Balancing Privacy and Accuracy in Federated Learning for Speech Emotion Recognition

Samaneh Mohammadi[1,3], Mohammadreza Mohammadi[1,2], Sima Sinaei[1], Ali Balador[3], Ehsan Nowroozi[4],
Francesco Flammini[3], and Mauro Conti[2]
[1] RISE Research Institutes of Sweden, Västerås, Sweden.
Email: {samaneh.mohammadi, mohammadreza.mohammadi, sima.sinaei}@ri.se
[2] University of Padua, Padua, Italy. Email: mauro.conti@unipd.it
[3] Mälardalen University, Västerås, Sweden. Email: {ali.balador, francesco.flammini}@mdu.se
[4] Queen's University Belfast, Centre of Secure Information Technologies, Belfast, Northern Ireland,
United Kingdom. Email: e.nowroozi@qub.ac.uk

Abstract—*Context:* **Speech Emotion Recognition (SER) is a valuable technology that identifies human emotions from spoken language, enabling the development of context-aware and personalized intelligent systems. To protect user privacy, Federated Learning (FL) has been introduced, enabling local training of models on user devices. However, FL raises concerns about the potential exposure of sensitive information from local model parameters, which is especially critical in applications like SER that involve personal voice data. Local Differential Privacy (LDP) has prevented privacy leaks in image and video data. However, it encounters notable accuracy degradation when applied to speech data, especially in the presence of high noise levels. In this paper, we propose an approach called LDP-FL with CSS, which combines LDP with a novel client selection strategy (CSS). By leveraging CSS, we aim to improve the representatives of updates and mitigate the adverse effects of noise on SER accuracy while ensuring client privacy through LDP. Furthermore, we conducted model inversion attacks to evaluate the robustness of LDP-FL in preserving privacy. These attacks involved an adversary attempting to reconstruct individuals' voice samples using the output labels provided by the SER model. The evaluation results reveal that LDP-FL with CSS achieved an accuracy of 65-70%, which is 4% lower than the initial SER model accuracy. Furthermore, LDP-FL demonstrated exceptional resilience against model inversion attacks, outperforming the non-LDP method by a factor of 10. Overall, our analysis emphasizes the importance of achieving a balance between privacy and accuracy in accordance with the requirements of the SER application.**

Index Terms—**Federated Learning, Privacy-preserving Mechanism, Differential Privacy, Speech Emotion Recognition**

I. Introduction

Speech Emotion Recognition (SER) is a cutting-edge technology that detects and interprets emotions conveyed through spoken language [1]. Its potential impact spans multiple sectors, including customer service, mental health, education, and entertainment [2], [3]. However, the traditional centralized SER model, which involves gathering user speech data and training a single model on a central server, poses privacy risks. Analyzing speech data can expose sensitive information like biometric identity, personality traits, location, emotional state, age, gender, and overall health [4]. To address these ethical and privacy concerns, regulation like the General Data Protection Regulation (GDPR) [5] has been implemented to safeguard personal data. Privacy must be a top priority when developing and implementing SER applications across various domains.

Federated learning (FL) has emerged as a promising solution to privacy concerns in various fields and applications [6]. FL maintains local data on end devices and trains ML models on local client devices without transferring raw data to a central server. This preserves data privacy and ensures compliance with regulations such as GDPR. For SER applications, the initial processing of speech data and training perform on clients' devices, and only local model parameters are sent to the central server for model aggregation [7].

However, FL faces new privacy concerns when it comes to transmitting local model parameters between clients and servers [8]. This is a significant concern because the transmission parameters can be exploited by third parties, enabling them to launch attacks that reconstruct raw speech data or features, thereby revealing sensitive information [9]. To address this issue, additional privacy mechanisms have been proposed together with FL to safeguard such applications.

One promising mechanism that has emerged is differential privacy (DP), which offers a potential solution to protect individual data points in certain cases. DP can be implemented in FL through two forms: local differential privacy (LDP) applied on the client side and global differential privacy (GDP) used on the server side [10]. LDP, a well-established technique, involves the addition of carefully calibrated noise to each client's model parameters before transmitting them to the central server [11], [12], [13].

Integrating LDP in FL for SER applications offers several benefits. It ensures the privacy and confidentiality of individual speech data, protecting sensitive information from unauthorized access and potential attacks. By introducing noise to the model parameters, LDP prevents the reconstruction of raw speech data or features by malicious third parties, thereby preserving users' privacy. Despite the promising potential of LDP in FL for SER applications, there is a noticeable lack of research published in reputable conferences or journals specifically addressing the utilization of LDP in this context.

This gap highlights the need for further investigation and exploration to fully understand the effectiveness and practical implications of integrating LDP into FL for SER.

However, when applied to SER applications, LDP does not offer acceptable accuracy due to the adverse effects of adding noise to voice data, which can distort the audio signal [14]. Furthermore, adding noise to SER model parameters can affect the model's utility by distorting or misaligning the parameters, leading to errors in the model's output [15]. This compromise in accuracy is especially detrimental to most SER applications, which rely on precise results for industrial use [16]. Therefore, when developing LDP mechanisms in FL for SER, it is necessary to find a concrete solution that effectively mitigates the impact of noise on SER accuracy while still maintaining robust privacy protections.

This paper proposes a method, referred to as LDP-FL with CSS, which integrates LDP with a novel client selection strategy (CSS) to enhance privacy while preserving the acceptable accuracy of SER in the FL system. LDP is utilized to protect clients' speech datasets, while CSS is employed to minimize the negative impact of noise scaling on the model updates, resulting in more representative updates and improved accuracy.

Moreover, our study focuses on adapting the model inversion attack, initially developed for facial recognition models [17], for the SER model through appropriate configuration adjustments. This attack attempts to reconstruct speech features by considering the adversary's knowledge of a particular client's emotion label and their local SER model. The primary goal is to evaluate the effectiveness of the LDP method in safeguarding against such attacks within the FL setup. Finally, we comprehensively evaluated the LDP-FL with CSS approach, specifically focusing on assessing its alignment with SER requirements and analyzing the trade-off between accuracy and privacy.

The novel contributions of this paper can be summarized as follows:

- We introduce a novel approach that combines local differential privacy in federated learning (LDP-FL) with a client selection strategy (CSS) to enhance privacy while mitigating the impact of noise on SER accuracy.
- We implement model inversion attacks to assess the robustness of LDP-FL and determine its effectiveness in preserving privacy. These attacks involve an adversary attempt to reconstruct individuals' voice samples based on the output labels provided by the SER model.
- We conduct a comprehensive evaluation of the LDP-FL with CSS approach on public SER datasets, considering important parameters such as privacy budget, noise scale, failure probability, and clipping threshold value. Our evaluation focuses on assessing how well our method meets SER requirements and analyzing the balance between accuracy and privacy.

The rest of the paper is structured as follows. Section II covers background and related work on privacy-preserving FL and SER. A reference system description is provided in Section III, including the SER non-functional requirements, threat model, the proposed method, and implementation of the model inversion attack. Section IV presents the experimental results obtained using the proposed approach. Lastly, Section V concludes the paper and provides insights for future developments.

II. Background and Related Works

In this section, we will provide an overview of the background and related work on LDP mechanisms in FL. We will then discuss the use of FL for SER applications and its related work.

A. *Privacy-preserving Federated Learning*

FL protects user privacy by decentralizing data from the central server to edge devices; however, sharing information with servers (e.g., model weights) can pose privacy threats [8]. Since FL requires central servers and clients to exchange model update parameters, attackers with white-box access obtain the model, its architecture, weight parameters, and any hyperparameters needed for predictions. When using black-box scenarios, the adversary can observe only the outputs of the model on arbitrary inputs [9].

LDP has become an increasingly popular technique for privacy-preserving in FL [10]. LDP can prevent individual devices' data from being leaked to the central server during the model training process [11], [12]. This technique involves adding artificial noise to each model's updated parameters before sharing it with the central server. Recent work proposed a framework called NbAFL that utilized LDP and demonstrated its capability to meet DP requirements under different protection levels by appropriately adapting various variances of artificial noise [12]. Another study proposed LDP-based stochastic gradient descent (SGD) that guarantees a given LDP level by providing a noise variance limit after multiple rounds of weight updates using a tight composition theorem [13].

B. *Speech Emotion Recognition using Federated Learning*

SER technology aims to recognize and understand human emotions through speech. SER systems analyze the audio signals from human speech and use ML algorithms to detect patterns and classify the emotional states conveyed by the speech [2]. Building SER models requires significant amounts of data, including sensitive personal information such as speech signals and emotions. However, centralized storage of this data presents privacy risks. To mitigate these risks, FL is a promising solution that allows models to be trained collaboratively on decentralized devices without the need to transfer raw data [7].

The paper [18] introduces an FL-based approach for building a private decentralized SER model. The proposed method utilizes data-efficient federated self-training to train SER models with minimal on-device labelled samples. However, the proposed method only relies on the FL framework as a privacy-preserving technique and does not consider any threat models from clients or servers in FL, nor does it

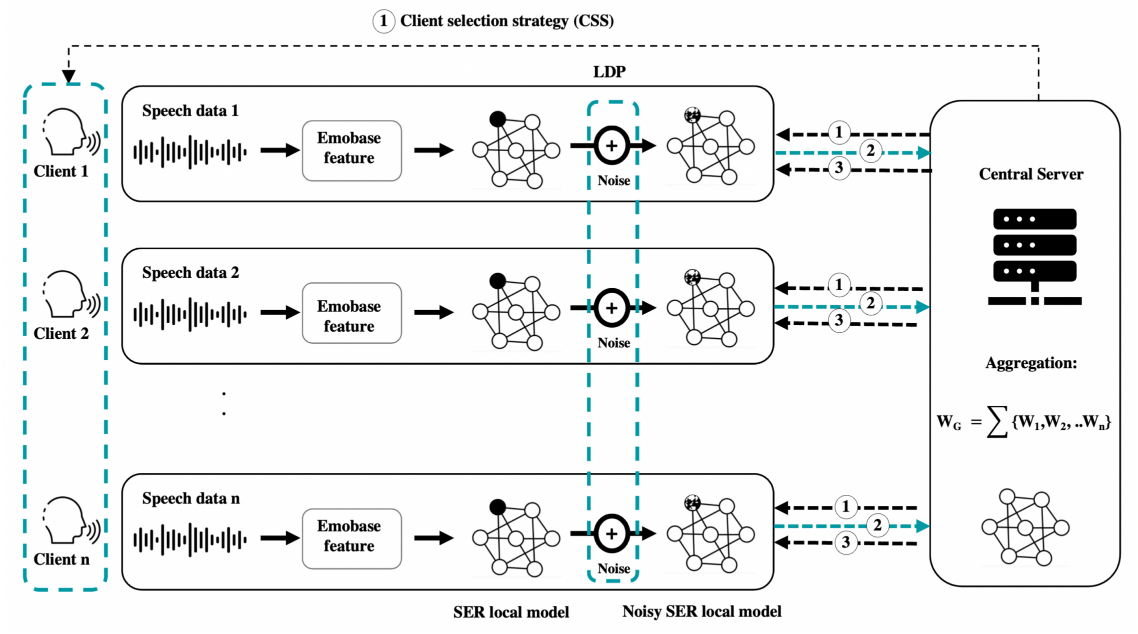


Fig. 1: An overall overview of LDP-FL with CSS for SER application.

consider any other privacy-preserving techniques. Similarly, another work [19] proposes a federated adversarial learning framework to protect both data and deep neural networks in SER. The framework comprises an FL framework for data privacy and adversarial training during the training stage for model robustness. However, like the previous method, it only relies on the FL framework for privacy preservation and does not consider other privacy-preserving techniques in FL.

III. System Description

In this section, we will address the non-functional requirements of the SER application (III-A), discuss the associated threat model (III-B), present the proposed LDP-FL with CSS method (III-C) along with algorithms and details, and finally describe the model inversion attack for speech features using algorithms (III-D).

A. *Non-functional Requirements of Speech Emotion Recognition Application*

Non-functional requirements refer to the characteristics or qualities of a system that are related to its performance rather than its specific functionality. In the context of SER applications, important non-functional requirements include privacy and accuracy. Satisfying these requirements is critical to ensure user needs and expectations while complying with legal requirements. In this part, we provide a more detailed explanation and explain how we meet these requirements in the evaluation section IV.

1) *Privacy:*
   a) Personal speech data must be kept on local devices only [5].
   b) The central server or a potential eavesdropper must be not able to infer sensitive information from the local model parameters
2) *Accuracy:*
   a) The level of accuracy of SER applications must be kept high enough to identify the correct emotions from speech samples reliably. We can consider a baseline accuracy of a minimum 70% in detecting the four basic emotions - neutral, sad, happy, and angry [20].

It is important to highlight that those requirements can be highly interdependent. For instance, privacy-preserving approaches can have an impact on accuracy due to e.g. the distributed setup the usage of FL, applied noise of the LDP method, etc. Additionally, when implementing a SER in an FL setup, it has been demonstrated in reference [18] that there is a potential for a 0-5% accuracy drop.

B. *Threat Model*

In this paper, we assume the server follows the honest-but-curious (HBC) paradigm. Under this paradigm, the server is not malicious and adheres to the FL protocol, but may still possess a curiosity about the data or models of other clients [21]. While the individual client datasets are kept locally in FL, the intermediate parameter w_i needs to be shared with the server, which can potentially expose clients' private information, as evidenced by model inversion attacks. For instance, in [17], researchers demonstrated a model inversion attack capable of reconstructing images from a facial recognition system.

C. *Proposed Method: LDP-FL with CSS*

We present "LDP-FL With CSS," a novel approach that combines local differential privacy (LDP) and a client selection strategy (CSS) in federated learning to balance client privacy and the accuracy of the SER model. Our method addresses privacy requirement 1.a by processing and training clients' speech data locally on their devices within the FL setup. By leveraging LDP techniques, Gaussian noise is incorporated into the local updates before transmitting them to the central server. This implementation provides robust protection against the inference of sensitive information, thereby fulfilling the requirement of 1.b and mitigating potential risks in the threat model. To meet the accuracy requirements 2.a, we incorporate CSS, prioritizing clients with larger data pools and involving them in each training round. This strategy aims to mitigate the potential negative impact of LDP on accuracy while enhancing it to meet the desired accuracy levels.

Figure 1 outlines the proposed method, which consists of three main steps. Firstly, the server broadcasts the initialized global SER model and uses the CSS method to select clients for training. In the second step, the chosen clients analyze speech data, extracting Emobase features (explained in Sec. IV-A) and train their models. They also update their local model parameters using the global model. Privacy is ensured by applying the LDP method to each parameter before sharing them with the server. The clients then share the noisy model parameters with the server. In the third step, the server aggregates the received noisy local model parameters and returns the updated global model to the clients. We provide Algorithm 1 as a comprehensive outline of the LDP-FL with CSS approach. Subsequently, we will delve deeper into the concepts of LDP and CSS in the context of FL.

Algorithm 1: LDP-FL with CSS

```
Input: Number of iterations: T, Number of selected
       clients: K, Local minibatch size: B, Initial
       global model: $w_0$, Learning rate: $\eta$, Clipping
       threshold: C, LDP parameters: $\epsilon$ and $\delta$
1  Initialization:
2  Initialize the global model parameters $w_0$
   for $t \leq T$ do
3      The server broadcasts current model $w_t$
4      K: Client Selection Strategy (CSS)
5      Clients-side:
6      for $i \in 1, 2, ..., K$ do
7          for each batch $b \in B_i$ do
8              Compute gradient g(b) $\leftarrow \nabla_w L^i(w_t; b)$
9          Clip gradient $\overline{g}(b) \leftarrow g(b)/Max(1, \frac{\|g(b)\|}{C})$
10         Add Noise
           $\tilde{g}_i = \frac{1}{|B|}(\sum_{b \in B} \overline{g}(b) + N(0, \sigma^2 C^2 I)$
11         Share $\tilde{g}_i$ with server
12     Server-side:
13     Aggregate $\tilde{g} = \frac{1}{K}\sum_{i=1}^{K} \tilde{g}_i$
14     Global model update $w_{t+1} \leftarrow w_t - \eta.\tilde{g}$
```

1) Local Differential Privacy (LDP): LDP is defined under the setting where the user does not trust anyone (not even the central data collector) [11]. In this setting, users themselves apply a random perturbation to protect their privacy. Each user runs a random perturbation algorithm, denoted as M, on their data and shares the perturbed results with the aggregator or central server. In LDP, the privacy budget, denoted as ϵ, represents the amount of privacy protection desired, with a higher value of ϵ implying a lower level of privacy. While δ represents the probability that an LDP mechanism fails to provide the specified privacy guarantee. Here is a formal definition of LDP:

Definition 1 ((ϵ, δ)-LDP [22]): A randomized mechanism M satisfies (ϵ, δ)-LDP if and only if for any pairs of input values v and v' in the domain of M, and for any possible output $y \in$ S , it holds:

$$Pr[M(v) = y] \leq e^{\epsilon} Pr[M(v') = y] + \delta. \tag{1}$$

Theoretically, (ϵ, δ)-LDP means that a mechanism M achieves (ϵ, δ)-LDP with probability at least $1 - \delta$.

To implement the LDP mechanism in a FL setup, we followed the approach described in reference [23]. Specifically, we incorporated artificial Gaussian noise into the clients' model parameters. In order to ensure that the given noise distribution $Z \sim N(0, \sigma^2 C^2 I)$ preserves (ϵ, δ)-LDP, for any $\epsilon < cq^2 T$, $\delta > 0$, and T number of epoch, we choose noise scale $\sigma \geq c\frac{q\sqrt{Tlog(1/\delta)}}{\epsilon}$, where the constant c and sampling probability q. In this result, Z is the value of an additive noise for client gradient.

In Algorithm 1, during time slot t, each selected client $i \in k$ trains its local dataset by minimizing the loss function ∇L^i (lines 6-8). For each client, the gradient $g(b)$ is calculated for each $b \in B_i$. To limit the impact of each gradient $g(b)$, we apply clipping using the $\|L\|_2$ norm. Specifically, $g(b)$ is replaced by $g(b)/\max(1, \frac{\|g(b)\|_2}{C})$ where C is the clipping threshold (line 7). This clipping mechanism ensures that if the norm $\|g\|_2$ is less than or equal to C, the gradient g remains unchanged. However, if $\|g\|_2$ exceeds C, it is scaled down to have a norm of C, thereby controlling the contribution of large gradients.

After clipping the gradient, we compute the average of all gradients in set B and add a scaled Gaussian noise $Z \sim N(0, \sigma^2 C^2 I)$ to each client's gradient to achieve LDP in lines 9-10. The resulting noisy gradient $\tilde{g}_i$ is then shared with the server in line 11. On the server side, upon receiving the noisy gradients $\tilde{g}_i$ from the selected clients, the server performs the FedSGD algorithm by aggregating the gradients $\tilde{g} = \frac{1}{K}\sum_{i=1}^{K} \tilde{g}_i$. Subsequently, the global model is updated using $W_{t+1} \leftarrow W_t - \eta \cdot \tilde{g}$ and utilized for the next iteration in lines 13-14.

2) Clients Selection Strategy (CSS): To mitigate potential noise effects and uphold the initial accuracy of SER models in FL, we introduce a refined client selection strategy named

(CSS). Our proposed approach involves carefully selecting clients for FL training, employing two distinct criteria.

```
Algorithm 2: Clients Selection Strategy (CSS)
Input: Number of iterations: T, Clients list: L,
       Number of selected Clients: K
Output: List of selected clients
1 for $t \leq T$ do
2     Half of selection: M = K / 2
3     $\mathcal{C}$ = sorted L in descending order by sample size
4     Selected clients = $\mathcal{C}$ [:M]
5     Remaining clients = randomly select $\mathcal{C}$ [M:]
6 Return K = selected clients + remaining clients
```

Firstly, we select half of the clients from a larger pool of candidates based on their sample size. This criterion ensures that clients with larger local datasets are given preference. By incorporating larger local datasets, which are more likely to yield accurate and representative model updates, we aim to enhance the overall model accuracy. Secondly, to mitigate selection bias, the remaining half of the clients are randomly chosen. This random selection mechanism introduces an element of diversity and reduces the potential bias that could arise from selecting clients based solely on their sample size.

Algorithm 2 outlines the client selection strategy (CSS) method used in each training round of the overall Algorithm 1. Our method selects the top half of the clients based on their sample size, giving those clients with the largest sample sizes a higher probability of being chosen for each round of training (line 4). To reduce bias in client selection, we combine our proposed method with random selection for the remaining clients (line 5). We then combine the two sets of selected clients to obtain the final selection (line 6).

We ensure that there is no overlap between the two sets of selected clients to guarantee that each client is selected precisely once per training round. By employing the CSS approach, we strike a balance between leveraging large local datasets for training and maintaining diversity within the FL system. This methodology effectively minimizes noise effects and fosters the preservation of the initial model accuracy in SER models trained through FL.

D. Model Inversion Attack for Speech Emotion Recognition Models

A model inversion attack takes place when an adversary gains access to a model's output and potentially its parameters, aiming to infer sensitive training data. In our paper, we adjust the existing work conducted in the field of face recognition [17] and adapt it for speech emotion recognition by changing some configurations. In this scenario, we assume that the attacker possesses knowledge of a single emotion label, such as neutral, sad, happy, or angry, as well as the model used by the clients. The objective of the adversary is to reconstruct the speech data features associated with a specific client and the corresponding emotion label.

The target of model inversion attack in this case is the inversion of speech features, which represent high-level statistical characteristics of a client's speech. Each intensity value in the features corresponds to a floating-point value. In our attack scenarios, we assume that the attacker does not possess exact knowledge of the feature values they are trying to infer. We consider feature vectors with n components and four emotion label classes, and we model each emotion recognition classifier as a function $\tilde{f} : [0,1]^n \rightarrow [0,1]^4$. The output of the model is a probability vector, where each component represents the probability that the feature vector belongs to a specific emotion label. We use the notation $\tilde{f}_{\text{label}}(x)$ to refer to the ith component of the output corresponding to the emotion label. The Algorithm 3 provides a comprehensive outline of the model inversion attack specifically designed for speech emotion recognition models.

```
Algorithm 3: Model inversion attack for speech emotion recognition models
Input: Number of iteration: $T$, Best score: $\gamma$, Target
       model: $\tilde{f}$, Learning rate: $\eta$
Output: Related speech features to target label
$c = 1 - \tilde{f}_{label}(x)$
$x_0 \leftarrow 0$
for $t \leq T$ do
    $x_t \leftarrow Process(x_{t-1} - \eta \cdot \nabla c(x_{t-1}))$
    if $c(x_t) \geq \max(c(x_{t-1}), \ldots, c(x_{t-\beta}))$ then
        break
    if $c(x_t) \leq \gamma$ then
        break
return $[\arg\min_{x_t}(c(x_t)), \min_{x_t}(c(x_i))]$
```

The algorithm utilizes gradient descent to minimize a cost function involving the emotion recognition model $\tilde{f}$ for model inversion. Gradient descent iteratively updates a candidate solution by moving towards the negative gradient direction. The cost function, denoted as c, is defined based on $\tilde{f}$. The model inversion attacks employ gradient descent for a maximum of T iterations with a step size of η. After each iteration, the resulting feature vector is processed using a post-processing function called Process, which can apply various manipulations to the speech features, such as denoising and sharpening, depending on the specific attack. The descent terminates if the cost does not improve within β iterations or if the cost exceeds a threshold γ. In such cases, the best candidate is returned as a result.

IV. Experiment Results

In this section, we present an industrial use case and the simulation setting. We evaluate the impact of the LDP-FL method on SER accuracy, considering parameters like noise scale, failure probability, and clipping threshold. We analyze the effect of CSS on SER accuracy within the LDP-FL framework and investigate the robustness of LDP-FL against model inversion attacks. Finally, we discuss the crucial task of achieving the optimal balance between privacy levels and accuracy.

A. Usecase Description and Simulation Setting

DAIS[1] (Distributed Artificial Intelligent System) [24] is a pan-European project that aims to provide trustworthy connectivity and interoperability by combining the IoT with AI into a distributed edge system for industrial applications. The project includes industry-driven use cases in domains such as digital life, digital industry, and smart mobility. One of the important use cases in DAIS is SER which is deployed on TV recommendation systems. The goal is to accurately capture users' emotions and provide personalized movie recommendations, leading to higher levels of user satisfaction. Achieving this requires a distributed, efficient, private, and accurate SER application. This was one of the main motivations for exploring the potential of LDP-FL with CSS in SER.

As part of this study, we evaluated the proposed method on one of the most widely used SER datasets, namely CREMA-D [25]. CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from various races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified). To train the SER model, we chose the four most commonly occurring emotion labels (neutral, sad, happy, and angry) based on the possible emotions expressed in the sentences.

For speech processing and feature extraction, we generate the Emo-Base feature set using the OpenSMILE toolkit [26]. The Emo-Base feature set is a widely used set of features for SER tasks. These features are extracted from the speech signal and capture various acoustic characteristics of the signal that are associated with different emotions. The features are designed to be highly discriminative for emotion recognition and have been shown to achieve state-of-the-art performance in various SER tasks. After extracting the features, we utilized a multilayer perceptron (MLP) architecture for the SER model and trained it using the FedSGD algorithms. The model consists of two dense layers with layer sizes of [256, 128] and ReLU activation function, along with a 0.2 dropout rate. We set a local training batch size of 20 and a learning rate of 0.1 to accelerate convergence in the FedSGD algorithm.

For the FL training on the CREMA-D dataset, each speaker serves as a unique client since there are 91 distinct speakers in the dataset. We employed 80% of the data for local training at each client and reserved the remaining 20% for validation. To ensure the robustness of our approach, we conducted five experiments with different test folds, and we reported the average results of the five-fold experiments. The FL scenarios were conducted over 200 global training epochs. Our experiments were conducted on a Windows 10 Pro environment, featuring an Intel(R) Core(TM) i7 CPU @1.80GHz 1.99 GHz processor and 16.0 GB of RAM.

B. SER accuracy results across different parameters: noise scale, failure probability and clipping threshold

We conducted an analysis to assess the accuracy of SER in LDP-FL by examining the impact of various LDP parameters on accuracy. Our evaluation involved 50 randomly selected clients and 120 training epochs, as depicted in Figure 2. Specifically, we investigated the influence of the noise scale σ on accuracy (Figure 2(a)), the effect of varying failure probability δ on accuracy (Figure 2(b)), and the impact of the clipping threshold C on accuracy (Figure 2(c)).

The experimental results illustrated in Figure 2(a) indicate that the accuracy of LDP-FL gradually stabilizes with an increase in the number of training epochs, indicating convergence of the method. However, higher noise scales, such as $\sigma = 10$, can impede convergence due to the injection of larger amounts of noise during training, resulting in an unstable system. Figure 2(b) demonstrates that higher failure probabilities δ lead to faster convergence and higher accuracy but weaker privacy protection. Conversely, lower failure probabilities provide stronger privacy guarantees at the expense of reduced accuracy. For instance, a failure probability of $\delta = 10^{-3}$ achieved the highest accuracy while sacrificing some privacy for utility.

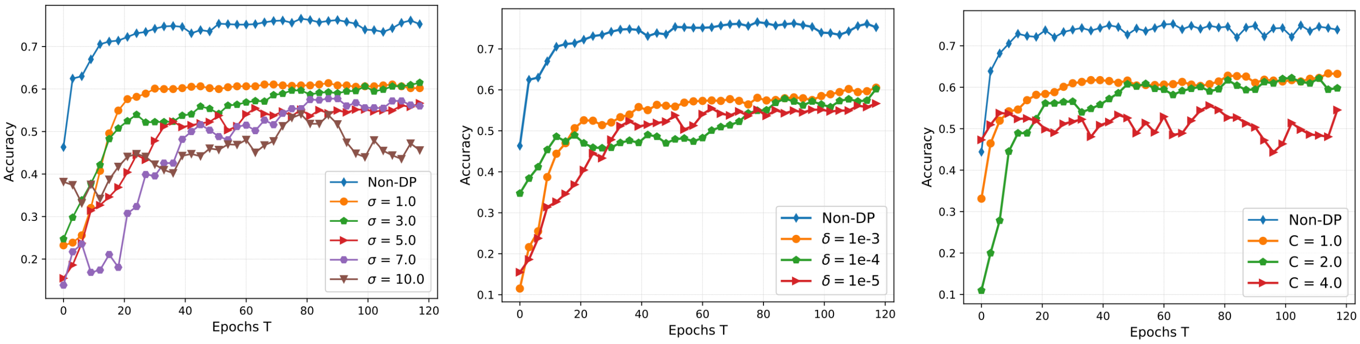
Our evaluation of LDP-FL's accuracy with different clipping thresholds showed that a threshold of 1.0 or 2.0 achieves high accuracy with fast convergence, as shown in Fig. 2(c). However, using a threshold beyond a certain point results in decreased accuracy due to excessive information loss during the clipping process. Hence, selecting the optimal clipping threshold is crucial to balance privacy preservation and model accuracy.

C. Effect of CSS on SER accuracy

To evaluate the effectiveness of LDP-FL with CSS for SER application, we conducted a comparative study between CSS and the commonly used random selection (RS) method in both LDP and non-LDP FL systems. Using parameters $\sigma = 1.0$, $C = 2$, $\delta = 10^{-5}$, and $K = 50$, we observed a significant improvement in accuracy from 60% to 70% when using CSS with LDP, as depicted in Figure 3 and meeting the accuracy requirements outlined in Section III-A. CSS proved to be an effective method for selecting clients, leading to more representative and larger datasets for training, resulting in more robust and accurate models. However, it is important to note that selecting clients with larger local datasets increases their exposure, potentially leading to data leakage. Therefore, a balance must be struck when employing CSS.

Interestingly, we observed that the choice of client selection method did not significantly impact the accuracy of non-LDP FL systems. This suggests that the accuracy improvement achieved by CSS is specific to the LDP-FL. Thus, adopting an

[1] https://dais-project.eu/



(a)

(b)

(c)

Fig. 2: Evaluation of the accuracy of the SER model using LDP-FL across different parameters: (a) noise scale σ, (b) failure probability δ, and (c) clipping threshold C.

TABLE I: MSE of reconstruction of speech features by model inversion attack in cases where FL has LDP or does not have LDP

Target model	Clipping Threshold (C)	Mean Squared Error (MSE)					
		LDP-FL					Non-LDP-FL
		$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	$\sigma = 7$	$\sigma = 10$	-
Client SER model	C=1	0.971	1.189	19.830	45.132	157.640	1.02
	C=2	1.028	9.189	78.910	1139.474	5807.308	1.02
	C=4	1.886	344.000	8508.620	59164.757	572765.191	1.02

efficient client selection strategy like CSS can be a valuable technique to enhance the performance of LDP-FL and mitigate the potential negative impact of LDP on accuracy.

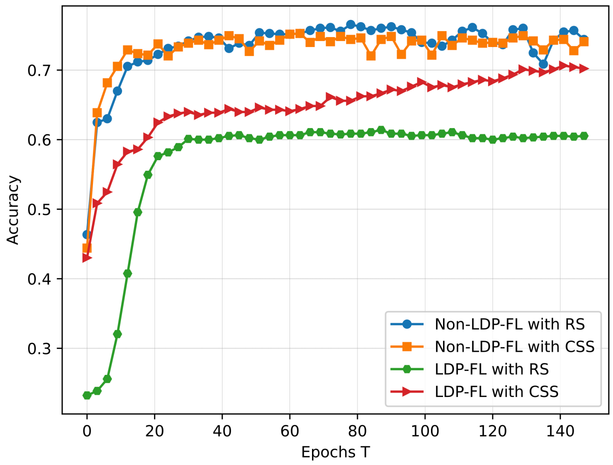


Fig. 3: Evaluation of different client selection methods based on accuracy.

D. Analyze the robustness of LDP-FL against model inversion attacks

We conducted a model inversion attack using the specified algorithms with the following settings: $T = 200$, $\eta = 0.1$, $\beta = 100$, and $\gamma = 0.99$. The attack was applied to two different settings in the system: LDP-FL with $C = [1, 2, 4]$ and $\delta = 10^{-5}$, and non-LDP in FL with $K = 7$. To ensure accurate results, we performed the attack on various client target models and target labels and reported the average outcomes.

The objective of model inversion attacks is to reconstruct the speech features of each client by exploiting the local SER model and its associated labels. To evaluate the effectiveness of these attacks on the FL system, we employed the Mean Squared Error (MSE) metric. The MSE was calculated by comparing the reconstructed speech features with the actual speech features of each specific client.

Table I illustrates the results obtained from the attack. When the noise scale σ was set to 1.0, the MSE values were similar for both LDP and non-LDP settings. However, as we increased the noise scale σ and the clipping threshold C, the MSE values significantly increased, indicating a decline in the attack effectiveness. These findings highlight the effectiveness of incorporating LDP as a robust privacy measure against model inversion attacks. Implementing LDP in the system significantly mitigates the risk posed by threat models and ensures compliance with the specified privacy requirements, particularly the 1.b privacy requirement. By introducing noise into the client models, the accuracy of predictions made by adversaries using these models is reduced, thereby impeding the reconstruction process of speech features associated with specific client labels.

E. Balancing privacy and accuracy

Achieving an optimal balance between privacy and accuracy is paramount when utilizing LDP for SER applications that require precise and accurate results. According to this reference [23], epsilon (ϵ) acts as a parameter that measures the level of privacy guarantee provided by the $(\epsilon, \delta) - LDP$

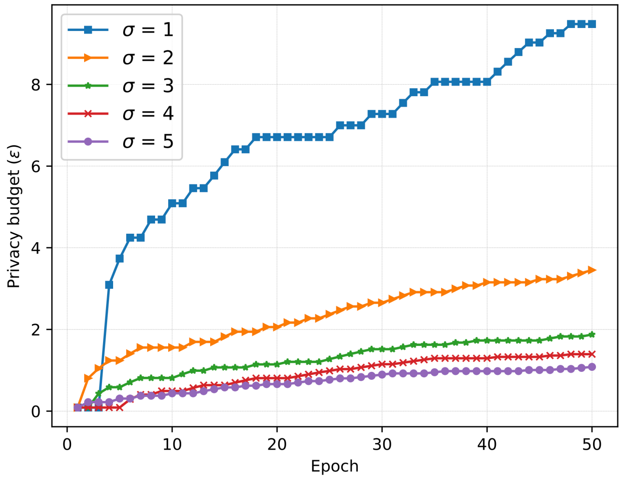


Fig. 4: With constant noise scale σ, changes in privacy budget (ϵ) with an increase in epochs

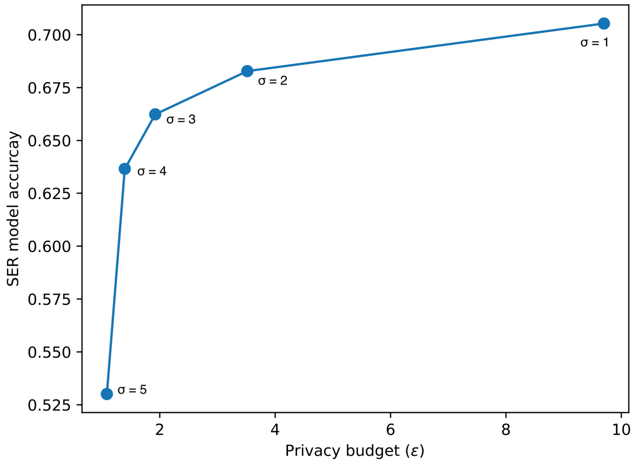


Fig. 5: An assessment of the accuracy of SER models based on privacy budgets and noise levels.

mechanism. It reflects the degree of privacy protection, with smaller epsilon values indicating stronger privacy guarantees. In our developed method, the value of ϵ is influenced by the number of epochs (T). Consequently, as the number of epochs increases, the value of ϵ changes, even when the noise scale remains constant. This association is illustrated in Figure 4.

In our evaluation of the LDP-FL with CSS mechanism for SER, we experimented with different noise scales σ, $k = 50$, failure probability parameter of $\delta = 10^{-5}$, clipping threshold $C = 2$ and a total of 50 epochs (T). The results, as illustrated in Figure 5 and Figure 4, revealed the following privacy levels and corresponding accuracy:

- For a noise scale of $\sigma = 5$, we achieved a privacy level of $(1.08, 10^{-5}) - LDP$, with an accuracy of approximately 54%.
- With a noise scale of $\sigma = 4$, we attained a privacy level of $(1.39, 10^{-5}) - LDP$, accompanied by an accuracy of around 64%.
- Employing a noise scale of $\sigma = 3$, we achieved a privacy level of $(1.92, 10^{-5}) - LDP$, while maintaining an accuracy of approximately 67%.
- A privacy level of $(3.51, 10^{-5}) - LDP$ was obtained by utilizing a noise scale of $\sigma = 2$, resulting in an accuracy of about 69%.
- Finally, with a noise scale of $\sigma = 1$, we achieved a privacy level of $(9.69, 10^{-5}) - LDP$, accompanied by an accuracy of roughly 70%.

Striking the right balance between privacy and accuracy is contingent upon specific system requirements. In the case of the SER application in the FL setup, where the specified acceptable accuracy range is 65-70% and privacy requirements are outlined in Section III-A, it is feasible to attain an acceptable level of privacy by utilizing a privacy parameter of $(1.92, 10^{-5})$-LDP, along with a noise scale of $\sigma = 3$, while maintaining the desired accuracy.

V. Conclusions and Future work

In this paper, we introduced LDP-FL with CSS, a novel approach for privacy-sensitive SER applications. Our objective is to ensure the privacy of clients' speech data while maintaining system accuracy. By combining LDP-FL and CSS, we mitigate the impact of noise scale on accuracy and improve it by selectively choosing clients based on their data size during each training round of FL. We evaluated our approach using the CREMA-D dataset. The evaluation results demonstrate that LDP-FL with CSS achieved an accuracy range of 65-70%, slightly lower than the initial SER model accuracy while maintaining a privacy level of $(1.92, 10^{-5})$-LDP. Our analysis highlights the importance of achieving a balance between privacy and accuracy, which aligns with the specific requirements of SER applications.

In the future, we plan to discuss personalized privacy with an adaptive noise scale of LDP mechanisms that are tailored to each client's privacy preference.

VI. Acknowledgements and disclaimer

This work was partially supported by EU ECSEL project DAIS that has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No.101007273. The work reflects only the authors' views; the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[3] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar, and M. Guizani, "Federated learning meets human emotions: A decentralized framework for human–computer interaction for iot applications," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6949–6962, 2020.

[4] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy implications of voice and speech analysis–information disclosure by inference," *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pp. 242–258, 2020.

[5] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[7] S. Latif, S. Khalifa, R. Rana, and R. Jurdak, "Federated learning for speech emotion recognition applications," in *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2020, pp. 341–342.

[8] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[9] M. S. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Security & Privacy*, vol. 19, no. 2, pp. 20–28, 2020.

[10] Z. Xiong, Z. Cai, D. Takabi, and W. Li, "Privacy threat and defense for federated learning with non-iid data in aiot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1310–1321, 2021.

[11] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam, "Local differential privacy-based federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8836–8853, 2020.

[12] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[13] M. Kim, O. Günlü, and R. F. Schaefer, "Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2650–2654.

[14] M. A. Pathak, *Privacy-preserving machine learning for speech processing*. Springer Science & Business Media, 2012.

[15] T. Feng, R. Peri, and S. Narayanan, "User-level differential privacy

[23] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in

against attribute inference attack of speech emotion recognition in federated learning," *arXiv preprint arXiv:2204.02500*, 2022.

[16] A. A. Alnuaim, M. Zakariah, A. Alhadlaq, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla, and R. Ratna, "Human-computer interaction with detection of speaker emotions using convolution neural networks," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[17] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[18] V. Tsouvalas, T. Ozcelebi, and N. Meratnia, "Privacy-preserving speech emotion recognition through semi-supervised federated learning," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 359–364.

[19] Y. Chang, S. Laridi, Z. Ren, G. Palmer, B. W. Schuller, and M. Fisichella, "Robust federated learning against adversarial attacks for speech emotion recognition," *arXiv preprint arXiv:2203.04696*, 2022.

[20] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems*, vol. 211, p. 106547, 2021.

[21] P. Liu, X. Xu, and W. Wang, "Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives," *Cybersecurity*, vol. 5, no. 1, pp. 1–19, 2022.

[22] R. Bassily, "Linear queries estimation with local differential privacy," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 721–729.
Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 2016, pp. 308–318.

[24] A. Balador, S. Sinaei, M. Pettersson, and I. Kaya, "Dais project - distributed artificial intelligence systems: Objectives and challenges," in *26th Ada-Europe International Conference on Reliable Software Technologies (AEiC'22)*, 2022.

[25] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.