 5)**			**Centralized Learning**			**Standard Deviation (FL-CL)**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Least Confidence	0.821	0.818	0.820	0.837	0.836	0.837	0.011	0.012	0.012
Margin Sampling	0.815	0.811	0.813	0.845	0.844	0.844	0.021	0.023	0.0219
Entropy Sampling	0.823	0.818	0.820	0.836	0.834	0.835	0.009	0.011	0.010
Vote Entropy	0.822	0.823	0.823	0.832	0.830	0.831	0.007	0.004	0.005
Consensus Entropy	0.826	0.823	0.824	0.825	0.825	0.825	0.0007	0.001	0.0007
Max Disagreement	0.826	0.839	0.832	0.835	0.834	0.835	0.006	0.004	0.002

(b) **Waste Classification Dataset**



(a) Natural Disaster Dataset

(b) Waste Classification Dataset

FIGURE 6: Trade-off between the number of clients and accuracy in FL. *The accuracy of the global model drops as we reduce the training samples per client by distributing the available training set among additional clients.*

the accuracy of the global model in FL. The main motivation of the experiment is to analyze how increasing the number of clients affect the performance of the global model. To this aim, we experimented with the manually annotated training set. We started with two clients, where the total available training samples are distributed between two clients, and increased the number of clients, iteratively. As depicted in the figure, a significant reduction has been observed in the performance on both datasets. At the initial stages with fewer clients, the accuracy is stable on both datasets, however, it is reduced more rapidly when the number of clients increases resulting in a significant reduction in the number of training samples at each client. Compared to natural disaster analysis use-case, the reduction in the accuracy of waste classification is slightly on a higher side due to the fewer training samples in the dataset. This experiment provides a basis for our third experiment, where we analyze the impact on the accuracy by extending the training set at each client via AL.

4) Trade-off analysis between accuracy and number of training samples

Figure 7 represents the results of our final experiment where we analyze the impact on the accuracy of the global model by extending the training set of each client in FL by keeping the number of clients constant. To this aim, we experimented on the manually annotated training set from both datasets. We started with a total of 1000 samples from each dataset, which were distributed among the five clients ensuring a sufficient

number of training samples per client. We kept extending the training set of each client by adding 50 samples per client manually, making an increase of 250 samples in the total number of training samples. As can be seen in Figure 7, the performance of the model is improved each time we increased the number of training samples. The variation is higher at the initial stages, and the rate of increase in the performances decreases as the number of training samples increases per client.

5) Comparison against the baselines

In order to show the effectiveness of the AL methods, we also compare the results of the AL methods against the two baselines in both FL and CL environments in terms of accuracy and F-score in Table 3a and Table 3b. On natural disaster images, in both FL and CL, comparable results have been achieved by the model trained on training samples annotated with the AL methods and manually annotated data. Moreover, in both cases, all the AL methods obtained better results compared to the model when trained on a loosely labeled training set, which shows the effectiveness of the methods. A similar trend has been observed in the waste classification where the AL methods provide comparable results with the baseline 1 (i.e., manually annotated dataset) while significant improvement could be observed in terms of both evaluation metrics over baseline 2.

In order to better highlight the changes in the performances of the models in FL and CL we also provide a standard deviation of the performances of the methods as well as the difference in the performance of the individual methods when deployed in CL and FL environments. The lower values of the standard deviation of the performances of the individual methods in CL and FL demonstrate the capabilities of the FL by achieving comparable results with improved privacy. On the other hand, slightly higher variation can be observed in the performance of the baseline and proposed methods. The main contributor in the variation of the performance is the baseline 2, where the results are lower compared to the baseline 1 and AL methods.

V. LESSONS LEARNED

The following lessons can be learned from the experimental results.

- Comparable results could be obtained with AL using fewer samples than the traditional passive learning.
- AL is equally beneficial in both federated and centralized learning environments.
- Some of the methods (or sampling/disagreement strategies) achieve the highest level of performance with fewer samples compared to others. Thus, the number of samples required to achieve the highest accuracy should be considered in the evaluation of AL methods/query selection schemes.
- Stopping criteria for AL schemes is an important factor to be considered in the success of the schemes in an application as after a certain number of iterations the algorithm is forced to pick less relevant samples, which might harm the performance.
- The performance of an algorithm is not affected much generally in FL; however, a significant amount of training samples at each client is required to properly train the local models. Performance could be significantly affected in the case of the unavailability of sufficient training samples at each client to train the local model. AL could be deployed in such cases to obtain relevant samples without involving human annotators.

VI. CONCLUSIONS AND FUTURE WORKS

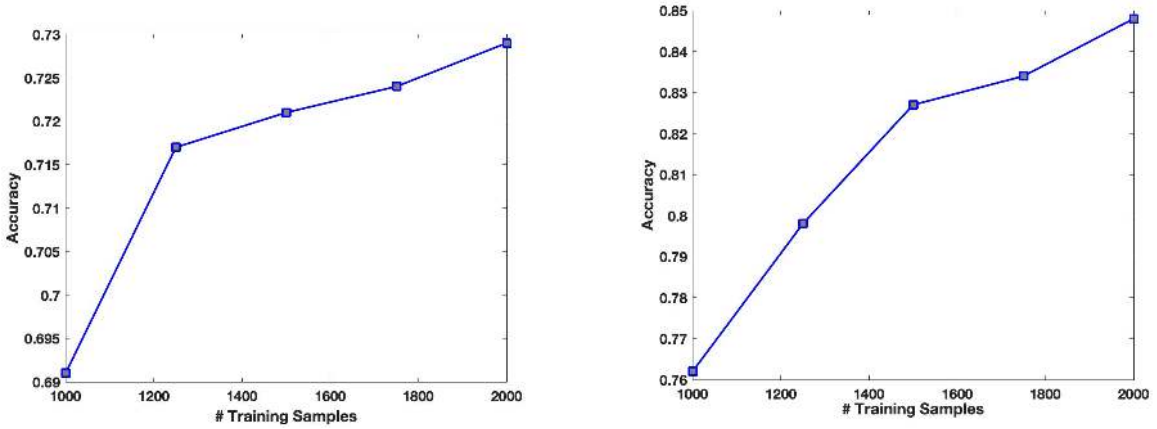
In this paper, we presented an AL-based FL framework to utilize unlabelled samples at clients for training local models in two interesting applications. A detailed evaluation of two different pool-based AL methods under several sampling and disagreement strategies have been provided. Moreover, we show that AL could be equally beneficial in federated and centralized learning in general and the applications lacking in large-scale annotated datasets. In addition, we analyze the impact of fewer training samples at clients on the performance of the global model. In the current implementation, we treated AL as an offline process to automatically annotate training samples at a client before participating in FL.

In the future, we aim to extend the framework to an online AL by exploring how unlabelled data can be incorporated in training local models during different communication rounds of FL, which seems a more challenging task due to several reasons. In such a case, there would be two main challenges. Firstly, if we keep the number of samples to be picked at each iteration very high, there will be higher variations in the performance of the global model, and on the other hand, keeping it low will increase the number of communication rounds.

REFERENCES

[1] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," IEEE Transactions on Knowledge and Data Engineering, 2019.
[2] K. Ahmad, K. Pogorelov, M. Riegler, O. Ostroukhova, P. Halvorsen, N. Conci, and R. Dahyot, "Automatic detection of passable roads after floods in remote sensed and social media data," Signal Processing: Image Communication, vol. 74, pp. 110–118, 2019.
[3] N. Said, K. Ahmad, M. Riegler, K. Pogorelov, L. Hassan, N. Ahmad, and N. Conci, "Natural disasters detection in social media and satellite imagery: a survey," Multimedia Tools and Applications, vol. 78, no. 22, pp. 31 267–31 302, 2019.
[4] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, "Social media and satellites," Multimedia Tools and Applications, vol. 78, no. 3, pp. 2837–2875, 2019.
[5] K. Ahmad, K. Khan, and A. Al-Fuqaha, "Intelligent fusion of deep features for improved waste classification," IEEE Access, 2020.
[6] X.-Y. Zhang, H. Shi, X. Zhu, and P. Li, "Active semi-supervised learning based on self-expressive correlation with generative adversarial networks," Neurocomputing, vol. 345, pp. 103–113, 2019.
[7] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 3, pp. 606–617, 2011.
[8] C. Mayer and R. Timofte, "Adversarial sampling for active learning," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3071–3079.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3038676, IEEE Access



(a) Natural Disaster Dataset

(b) Waste Classification Dataset

FIGURE 7: Trade-off between the number of training samples and accuracy in FL. The number of clients is fixed (i.e., 5) in this experiment. *The accuracy of the global model increases as we increase the training samples per client.*

TABLE 3: Comparison against the two baselines in terms of accuracy and F1-score. (*Baseline 1* represents the best case where each sample is manually analyzed and annotated while the *Baseline 2* represents a training set with irrelevant images.) *Promising results are obtained by the proposed method, outperforming Baseline 2 and comparable performance against Baseline 1 on both datasets.*

Methods	**Centralized Learning**		**Federated Learning**		**Standard Deviation (FL-CL)**	
	Accuracy	F-Score	Accuracy	F-Score	Accuracy	F-Score
Baseline 1	0.742	.750	0.720	0.727	0.015	0.016
Baseline 2	0.689	0.697	0.687	0.696	0.001	0
Proposed method (MD)	.720	0.730	0.713	0.719	0.004	0.007
Standard Deviation (among methods)	0.026	0.027	0.017	0.015	-	-

(a) **Natural Disaster Image Dataset**

Methods	**Centralized Learning**		**Federated Learning**		**Standard Deviation (FL-CL)**	
	Accuracy	F-Score	Accuracy	F-Score	Accuracy	F-Score
Baseline 1	0.851	0.851	0.830	0.832	0.014	0.013
Baseline 2	0.818	0.818	0.785	0.786	0.023	0.022
Proposed method (MS)	0.844	0.844	0.818	0.813	0.018	0.021
Proposed method (MD)	0.832	0.835	0.825	0.832	0.004	0.002
Standard Deviation (among methods)	0.014	0.014	0.020	0.021	-	-

(b) **Waste Classification Dataset**

[9] F. K. Nakano, R. Cerri, and C. Vens, "Active learning for hierarchical multi-label classification," Data Mining and Knowledge Discovery, pp. 1–35, 2020.

[10] K. Ahmad, M. L. Mekhalfi, and N. Conci, "Event recognition in personal photo collections: An active learning approach," Electronic Imaging, vol. 2018, no. 2, pp. 173–1, 2018.

[11] N. Said, K. Ahmad, N. Conci, and A. Al-Fuqaha, "Active learning for event detection in support of disaster analysis applications," arXiv preprint arXiv:1909.12601, 2019.

[12] Y. Sun and K. Loparo, "Context aware image annotation in active learning," arXiv preprint arXiv:2002.02775, 2020.

[13] W. Liu, X. Chang, L. Chen, D. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann, "Pair-based uncertainty and diversity promoting early active learning for person re-identification," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 2, pp. 1–15, 2020.

[14] G. T. Ngo, T. Q. Ngo, and D. D. Nguyen, "Image retrieval with relevance feedback using svm active learning," International Journal of Electrical and Computer Engineering, vol. 6, no. 6, p. 3238, 2016.

[15] J. Yuan, X. Hou, Y. Xiao, D. Cao, W. Guan, and L. Nie, "Multi-criteria active deep learning for image classification," Knowledge-Based Systems, vol. 172, pp. 86–94, 2019.

[16] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," IEEE Transactions on Geoscience and Remote Sensing, 2020.

[17] C. Mu, J. Liu, Y. Liu, and Y. Liu, "Hyperspectral image classification based on active learning and spectral-spatial feature fusion using spatial coordinates," IEEE Access, vol. 8, pp. 6768–6781, 2020.

[18] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in Advances in Neural Information Processing Systems, 2017, pp. 4424–4434.

[19] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[21] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," IEEE Communications Surveys & Tutorials, 2020.

[22] H. B. McMahan, E. Moore, D. Ramage, S. Hampson et al., "Communication-efficient learning of deep networks from decentralized data," arXiv preprint arXiv:1602.05629, 2016.

[23] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," arXiv preprint arXiv:1812.06127, 2018.

[24] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," arXiv preprint arXiv:1812.00984, 2018.
[25] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1175–1191.
[26] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv preprint arXiv:1712.07557, 2017.
[27] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in Esann, 2013.
[28] L. Huang, Y. Yin, Z. Fu, S. Zhang, H. Deng, and D. Liu, "Loadaboost: Loss-based adaboost federated machine learning on medical data," arXiv preprint arXiv:1811.12629, 2018.
[29] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," 2020.
[30] M. Johnson, D. Murthy, B. Roberstson, R. Smith, and K. Stephens, "Disasternet: Evaluating the performance of transfer learning to classify hurricane-related images posted on twitter," in Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020.
[31] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ACM, 2017, pp. 569–576.
[32] B. Bischke, P. Helber, Z. Zhao, J. De Bruijn, and D. Borth, "The multimedia satellite task at mediaeval 2018: Emergency response for flooding events," in 2018 Working Notes Proceedings of the MediaEval Workshop, MediaEval 2018. CEUR-WS. org, 2018, pp. 1–3.
[33] B. Bischke, P. Helber, E. Basar, S. Brugman, Z. Zhao, and K. Pogorelov, "The multimedia satellite task at mediaeval 2019: Flood severity estimation," in Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France.
[34] B. Bischke, P. Bhardwaj, A. Gautam, P. Helber, D. Borth, and A. Dengel, "Detection of flooding events in social multimedia and satellite imagery using deep neural networks," in Working Notes Proceedings MediaEval Workshop, 2017, p. 2.
[35] Y. Feng, S. Shebotnov, C. Brenner, and M. Sester, "Ensembled convolutional neural network models for retrieving flood relevant tweets," in Proceedings of the MediaEval 2018 Workshop, Sophia-Antipolis, France, Oct. 29-31, 2018.
[36] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and P. Halvorsen, "Cnn and gan based satellite and social media data fusion for disaster detection." in MediaEval, 2017.
[37] J. Bobulski and M. Kubanek, "Waste classification system using image processing and convolutional neural networks," in International Work-Conference on Artificial Neural Networks. Springer, 2019, pp. 350–361.
[38] O. Adedeji and Z. Wang, "Intelligent waste classification system using deep learning convolutional neural network," Procedia Manufacturing, vol. 35, pp. 607–612, 2019.
[39] A. H. Vo, M. T. Vo, T. Le et al., "A novel framework for trash classification using deep transfer learning," IEEE Access, vol. 7, pp. 178 631–178 639, 2019.
[40] G. E. Sakr, M. Mokbel, A. Darwich, M. N. Khneisser, and A. Hadi, "Comparing deep learning and support vector machines for autonomous waste sorting," in 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). IEEE, 2016, pp. 207–212.
[41] S. Frost, B. Tor, R. Agrawal, and A. G. Forbes, "Compostnet: An image classifier for meal waste," in 2019 IEEE Global Humanitarian Technology Conference (GHTC). IEEE, 2019, pp. 1–4.
[42] Y. Chu, C. Huang, X. Xie, B. Tan, S. Kamal, and X. Xiong, "Multilayer hybrid deep-learning method for waste classification and recycling," Computational Intelligence and Neuroscience, vol. 2018, 2018.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. of the IEEE CVPR, 2016, pp. 770–778.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. of the IEEE CVPR. IEEE, 2009, pp. 248–255.
[45] V. Ruiz, Á. Sánchez, J. F. Vélez, and B. Raducanu, "Automatic image-based waste classification," in International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer, 2019, pp. 422–431.
[46] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2285–2294.
[47] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, "Cnn-rnn: a large-scale hierarchical image classification framework," Multimedia Tools and Applications, vol. 77, no. 8, pp. 10 251–10 271, 2018.
[48] H. Zhu and Y. Jin, "Multi-objective evolutionary federated learning," IEEE transactions on neural networks and learning systems, 2019.
[49] M. Yang and G. Thung, "Classification of trash for recyclability status," CS229 Project Report, vol. 2016, 2016.
[50] W. Yuan, Y. Han, D. Guan, S. Lee, and Y.-K. Lee, "Initial training data selection for active learning," in Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, 2011, pp. 1–7.

LULWA AHMED is currently working as a project manager at Gulf Business Machines in Qatar. She received her Bachelors degree in Computer Science from Carnegie Mellon University, Qatar in 2013. Her research interests include Machine Learning, Active learning, Federated learning and Deep Learning in particular to the deployment of IoT in smart city infrastructures.

KASHIF AHMAD is currently working as a research fellow at the division of Computer Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. He also worked as Postdoctoral researcher at ADAPT, Trinity College, Dublin, Ireland. He received his PhD degree from University of Trento, Italy in 2017 where he worked with Multimedia lab in DISI. He received his Bachelors and Masters degrees from University of Engineering and technology, Peshawar, Pakistan in 2010 and 2013, respectively. He authored and co-authored more than 40 journal and conference publications. His research interests include Multimedia analysis, Computer Vision, Machine Learning and Signal processing applications in Smart Cities. He is program committee member of multiple international conferences including CBMI, ICIP, and MMSys.

NAINA SAID is currently working as a lecturer in the department of Computer Systems Engineering, University of Engineering and Technology Peshawar, Pakistan. She did her bachelors and Masters from the same department in 2015 and 2018, respectively. She also worked as a Research Assistant in the same department with a team of researchers working on the development of scalable P2P streaming application. Her research interests include image and video processing, multimedia streaming applications and machine learning.

BASHEER QOLOMANY received the Ph.D. and second master's en-route to Ph.D. degrees in Computer Science from Western Michigan University (WMU), Kalamazoo, MI, USA, in 2018. He also received his B.Sc. and M.Sc. degrees in computer science from University of Mosul, Mosul city, Iraq, in 2008 and 2011, respectively. He is currently an Assistant Professor at Department of Cyber Systems, University of Nebraska at Kearney (UNK), Kearney, NE, USA. Previously, he served as a Visiting Assistant Professor at Department of Computer Science, Kennesaw State University (KSU), Marietta, GA, USA, in 2018-2019; a Graduate Doctoral Assistant at Department of Computer Science, WMU, in 2016-2018; he also served as a Lecturer at the Department of Computer Science, University of Duhok, Kurdistan region of Iraq, in 2011-2013. His research interests include machine learning, deep learning, Internet of Things, smart services, cloud computing, and big data analytics. Dr. Qolomany has served as a reviewer of multiple journals, including IEEE Internet of Things journal, Energies — Open Access Journal, and Elsevier - Computers and Electrical Engineering journal. He also served as a Technical Program Committee (TPC) member and a reviewer of some international conferences including IEEE Globecom, IEEE IWCMC, and IEEE VTC.

JUNAID QADIR is the director of the IHSAN Research Lab and the Chairperson of the Electrical Engineering Department at the Information Technology University (ITU) of Punjab in Lahore, Pakistan. His primary research interests are in the areas of computer systems and networking, applied machine learning, using ICT for development (ICT4D); and engineering education. He has published more than 100 peer-reviewed articles at various high-quality research venues including more than 50 impact-factor journal publications at top international research journals including IEEE Communication Magazine, IEEE Journal on Selected Areas in Communication (JSAC), IEEE Communications Surveys and Tutorials (CST), and IEEE Transactions on Mobile Computing (TMC). He was awarded the highest national teaching award in Pakistan—the higher education commission's (HEC) best university teacher award—for the year 2012-2013. He has been appointed as ACM Distinguished Speaker for a three-year term starting from 2020. He is a senior member of IEEE and ACM.

ALA AL-FUQAHA [S'00-M'04-SM'09] received Ph.D. degree in Computer Engineering and Networking from the University of Missouri-Kansas City, Kansas City, MO, USA, in 2004. He is currently a professor at Hamad Bin Khalifa University (HBKU). His research interests include the use of machine learning in general and deep learning in particular in support of the data-driven and self-driven management of large-scale deployments of IoT and smart city infrastructure and services, Wireless Vehicular Networks (VANETs), cooperation and spectrum access etiquette in cognitive radio networks, and management and planning of software defined networks (SDN). He is a senior member of the IEEE and an ABET Program Evaluator (PEV). He serves on editorial boards of multiple journals including IEEE Communications Letter and IEEE Network Magazine. He also served as chair, co-chair, and technical program committee member of multiple international conferences including IEEE VTC, IEEE Globecom, IEEE ICC, and IWCMC.