2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) | 979-8-3503-6944-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/DCOSS-IoT61029.2024.00025

FedMIL: Federated-Multiple Instance Learning for Video Analysis with Optimized DPP Scheduling

Ashish Bastola
School of Computing
Clemson University
Clemson, USA
abastol@clemson.edu

Hao Wang
School of Computing
Clemson University
Clemson, USA
hao9@clemson.edu

Xiwen Chen
School of Computing
Clemson University
Clemson, USA
xiwenc@clemson.edu

Abolfazl Razi
School of Computing
Clemson University
Clemson, USA
arazi@clemson.edu

***Abstract*—Many AI platforms, including traffic monitoring systems, use Federated Learning (FL) for decentralized sensor data processing for learning-based applications while preserving privacy and ensuring secured information transfer. On the other hand, applying supervised learning to large data samples, like high-resolution images requires intensive human labor to label different parts of a data sample. Multiple Instance Learning (MIL) alleviates this challenge by operating over labels assigned to the 'bag' of instances. In this paper, we introduce Federated Multiple-Instance Learning (FedMIL). This framework applies federated learning to boost the training performance in video-based MIL tasks such as vehicle accident detection using distributed CCTV networks. However, data sources in decentralized settings are not typically Independently and Identically Distributed (IID), making client selection imperative to collectively represent the entire dataset with minimal clients. To address this challenge, we propose DPPQ, a framework based on the Determinantal Point Process (DPP) with a quality-based kernel to select clients with the most diverse datasets that achieve better performance compared to both random selection and current DPP-based client selection methods even with less data utilization in the majority of non-IID cases. This offers a significant advantage for deployment on edge devices with limited computational resources, providing a reliable solution for training AI models in massive smart sensor networks.** [1]

***Index Terms*—Federated Learning, Multiple Instance Learning, Determinantal Point Process, Non-IID Distribution, Traffic Analysis, Crash Detection, Smart Transportation.**

I. Introduction

In recent years, video-based AI platforms have seen prolific growth and unprecedented applications across diverse sectors. Smart transportation systems, by using Autonomous Vehicles (AVs) and AI-based traffic control platforms, are revolutionizing mobility with enhanced efficiency and safety[1, 4, 5, 6, 9, 27]. Specifically, accident detection plays a significant role in ensuring safety during unforeseen circumstances. Prompt detection of accidents enables emergency services to respond on time, while also alerting ongoing traffic to prevent further catastrophic incidents. However, even with millions of decentralized CCTV cameras across the US infrastructures, accident detection remains an often overlooked area [29, 31].

This paper focuses on developing data selection strategies to construct generalizable learning-based models for traffic analysis that require access to a minimal number of roadside units (RSUs) that collectively best represent the entire traffic. Building learning-based inference and control models typically requires access to a sheer number of data sources (roadside cameras, in our case) to collect a massive amount of data to represent various situations. Our approach of selective access to the best set of data sources not only minimizes the processing and communication costs but also facilitates the development of reliable models with limited access to a handful of roadside infrastructures.

Centralized data storage and processing usually face several privacy challenges. While training foundational models necessitates extensive data collection of these sensitive data, it is almost impossible to gather this volume of data without potentially infringing upon user privacy with current technologies [21, 32]. The risk of data breaches and unauthorized access is a constant threat in such situations [11, 30]. Secondly, centralized processing can be inefficient due to the high computational and storage demands placed on central servers, also an issue of latency in data transfer when exchanging a massive amount of data, which prevents the updating capabilities of these systems[18].

For these reasons, most AI platforms depart from central learning and bare minimum learning on edge servers towards Federated Learning (FL), with the core idea of sharing local model parameters to implement a global model instead of directly sharing raw data. Using this system, each client performs training on its own dataset, and then shares model parameters to build an inclusive global model that can operate in wide geographic areas with varying types, rates, and frequency of crash types.

On the other hand, Multiple Instance Learning (MIL) is a weakly supervised learning paradigm that trains models on bags of data samples that share the same label, which can be applied to various fields such as medical image classification and anomaly detection [28, 36, 37]. It also has gained widespread use in video-based incident detection applications by treating each video frame as an individual instance and the entire video as a bag of such instances [3, 26]. Furthermore, efforts have

[1]This material is based upon work supported by the National Science Foundation under Grant Numbers CNS2204721 and CNS-2204445.

2325-2944/24/$31.00 ©2024 IEEE
DOI 10.1109/DCOSS-IoT61029.2024.00025

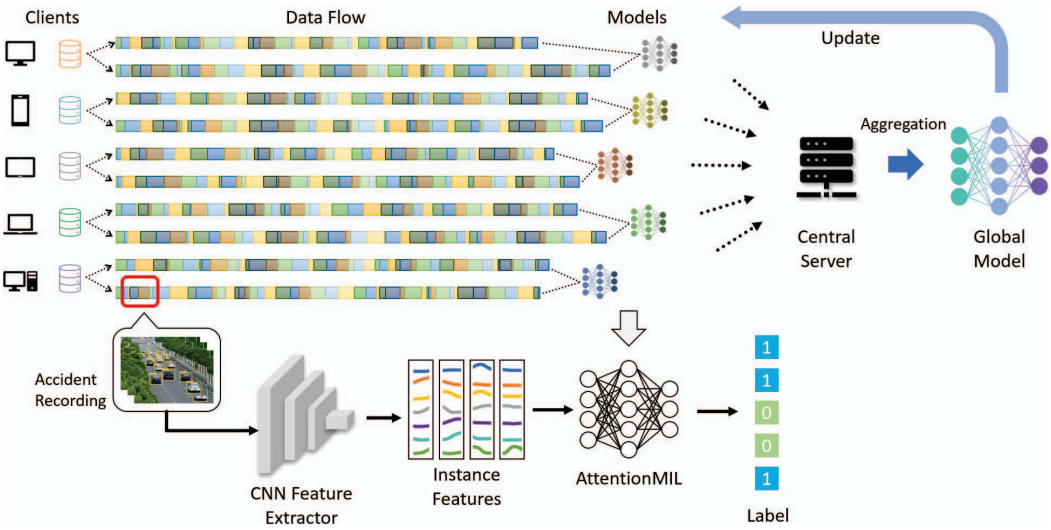

Fig. 1. The architecture of FedMIL. Compared to classical central learning, the training process is broken down to the client level to prevent massive data exchange. In advance, with a pre-trained CNN model to extract the video feature and a lightweight MIL model for bag-level classification, the computational cost for edge devices is greatly reduced.

been made to enhance these models by focusing on multi-scale continuity [13]. This enhancement is based on the understanding that anomalous events in real-world videos typically unfold continuously, necessitating their evaluation across the entire bag of instances. The use of multiple instance learning significantly reduces the need for extensive data preprocessing and feature engineering, which reduces both the labor cost and the demand for computational resources to train AI models on edge devices.

To combine both advantages, we propose FedMIL, the federated version of MIL, as a technical contribution. As shown in Figure 1, MIL provides the benefit of developing weakly supervised models to reduce the computation cost on the client side, and FL provides the benefit of avoiding massive data exchange. Different than another federated MIL setting that deploys models on only a few sites for weakly-supervised image classification in [24], our study focuses on the video-based analysis among hundreds of data nodes, where the total data flow and computational cost is thousands of times than image-based tasks. Furthermore, compared to image classification tasks, our proposed framework can be extended to other tasks that are more practical, such as video anomaly detection, traffic volume analysis, human action recognition, and deepfake detection [6, 23, 26].

Although the combination of FL and MIL eliminates the need for massive data exchange and reduces computation power requirements for edge devices for complex tasks, orchestrating model parameter exchange among a sheer number of clients might not be feasible for practical reasons. For instance, the FL scheduler system only allows a limited number of communication ports at one time, and the network bandwidth is highly constrained. Meanwhile, due to the inherent diversity of data across different clients [35], the rate and types of crashes can be extremely different across different cameras in the real world. For instance, an RSU located in a less crowded area or a safe zone might not witness frequent accidents, thus the locally trained model may fail to detect accidents if one occurs [6]. Thus, these non-IID data distributions may greatly impact the model's performance due to weight divergence issues [7, 25].

Therefore, developing client selection strategies to reduce the negative impact of uneven data distribution, and improve the performance under constrained communication or limited local processing power becomes a key challenge in this research area.

Intuitively, selecting RSUs with more diversity in terms of data and label distributions is more desirable. In recent studies, the Determinantal Point Process (DPP) achieved significant gains in federated learning applications for capturing diversity [19]. DPP enforces repulsion to alike items and prioritizes selecting clients whose data distributions are varied, aiming to reduce redundancy in the training process. With classical DPP-based client selection, the central model tends to collect model parameters from clients with more diverse data, avoiding convergence at lower accuracy [34].

We take this step further by proposing a power-of-choice [8] version of DPP embedded as a quality matrix to achieve the best of both diversity and loss gradients, which makes the selection even more robust[8]. Thus, the second contribution of this work is the demonstration of the efficacy of selecting a diverse subset of clients that accurately represent the overall dataset, which ensures the results are robust and broadly applicable across various clients in real-world settings.

II. Problem Formulation

The key aspect of the problem in this study is the data distribution among clients. While IID sampling offers a baseline approach where each client receives equitable and balanced

representative samples, non-IID sampling models the skewness in data distribution, which aligns more closely with real-world data distribution scenarios[20, 26]

To simulate non-IID distribution, we introduce class and feature-space-based imbalance, as illustrated in Figure 2. For the first type, we use power-law to assign imbalanced crash labels to each client [2, 16, 22]. This mirrors the situations where some RSUs may observe crashes or specific types of crashes more frequently than others. For the second type, we introduce an underlying distribution-based imbalance such that each client receives a subset whose data type is imbalanced according to a Dirichlet distribution [20]. For instance, the traffic compositing on a highway contains more trucks than at a city intersection. Therefore, their features extracted from video frames may follow statistically different distributions.

Specifically, if labeled video pairs are shown by $(V_i, y_i) = ([F_i(1), F_i(2), \cdots, F_i(M)], y_i)$ with V_i being a video clip, $F_i(j)$ being its j^{th} frame, and y_i, being the class label, then we impose imbalance to both V_i and y_i.We apply power-law to class labels y_i and Dirichlet distribution to samples V_i as detailed next, to divide the data into non-IID subsets among clients.

A. Label-based Imbalance (Type I Non-IID)

We follow the imbalance strategy for the binary distribution by changing the ratio of class labels for each client. We vary the ratio of class labels client-wise by tuning the power-law.

Specifically, we use:

$$n_p = V \cdot (p + H)^{\beta} \tag{1}$$

where n_p is the proportion of class 1 counts for client p, V and H denote the vertical and horizontal scaling, respectively, and the exponent β is a tuning factor to control the the ratio of class labels. We fix the horizontal and vertical scaling parameters to certain values to account for the unequal count of either of the classes before beginning the distribution, then vary β to generate different levels of disparity. To add non-uniformity in label assignment, we select a hold-out set from both classes in the two selected distributions of β, so some clients are assigned only one label type. Here, we assume binary labels for crash detection, but it easily extends to multi-level cases.

B. Distribution-based Imbalance (Type II Non-IID)

While class-based imbalance provides disparity of labels, the distribution-based imbalance better captures the disparity among the attributes of video frames that represent real-world factors, such as climate conditions (sunny, snowy, and cloudy), road features (off-road, highway, or regular), environment (urban and rural), light conditions, etc.

The process includes the following steps. Assume a dataset offers video clips V_i and a set of feature vectors $\phi_W(F_j)$ for each frame F_j extracted using a pre-trained VGG network. These feature vectors are clustered using k-means [14] to identify $C = 10$ different underlying classes as shown in Fig. 2. Then, we perform majority voting on the obtained feature vectors to label the sample (aka video clip V_i). Finally, we apply Dirichlet distribution to create the imbalance corresponding to these cluster labels. We follow the strength variation of this distribution as followed by [20], where we sample $\vec{n}_p = [n_p^1, n_p^2, \cdot, n_p^C] \sim \mathrm{Dir}(\vec{n}; \vec{\alpha})$ using

$$\mathrm{Dir}(\vec{n}; \vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{c=1}^{C} n_p^{\alpha_c - 1} \tag{2}$$

and allocate n_p^c proportion of the instances of class c to client p, where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution and $\vec{\alpha} = [\alpha_1, \alpha_2, \cdot, \alpha_C]$ denotes the concentration parameter ($\alpha_c > 0$). Here, $\mathrm{B}(\vec{\alpha})$ is the multinomial beta function given by:

$$\mathrm{B}(\vec{\alpha}) = \frac{\prod_{c=1}^{C} \Gamma(\alpha_c)}{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)} \tag{3}$$

where L is the gamma function given by $\Gamma(\alpha_c) = (\alpha_c - 1)!$. The major advantage of this approach is that we can change the imbalance level by varying the concentration parameter α. The smaller the α, the more unbalanced the partition.

III. Algorithm Design

Our contribution can be divided into two folds. First, we optimized the existing FL framework to better serve the MIL specification. In our case, the FL architecture is specifically designed for video analysis, where each frame (instance) has multiple features. Second, we propose the client selection algorithm based on the DPPQ method. Proposed by [19], the quality-diversity decomposition is the special reparameterization of the normal DPP-based method. This method has been proven to achieve a significant gain in video summarization applications [12]. We show that it can be applied to similar tasks, such as the video classification in this work.

A. Federated Learning Framework

1) Individual Client Model and Training: Each client p trains the model on their local data D_p by minimizing the local loss function $L_p(w)$ using stochastic gradient descent:

$$L_p = \frac{1}{|D_p|} \sum_{k=1}^{|D_p|} L_k(g_w(\vec{z}_k), \vec{y}_k) \tag{4}$$

where L_p is the loss function for each client p. $\vec{y}_k$ is the label, $\vec{z}_k$ is the bag level representation(vector in our case) and L_k is the loss for each sample k, and g_w is the model parameterized by weights w.

2) Global Aggregation: After local training, clients send their model weights w_p to the central server. The server performs a weighted average to update the global model:

$$w_g^{t+1} = \sum_{p=1}^{P} \frac{n_p}{n} w_p^t \tag{5}$$

where, w_g^{t+1} is the aggregated global model ready for $t+1$-th round, n_p is the number of samples at client p, n is the total number of samples, and t denotes the current round.

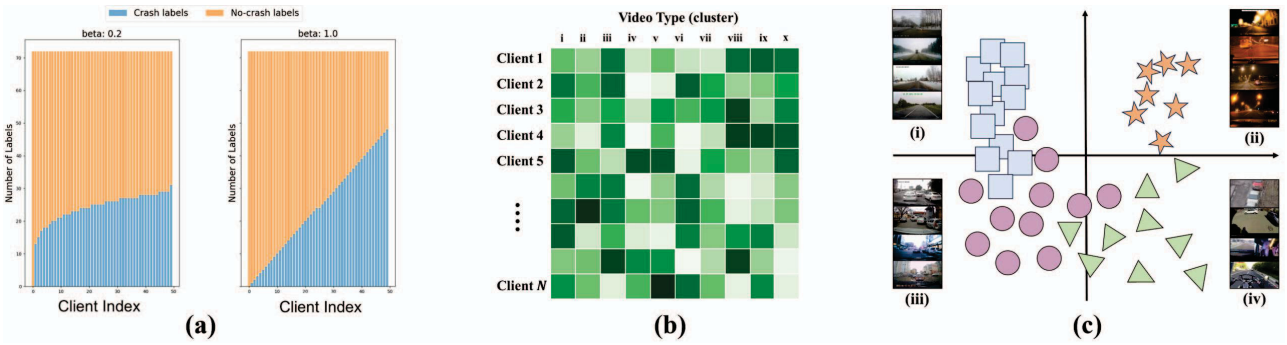

Fig. 2. Simulating real-world data distribution: (a) the data distribution of different clients is imbalanced directly at the class level; (b) data imbalance based on Dirichlet distribution over underlying data clusters, and (c) K-means clustering on VGG features to identify underlying clusters, where (i) represents the video condition under snow day, (ii) represents the night snippets, (iii) is the urban recording, and (iv) represent other scenarios.

B. Model Design for FedMIL

We apply AttentionMIL [15], A multi-instance learning model that processes bags of instances. It assigns attention weights to instances, enabling the model to focus on the most informative instances in each bag.

Thus, for the weighted average of instances where the image feature determines the weights, the extracted embeddings $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ are the bag of K embeddings for K samples(which have n instances each) that are the output of the feature extractor (which is a Fully connected layer). Thus, as proposed by [10, 15], the MIL Pooling or the bag-level aggregation to form the bag representation for the kth sample is given by:

$$\vec{z}_k = \sum_{m=1}^{n} a_m \vec{h}_m \tag{6}$$

where a_m is the gated attention with $\tanh(\cdot)$ non-linearity given by,

$$a_m = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\vec{h}_m^\top\right) \odot \text{sigm}\left(\mathbf{U}\vec{h}_m^\top\right)\right)\right\}}{\sum_{j=1}^{n} \exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{V}\vec{h}_j^\top\right) \odot \text{sigm}\left(\mathbf{U}\vec{h}_j^\top\right)\right)\right\}} \tag{7}$$

where $\mathbf{U} \in \mathbb{R}^{L \times M}$ are parameters, $\odot$ is an element-wise multiplication and $\text{sigm}(\cdot)$ is the sigmoid non-linearity. The gating mechanism introduces a learnable non-linearity that potentially removes the troublesome linearity in $\tanh(\cdot)$.

The Attention-based MIL handles the multi-instance nature of the video data. We utilize the stochastic gradient descent (SGD) optimizer with a Lookahead mechanism for robust optimization [33]. The model was trained using a cross-entropy loss function suitable for the binary classification task (accident vs. non-accident). Thus, the global loss function is given by:

$$L_g = \sum_{p=1}^{P} \frac{n_p}{\sum_{p \in \mathcal{P}} n_p} L_p \tag{8}$$

where P is the total number of selected clients, L_p is the loss for p-th client defined in 4 (however $\vec{z}_k$ now is the attention-based bag-level aggregation defined in 6 for k-th bag). Based on this, we formulate the Federated learning architecture to train multiple clients in a decentralized fashion.

C. Client Profiling

Assume each sample of the dataset is represented as an image feature. Considering the p-th client, let the k-th sample be represented as x_k. Each sample contains 50 features corresponding to the image frame. Then, the i-th feature corresponding to i-th frame undergoes the following feature extraction $\vec{h}_i = \text{ReLU}(\mathbf{W}\mathbf{x}_i + \mathbf{b})$. Note that this intermediary feature extractor helps further ensure security, which makes them robust to inference-based attacks. Thus, for the k-th sample, we have $\mathbf{h_k} = \left\{\vec{h}_1, \ldots, \vec{h}_n\right\}$ where $n = 50$ is the total number of frames in a single video sample. Since we employ SGD, the batch size is the total number of samples for each client after the distribution. Thus for each pth client with dataset D_p, we compute the average feature profile $\mathbf{h_p}$ given by:

$$\mathbf{h_p} = \frac{1}{|D_p|} \sum_{k=1}^{|D_p|} \mathbf{h}_k \tag{9}$$

Further, we profile the average loss of each client as follows,

$$L_p = \frac{1}{|D_p|} \sum_{k=1}^{|D_p|} L_k \tag{10}$$

where L_p is the loss profile for the p-th client and L_k is the loss value computed for each sample k.

D. Client Selection via Quality-Diversity Decomposition

Let $\mathcal{Y} = \{1, 2, \cdots, N\}$ be the ground set of N clients. We then profile each p-th client C_p using the average output of their feature representation given by Eq. 9 and their average loss given by Eq. 10, which are the only data shared to the central server as their representative information and only during the initialization phase.

With this information, the central server first computes the similarity matrix $\mathbf{S} = \{s_{r,t}\}_{C \times C}$ as implemented in [19, 34] using the following:

$$s_{r,t} = 1 - \left(\frac{s^0_{r,t} - \min(\mathbf{S}^0)}{\max(\mathbf{S}^0) - \min(\mathbf{S}^0)} \right) \tag{11}$$

where, $s^0_{r,t} = \|\mathbf{h}_r - \mathbf{h}_t\|_2$ with $\mathbf{h}_r, \mathbf{h}_t \forall r, t \in C$, and $\mathbf{S^0}$ is the full difference matrix of every client combination defined by $\{s^0_{r,t}\}_{C \times C}$.

Similarly, the quality matrix is defined as the diagonal matrix of loss values $\mathbf{L_p}$ pertaining to each client $C_p \in \mathcal{Y}$ i.e.

$$\mathbf{Q} = \begin{pmatrix} q_1 & 0 & \cdots & 0 \\ 0 & q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_N \end{pmatrix} \tag{12}$$

where, q_p for $p \in \{1, 2, \cdots, N\}$ is the min-max scaled quality value of the p-th client with lowerbound ϵ given by eq 13. This is important to prevent a loss value of 0 that ruins the positive semidefinite property of the kernel matrix.

$$q_{\mathrm{p}} = \epsilon + \left(\frac{L_p - L_{\min}}{L_{\max} - L_{\min}} \times (1 - \epsilon) \right) \tag{13}$$

We finally calculate the kernel matrix given by $\mathbf{L} = \mathbf{Q}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{Q}$. And following [19], we can express the probability of selecting subset G from $\mathcal{Y}$ as,

$$P_L(G \subseteq \mathcal{Y}) \propto det(\mathbf{L_G}), \tag{14}$$

Thus, from Eq. 14, we can infer the select clients with higher diversity-quality score. i.e. the loss value of the client L_p and the diversity of the extracted features $\mathbf{h_k}$. Thus our method incorporates a selection of clients with higher loss values and are similarly highly diverse. Here, we sample P clients and denote their index set as G. The more detailed working of the algorithm is mentioned in Algorithm 1.

Algorithm 1 FL-DPPQ: Federated Learning with DPP-based Quality Diversity Decomposition

Input: Clients with global model parameters $\mathbf{w}_g^{(0)}$ having their own data distributed using some non-IID distributions defined in Eq. 3 and 1
for each client $c \in \mathcal{Y}$ **do**
Calculate the average loss and average of extracted features using Eqs. 10 and 9 and upload it to the server.
end for
The Central Server calculates the similarity matrix S according to Eq. 11 and quality matrix Q according to 12 and constructs k-DPPQ and selects G of $\mathcal{Y}$ clients with probability given by Eq. 14.
for t **in** $1 : T$ **do**
for each client $c \in G$ **do**
for each local epoch $t' \in 1 : T'$ **do**
Update $\mathbf{w}_c^{(t)}$ using back prop. while minimizing the local objective function given by Eq. 4.
end for
Send updated weight $\mathbf{w}_c^{(t)}$ to the server.
end for
Central server performs the weight aggregation using Eq. 5 and distributes the updated global model to all the clients.
end for
Output: The trained Global Model $\mathbf{w}_g^{(T)}$.

TABLE I
MODEL COMPARISON IN MINIST

Category	Test Accuracy
Random	0.7869
DPP	0.8370
DPPQ	**0.9084 (+8.5%)**

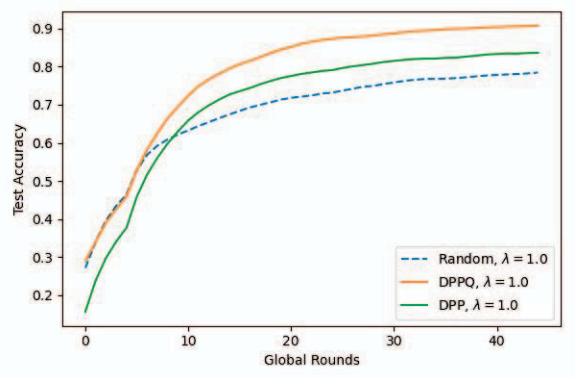

Fig. 3. Training comparison in MNIST dataset with Type I Non-IID distribution of strength ($\alpha = 0.5$) and 100% data utilization

IV. EXPERIMENTS

We show the proposed work on CCD, a typical multi-instance annotated dataset, and we compare the model performance under different settings (e.g., data utilization rate, strength of data imbalance, and non-IID type). The results are averaged over 20 runs. The model was evaluated on a test set with performance metrics including loss, accuracy, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [17]. The result shows that model performance varies in terms of data distribution type. Moreover, our implementation of FedMIL improved the model performance by selecting quality clients via the proposed DPPQ, especially when the data utilization decreased.

a) Car Crash Dataset (CCD) for Traffic Accident Analysis: The Car Crash Dataset (CCD) is a specialized collection for traffic accident analysis. It comprises 1,500 dashcam-captured accident videos and 3,000 normal videos from YouTube and the BDD100K dataset, each containing 50 frames at 10 fps [3]. Advanced feature extraction is employed for all videos using VGG-16 for frames and bounding boxes, detected via Cascade R-CNN with a ResNeXt-101 backbone. Thus, each video is annotated with frame-level binary labels (crash, no crash)

b) Evaluation on MNIST Dataset: To validate the performance of the proposed DPPQ method in general, we test our method on the MNIST dataset that has 10 different classes. We perform an average of 10 runs for extreme non-IID data distribution with Type I Non-IID distribution of strength ($\alpha = 0.5$) and full data utilization ($\lambda = 1.0$). As shown in table I, the proposed DPPQ outperforms both random selection and DPP

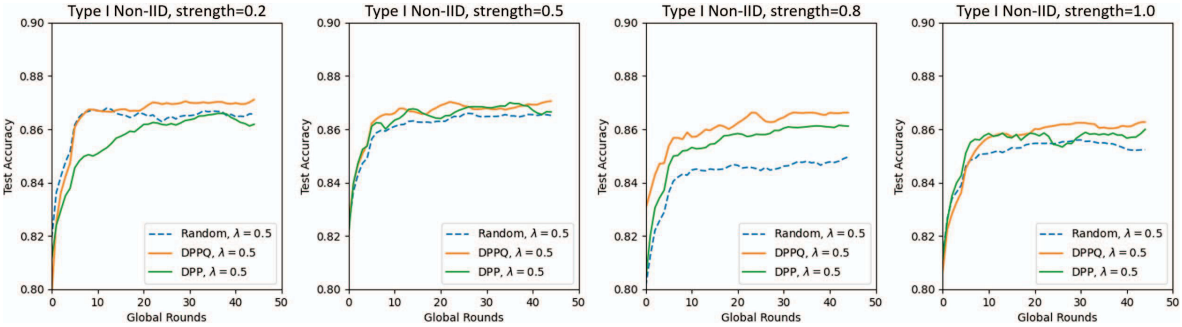

Fig. 4. Model performance comparison in terms of data imbalance based on class labels.

TABLE II
TYPE I NON-IID MODEL COMPARISON

		Strength = 0.2			Strength = 0.5			Strength = 0.8			Strength = 1.0		
Data Utilization	Method	aucs	f1	test acc	aucs	f1	test acc	aucs	f1	test acc	aucs	f1	test acc
0.5	DPPQ	**0.9443**	**0.8689**	**0.8703**	0.9411	**0.8675**	**0.8700**	**0.9425**	**0.8645**	**0.8658**	0.9353	**0.8614**	**0.8621**
	DPP	0.9394	0.8619	0.8632	0.9425	0.8660	0.8676	0.9425	0.8638	0.8624	0.9382	0.8581	0.8589
	Random	0.9404	0.8649	0.8657	0.9386	0.8626	0.8632	0.9277	0.8473	0.8485	0.9335	0.8507	0.8463
0.8	DPPQ	0.9449	**0.8738**	**0.8773**	0.9535	0.8794	0.8812	0.9505	**0.8799**	**0.8791**	**0.9511**	**0.8777**	**0.8765**
	DPP	0.9432	0.8684	0.8720	0.9524	0.8770	0.8778	0.9576	0.8730	0.8712	0.9462	0.8700	0.8695
	Random	0.9503	0.8736	0.8735	0.9578	0.8884	0.8880	0.9441	0.8652	0.8652	0.9488	0.8755	0.8754
1.0	DPPQ	0.9564	**0.8859**	**0.8835**	0.9568	0.8851	0.8854	0.9508	0.8806	0.8791	0.9508	0.8779	**0.8773**
	DPP	0.9579	0.8843	0.8829	0.9585	0.8848	0.8861	0.9562	0.8833	0.8831	0.9523	0.8784	0.8768
	Random	0.9562	0.8829	0.8822	0.9584	0.8918	0.8935	0.9518	0.8733	0.8717	0.9517	0.8756	0.8726

with a considerable gain. Furthermore, Figure 3 shows that DPPQ can reach a higher accuracy in the training process, which demonstrates the benefit of client selection with quality consideration.

A. Type I Non-IID Evaluation

Figure 4 shows the comparison between simple DPP-based sampling as presented in [34], random and our DPPQ-based sampling in the CCD dataset. The different strength values in this case are the tuning parameter for β as defined in Eq. 1 with the data utilization of 50%. From the results, we can infer that DPPQ achieved faster convergence than DPP when analyzed over a strength value of 0.2. We also see random performs faster but fails to maintain the same accuracy as DPPQ after the 15th epoch. We observed similar faster convergence in strength value of 0.8, achieving nearly 2% accuracy gain compared to random selection.

B. Type II Non-IID Evaluation

Figure 5 shows the comparison between random selection, DPP, and the proposed DPPQ method in the CCD dataset under the Dirichlet distribution with a data utilization rate of 50%, where the strength represents the data imbalance ratio. In advance, $\alpha = 0.2$ represents extremely imbalanced, $\alpha = 0.5$ represents moderately imbalanced, $\alpha = 0.8$ represents almost balanced, and $\alpha = 1.0$ indicates the data volume for each client is almost even.

As a result, the overall performances of all models are dropped due to the lower data utilization. Nevertheless, DPPQ outperforms DPP and random selection regardless of the data imbalance ratio. This is a significant advantage when the client's device is busy or has limited computation resources.

TableIII on the other hand, shows the complete results of all experiments. As the data utilization is controlled to 100%, DPPQ shows identical performance to the DPP and random selection when Dirichlet strength α is 0.5 and 1.0, this is because the full-utilized client data enhanced the overall diversity, and the reduction of data imbalance (increase of α) will improve all model performance, especially the random selection. However, when data become more imbalanced ($\alpha = 0.2$), the proposed DPPQ shows better performance than DPP and random selection. This is because the proposed DPPQ not only considered the data volume but also the data quality of each client. With the data utilization dropped to 80%, the overall performances of all models are dropped due to the lower data utilization. Nevertheless, the proposed DPPQ and DPP still retain a better performance than the random selection.

C. Data Utilization Analysis

Figure 6, on the other hand, shows the model performance in terms of data utilization ratio. The result is using type-II non-IID (Dirichlet distribution) with strength $\alpha = 0.5$. Since the proposed DPPQ is optimized for Dirichlet distribution, its performance is better and more stable than DPP and random selection. As the data utilization keeps dropping, the proposed DPP still outperforms random selection, proving its ability for low data-flow scenarios.

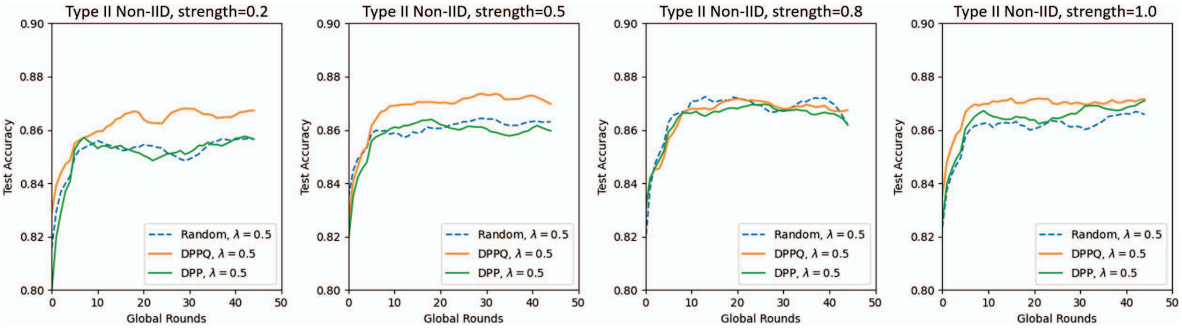

Fig. 5. Model performance comparison in terms of data imbalance based on Dirichlet distribution.

TABLE III
TYPE II NON-IID MODEL COMPARISON

		Strength = 0.2			Strength = 0.5			Strength = 0.8			Strength = 1.0		
Data Utilization	Method	aucs	f1	test acc	aucs	f1	test acc	aucs	f1	test acc	aucs	f1	test acc
0.5	DPPQ	**0.9389**	**0.8630**	**0.8660**	**0.9411**	**0.8676**	**0.8713**	0.9439	0.8669	**0.8683**	**0.9444**	**0.8703**	**0.8710**
	DPP	0.9316	0.8538	0.8561	0.9378	0.8604	0.8596	0.9465	0.8538	0.8643	0.9437	0.8683	0.8693
	Random	0.9276	0.8531	0.8563	0.9378	0.8579	0.8630	0.9436	0.8685	0.8670	0.9374	0.8628	0.8658
0.8	DPPQ	0.9475	**0.8719**	**0.8733**	0.9532	**0.8789**	**0.8782**	0.9563	**0.8837**	**0.8855**	**0.9525**	**0.8782**	**0.8792**
	DPP	0.9485	0.8705	0.8705	0.9534	0.8786	0.8776	0.9574	0.8802	0.8809	0.9525	0.8779	0.8785
	Random	0.9413	0.8603	0.8589	0.9510	0.8753	0.8761	0.9539	0.8825	0.8820	0.9511	0.8751	0.8765
1.0	DPPQ	**0.9581**	**0.8857**	**0.8860**	0.9564	**0.8870**	**0.8870**	0.9527	0.8815	0.8842	0.9575	0.8871	**0.8880**
	DPP	0.9531	0.8816	0.8830	0.9568	0.8851	0.8858	0.9586	0.8853	0.8856	0.9602	0.8878	0.8880
	Random	0.9465	0.8710	0.8694	0.9543	0.8844	0.8856	0.9560	0.8869	0.8874	0.9544	0.8825	0.8844

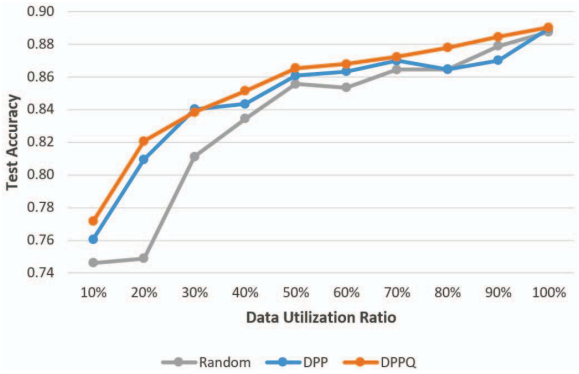

Fig. 6. Model performance comparison in terms of data utilization ratio.

V. CONCLUSION

In this study, we presented an approach for accident detection based on traffic videos, integrating a federated learning framework with multiple instance learning to boost the training process on edge devices. Furthermore, we introduce an improved client selection strategy based on DPPQ, which prioritizes diversity and loss gradient. This strategy not only enhances the model's generalization capabilities but also provides robust performance under low-data utilization compared to the classical DPP-based approach, which we also validated in the MNIST dataset as a general benchmark. The test results of the CCD dataset prove that the proposed DPPQ model retains identical performance even though the data become extremely imbalanced. More importantly, the proposed DPPQ model performs better than classical DPP and random selection even with less data utilization. Our implementation addresses the challenges of both non-IID data distributions and limited computation power in practical applications in the real world.

In advance, the proposed FedMIL framework can be easily extended to more practical video-based tasks such as video anomaly detection, traffic volume analysis, human action recognition, and even deepfake detection, with ensured privacy and security.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers CNS2204721 and CNS-2204445.

REFERENCES

[1] Adriano Alessandrini, Andrea Campagna, Paolo Delle Site, Francesco Filippi, and Luca Persia. Automated vehicles and the rethinking of mobility and cities. Transportation Research Procedia, 5:145–160, 2015.

[2] Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeff Bilmes. Diverse client selection for federated learning via submodular maximization. In International Conference on Learning Representations, 2022.

[3] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In ACM Multimedia Conference, May 2020.

[4] Ashish Bastola, Md Atik Enam, Ananta Bastola, Aaron Gluck, and Julian Brinkley. Multi-functional glasses for the blind and

visually impaired: Design and development. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 67, pages 995–1001. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[5] Ashish Bastola, Julian Brinkley, Hao Wang, and Abolfazl Razi. Driving towards inclusion: Revisiting in-vehicle interaction in autonomous vehicles. arXiv preprint arXiv:2401.14571, 2024.

[6] Xiwen Chen, Hao Wang, Abolfazl Razi, Brendan Russo, Jason Pacheco, John Roberts, Jeffrey Wishart, Larry Head, and Alonso Granados Baca. Network-level safety metrics for overall traffic safety assessment: A case study. IEEE Access, 11:17755–17778, 2022.

[7] Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Yağan. Bandit-based communication-efficient client selection strategies for federated learning. In 2020 54th Asilomar Conference on Signals, Systems, and Computers, pages 1066–1069. IEEE, 2020.

[8] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. arXiv preprint arXiv:2010.01243, 2020.

[9] Pierluigi Coppola and Fulvio Silvestri. Autonomous vehicles and future mobility solutions. In Autonomous vehicles and future mobility, pages 1–15. Elsevier, 2019.

[10] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In International conference on machine learning, pages 933–941. PMLR, 2017.

[11] Douglas Everson, Ashish Bastola, Rajat Mittal, Siddheshwar Munde, and Long Cheng. A comparative study of log4sheil test tools. In 2022 IEEE Secure Development Conference (SecDev), pages 16–22. IEEE, 2022.

[12] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. Advances in neural information processing systems, 27, 2014.

[13] Yiling Gong, Chong Wang, Xinmiao Dai, Shenghao Yu, Lehong Xiang, and Jiafei Wu. Multi-scale continuity-aware refinement network for weakly supervised video anomaly detection. In 2022 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2022.

[14] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1):100–108, 1979.

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In International conference on machine learning, pages 2127–2136. PMLR, 2018.

[16] Hadi Jamali-Rad, Mohammad Abdizadeh, and Anuj Singh. Federated learning with taskonomy for non-iid data. IEEE transactions on neural networks and learning systems, 2022.

[17] Ammar Mansoor Kamoona, Amirali Khodadadian Gostar, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications, 214:119079, 2023.

[18] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.

[19] Alex Kulesza and Ben Taskar. Learning determinantal point processes. 2011.

[20] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In 2022 IEEE 38th International Conference on Data Engineering (ICDE), pages 965–978. IEEE, 2022.

[21] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 2021.

[22] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189, 2019.

[23] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In Proceedings of the 28th ACM international conference on multimedia, pages 1864–1872, 2020.

[24] Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. Medical image analysis, 76:102298, 2022.

[25] WANG Luping, WANG Wei, and LI Bo. Cmfl: Mitigating communication overhead for federated learning. In 2019 IEEE 39th international conference on distributed computing systems (ICDCS), pages 954–964. IEEE, 2019.

[26] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8022–8031, 2023.

[27] Cristina Olaverri-Monreal. Autonomous vehicles and smart mobility related technologies. Infocommunications Journal, 8(2):17–24, 2016.

[28] Peijie Qiu, Pan Xiao, Wenhui Zhu, Yalin Wang, and Aristeidis Sotiras. Sc-mil: Sparsely coded multiple instance learning for whole slide image classification. arXiv preprint arXiv:2311.00048, 2023.

[29] Abolfazl Razi, Xiwen Chen, Huayu Li, Hao Wang, Brendan Russo, Yan Chen, and Hongbin Yu. Deep learning serves traffic safety analysis: A forward-looking review. IET Intelligent Transport Systems, 17(1):22–71, 2023.

[30] Daniel J Solove and Paul M Schwartz. Information privacy law. Aspen Publishing, 2020.

[31] Chen Wang, Yulu Dai, Wei Zhou, Yifei Geng, et al. A vision-based video crash detection framework for mixed traffic flow environment considering low-visibility condition. Journal of advanced transportation, 2020, 2020.

[32] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (CSUR), 54(6):1–36, 2021.

[33] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. Advances in neural information processing systems, 32, 2019.

[34] Yuxuan Zhang, Chao Xu, Howard H Yang, Xijun Wang, and Tony QS Quek. Dpp-based client selection for federated learning with non-iid data. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[35] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. Neurocomputing, 465:371–390, 2021.

[36] Wenhui Zhu, Peijie Qiu, Oana M Dumitrascu, and Yalin Wang. Pdl: Regularizing multiple instance learning with progressive dropout layers. arXiv preprint arXiv:2308.10112, 2023.

[37] Wenhui Zhu, Peijie Qiu, Natasha Lepore, Oana M Dumitrascu, and Yalin Wang. Self-supervised equivariant regularization reconciles multiple-instance learning: Joint referable diabetic retinopathy classification and lesion segmentation. In 18th International Symposium on Medical Information Processing and Analysis, volume 12567, pages 100–107. SPIE, 2023.