

ProtoFL: Unsupervised Federated Learning via Prototypical Distillation

Hansol Kim^{*1}, Youngjun Kwak^{*12}, Minyoung Jung³, Jinho Shin¹, Youngsung Kim⁴, and Changick Kim^{†2}

¹KakaoBank Corp., South Korea

²Department of Electrical Engineering, KAIST, South Korea

³KETI, Korea Electronics Technology Institute, South Korea

⁴Inha University, South Korea

{hans.kim, vivaan.yjkwak, william.shin}@lab.kakaobank.com, {yjk.kwak, changick}@kaist.ac.kr, minyoung.jung@keti.re.kr, y.kim@inha.ac.kr

Abstract

Federated learning (FL) is a promising approach for enhancing data privacy preservation, particularly for authentication systems. However, limited round communications, scarce representation, and scalability pose significant challenges to its deployment, hindering its full potential. In this paper, we propose ‘ProtoFL’, Prototypical Representation Distillation based unsupervised Federated Learning to enhance the representation power of a global model and reduce round communication costs. Additionally, we introduce a local one-class classifier based on normalizing flows to improve performance with limited data. Our study represents the first investigation of using FL to improve one-class classification performance. We conduct extensive experiments on five widely used benchmarks, namely MNIST, CIFAR-10, CIFAR-100, ImageNet-30, and Keystroke-Dynamics, to demonstrate the superior performance of our proposed framework over previous methods in the literature.

1. Introduction

In recent years, there has been a growing concern about privacy, leading people to hesitate when it comes to uploading their biological data to central data servers. Moreover, companies that possess personal information from users are strictly bound by the General Data Protection Regulation (GDPR) [44]. To address these privacy issues, Federated Learning (FL), an emerging distributed data parallel machine learning approach, has been proposed. FL leverages the decentralized data available on individual clients to collaboratively train a shared global model on a mediator server without the need to share personal data.

^{*}These authors contributed equally to this work

[†]Corresponding author

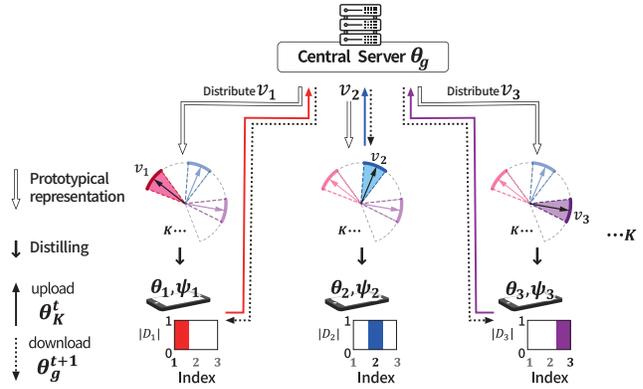


Figure 1: Visualization of **extreme** non-i.i.d. ($\alpha = 0.0$ is a concentration parameter of Dirichlet distribution) data based FL schema shows that each device has only the data from its target class not shared among all joined devices, and prototypical representation is distributed separately to corresponding joined clients in secure from a central server.

In the context of FL, where data is decentralized across individual clients, one-class classification (OCC) can be used. This is because that OCC can determine whether a new example belongs to the target distribution or not. Despite not using data from the non-target class, OCC has shown impressive performance [31, 38, 43, 42, 39, 40, 15, 28]. In computer vision applications, OCC is particularly useful for detecting fraud, defects, and unauthorized users.

Recent advancements in biometric authentication have highlighted the importance of FL-based OCC in computer vision (e.g., user-defined embedding-based methods [17, 22] and data-driven methods [2, 25, 32]). However, FL-based OCC methods face significant challenges, including high communication costs, limited representation, and unstable learning processes. Additionally, centralized server-based methods [31, 42, 43] may not be suitable for FL-

based OCC due to limited computing resources on each device, which are a major obstacle for centralized methods that require substantial computing power.

The limitations inherent in FL-based OCC necessitated the development of a novel approach to address these challenges. In this paper, we aim to attain a more expressive global model without the need for frequent re-training processes between the central server and client devices on a large scale. To achieve this, we propose **Prototypical representation distillation based unsupervised Federated Learning (ProtoFL)** that distills representation from an off-the-shelf model learned using off-the-shelf datasets, regardless of individual client data. In contrast to traditional FL-based OCC, our proposed ProtoFL approach does not transfer parameters directly to the client, as depicted in Fig. 1. This resolves the issues of frequent round communication costs and the need for extra data due to one-time prototype representation distillation. ProtoFL provides a novel solution to the existing challenges of FL-based OCC, enabling efficient and effective global model updates.

Subsequently, we suggest a novel approach for estimating the density of a target class in a distributed setting using a flow-based one-class classifier [20]. To achieve this, we conduct the estimation on individual client devices, using augmented latent variables for training their distributed models. Our approach leverages two key techniques: maximum likelihood estimation with log-likelihood and a probabilistic similarity loss function that includes \mathcal{KL} -divergence. By combining the distillation and one-class classification phases, we can effectively handle complex data distributions that are non-independent and non-identically distributed across individual clients. Our two-phase learning framework is inspired by the success of flow-based models in various applications [20, 11, 40, 1], and we demonstrate its effectiveness in handling complex data distributions in the distributed settings.

The experimental findings of our proposed method demonstrate superior classification performance compared to both server-based and client-based methods on both image and tabular datasets. Additionally, we have validated the scalability of the learned representation and have shown that the global model learned by the ProtoFL is compatible with existing one-class classifiers as well as our one-class classifier based on the benchmark datasets. Our results indicate that our method is a promising approach for large-scale machine learning tasks that require robust and scalable classification capabilities.

Our contributions are described as follows:

- We propose a novel unsupervised federated learning framework that effectively addresses the challenge of insufficient local training data. By leveraging normalizing flows for local classifier learning and prototypical representation distillation, our approach enables effi-

cient and effective global model updates.

- We propose a novel prototype-based representation learning method for distilling normal data representation using an off-the-shelf model and dataset. Our approach demonstrates the scalability of the global model, which can be verified by adding new clients in FL-based OCC.
- We propose new federated and centralized learning methods for one-class classification, which we evaluate on five widely-used benchmarks. Our experiments show that our methods achieve superior performance compared to existing approaches.

2. Related Work

One-class classification Various one-class classification approaches have been proposed and categorized into description-based and representation-based learning. In the description-based methods, Deep-SVDD [31] performs one-class detection by learning a model to map target samples into a center in the latent space, otherwise non-target samples are mapped far from the center. And FCDD [27] proposes an explainable one-class classifier by upsampling based on gaussian sampling. In contrast, representation-based learning methods (DROC [42] and CSI [43]) present a two-step learning framework to learn a representation model through excessive data-augmentation and contrastive loss. The framework either employs a classifier or defines a score function for detecting a target class. Unlike previous centralized server-based approaches, our proposed approach for one-class classification on decentralized learning avoids the risk of personal data leakage by constructing a model using distributed data. To the best of our knowledge, our method is the first to directly address this problem in a decentralized setting.

Federated learning for local one-class classifier Federated Learning (FL) is a distributed machine learning paradigm that enables collaborative model training without direct data sharing. FL has been applied to one class classification (OCC) tasks for user verification and authentication, with several studies demonstrating promising results. [17, 32, 34]. FedAwS [45] introduces a geometric regularization to learn a global model by utilizing uploaded latent variables of joined clients in a central server. Since the latent variables shared on a server are private data, the FedAwS violates in the setting of FL. FedUV [17] employs independent secret error correcting codes (ECCs) to train a one-class classifier by preventing personal data from sharing among joined clients. The secret codes induce the concise objective of FedUV so that only positive examples are required. The FedUV approach was further refined to estimate a center of distribution by FedAA [32] and FedMet-

ric [34], instead of defining hand-designed codes. However, local data on each device is not sufficient to represent a centroid of local data distribution.

Federated learning for central and client classifiers In authentication systems, client data is often partitioned according to a target class. This results in highly or extremely non-i.i.d. data that poses a significant challenge for FL. This **extreme** non-i.i.d. means that a concentration parameter of Dirichlet distribution approaches to zero ($\alpha = 0.0$) while many federated learning methods [6, 9, 13] assumes non-i.i.d. data with non-zero ($\alpha \neq 0.0$) settings [18, 24, 29, 6, 13, 9].

Unsupervised federated learning methods [24, 29] aggregate local representation models and centroids for achieving a global model and centroid. FedRep [6] shares partitions of a global model to adapt each local heterogeneous data. FedX [13], which utilizes structure knowledge distillation between local and global knowledge relationships, learns meaningful data representation without sharing external data. SphereFed [9] introduced a learned matrix, which is a fixed-classifier for sharing among all participating clients, to transform each local data distribution into the predefined latent space. But, SphereFed requires enormous cost to re-construct and re-distribute a new learned matrix whenever a new client joins. Therefore, we first propose a prototypical representation distillation learning to save communication cost in FL with **extreme** non-i.i.d. ($\alpha = 0.0$) data settings.

Normalizing flows as classifiers Normalizing flows (NFs) normalize entangled data distributions into disentangled distributions by composing invertible and differentiable transformations. NFs have been applied for AD in image and video tasks [11, 40, 1] because FOOD [20] finds that NFs transforming latent to latent space outperforms those transforming data to latent space. For instance, FLOW [40] utilizes a fixed feature extractor to train NFs by either maximizing the log-likelihood on a target-class or minimizing the log-likelihood on outlier-exposure (OE) data. ITAE [1] employs NFs to estimate the density by learning appearance and motion latent features in videos. By the effectiveness of AD using NFs, we apply the NFs minimizing log-likelihood to our global model for one-class classification.

3. Preliminary

Federated Average Learning FedAVG [30], composed of participating clients and a central mediator server, trains a global model by aggregating locally-computed parameters, and broadcasts the updated global model to the clients. In each round, FedAVG updates the global model parameters

with local model parameters as follows:

$$\theta_g^{t+1} = \sum_{k=1}^K \frac{|D_k|}{\sum_{k=1}^K |D_k|} \theta_k^t, \quad (1)$$

where $k \in \{1, \dots, K\}$ indicates an index of a local client. θ_g^{t+1} is the parameters of a global model, and θ_k^t is the parameters of the k^{th} local model at a round t . For the balanced updating of the global model θ_g^{t+1} , $|D_k|$ is the number of samples on dataset D_k .

Unsupervised Contrastive Learning FL with **extreme** non-i.i.d. ($\alpha = 0.0$) data setting is unavailable to access samples of the other clients. Thus, we simplify the contrastive loss, $\mathcal{L}_{ctr} = y \times (1 - d(\bullet, \hat{\bullet})) + (1 - y) \times \max(0, d(\bullet, \hat{\bullet}))$, as follows:

$$\mathcal{L}_d = 1 - d(\bullet, \hat{\bullet}), \quad (2)$$

where $d(\bullet, \hat{\bullet})$ is the cosine-similarity $\frac{\bullet \cdot \hat{\bullet}}{\|\bullet\| \cdot \|\hat{\bullet}\|}$. We utilize L_d as our positive cosine similarity loss for our unsupervised learning, instead of \mathcal{L}_{ctr} .

Normalizing Flows Normalizing flows (NFs) [8] are the statistical methods using the change-of-variable law of probabilities to fit an arbitrary target density $p_R(r)$ by a tractable base distribution with density $p_Z(z)$ and a bijective invertible mapping $t^{-1} : R^D \rightarrow Z^D \Leftrightarrow t : Z^D \rightarrow R^D$. According to [33], \mathcal{KL} -divergence and log-likelihood estimation are employed to optimize the invertible flow-based model $z = t^{-1}(r; \psi)$ and $r = t(z; \psi)$ by the target distribution $p_R(r)$. The bijective invertible mapping is drawn as:

$$\begin{aligned} \mathcal{D}_{\mathcal{KL}}[\hat{p}_R(r) \| p_R(r; \psi)] \\ \approx -\mathbb{E}_{\hat{p}_R(r)}[\log p_Z(t^{-1}(r; \psi)) + \log |\det j_{t^{-1}}|] + cont, \end{aligned} \quad (3)$$

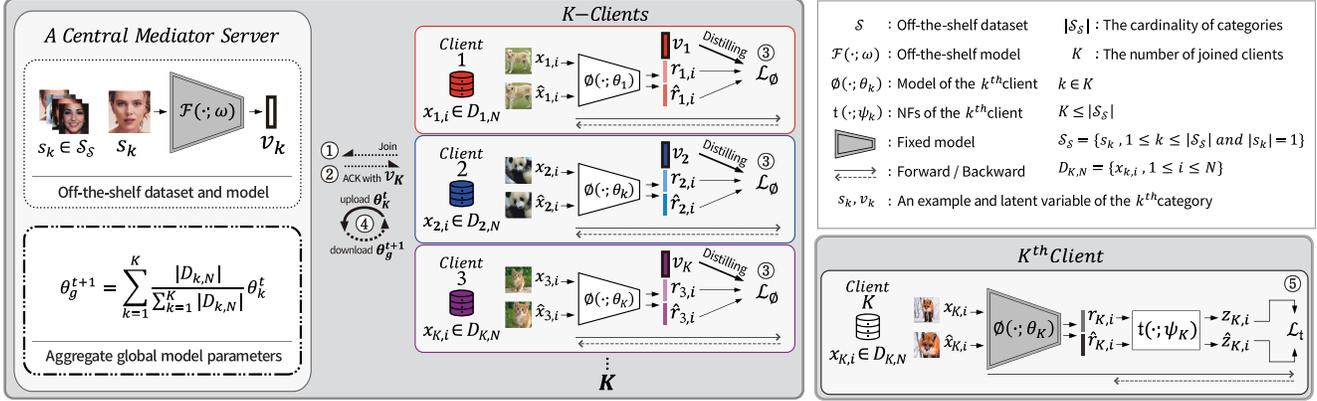
where $p_R(r; \psi)$ denotes the likelihood of a model. \hat{p}_R is the target distribution to learn the model by \mathcal{KL} -divergence. $p_R(r; \psi)$ only remains for learning parameters ψ , and then is replaced by $\frac{1}{N} \sum_{i=1}^N \left[\frac{\|t^{-1}(r_i)\|_2^2}{2} - \log |\det j_{t^{-1}}| \right]$ through the training dataset D_k as explained by [11, 20]. The re-written equation is applied to our objective to estimate the log-likelihood of given local data on each device.

4. Proposed Method

In this section, we propose a two-phase unsupervised learning framework that combines unsupervised federated learning via distillation of prototypical representations with local classifier learning via leveraging normalizing flows.

4.1. Problem Formulation

Participating K clients aim to train a global model by aggregating locally-computed parameters $\phi(\bullet, \theta)$ for an one-class classification task without sharing private data among



(a) Phase 1: Unsupervised federated learning via prototypical representation distillation

(b) Phase 2: Local one-class classifier based on NFs

Figure 2: Overview of our proposed two-phase learning architecture. ① ~ ④ indicates the workflows of ProtoFL with all the joined clients to upload parameters of each client and to download the aggregated global model, and ⑤ represents a local training for the OC-NF. The details of sequences ① ~ ⑤ are described in Section 4.2. NFs denote normalizing flows.

joined clients. Each client has a training dataset $D_{K,N} = \{x_{k,i}, 1 \leq i \leq N, 1 \leq k \leq K\}$ and N is the cardinality of the local data $D_{k,N}$. Let $\phi(x_{k,i}; \theta_k)$ denote the trainable model of the k^{th} client. In particular, we follow the **extreme** non-i.i.d. ($\forall m, n, D_{m,N} \cap D_{n,N} = \emptyset$) data setting in our FL-based OCC. Different from prior FL settings utilizing either hand-designed codes or additional training datasets, we invent generic representation of the local model ϕ by leveraging the off-the-shelf model \mathcal{F} and dataset \mathcal{S} , and we employ the flow-based model ψ for the local one-class classifier.

4.2. Sequence of procedures

As shown in Fig. 2, we describe our proposed method workflow in each phase. ① Whenever a new client participates in our unsupervised federated learning, all clients join a central mediator server. ② For the k^{th} client, a categorical image s_k is chosen from \mathcal{S} and transformed into a latent variable v_k by the central mediator server. Both ① and ② occur only once at the first time, and $\mathcal{V} = \{v_k, 1 \leq k \leq K\}$ is secretly distributed to the corresponding client. ③ We train the k^{th} local model θ_k via the local data $D_{k,N}$, the distributed prototype representation v_k , cosine similarity loss, and \mathcal{KL} -divergence loss. ④ A global model θ_g is aggregated with all the uploaded local models by FedAVG, and the global model is distributed to all participating clients. ⑤ Once federated representation learning has finished, we train a local one-class classifier based on normalizing flows (OC-NF) by using maximum likelihood and cosine similarity losses. We describe the details of our procedures (① ~ ⑤) in the following subsections.

4.3. Prototypical representation distillation based unsupervised federated learning (ProtoFL)

We assume $\mathcal{F} : R^{W \times H \times C} \rightarrow R^D$ and $\phi : R^{W \times H \times C} \rightarrow R^D$ indicate the off-the-shelf model and the local model,

respectively as illustrated in Fig. 2. For training the local model of the k^{th} client, we augment two views $x_{k,i}, \hat{x}_{k,i}$ from the i^{th} example of the local training data $D_{k,N}$, and subsequently transform the two samples as follows:

$$r_{k,i} = \phi(x_{k,i}; \theta_k); \hat{r}_{k,i} = \phi(\hat{x}_{k,i}; \theta_k), \quad (4)$$

where $r_{k,i}$ and $\hat{r}_{k,i}$ represent latent variables of the i^{th} example. ϕ is the client model with k^{th} learnable parameters θ_k . Inspired by SimCLR [5], the two latent variables r_k and \hat{r}_k must be very close to each other. Given the only positive local data and the constraint, we employ the positive cosine similarity loss (Eq. 2) to learn the local model as follows:

$$\mathcal{L}_p^\theta(\theta_k) = \frac{1}{N} \sum_{i=1}^N [1 - d(r_{k,i}, \hat{r}_{k,i})], \quad (5)$$

where θ_k is optimized by the local data $D_{k,N}$. To overcome the small cardinality of local data for training the local representative model, we propose the off-the-shelf model with pre-trained parameters $\mathcal{F}(\bullet; \omega)$ and the off-the-shelf dataset $s_k \in \mathcal{S}_s$. The off-the-shelf model and dataset are utilized on a central mediator server as depicted:

$$v_k = \mathcal{F}(s_k; \omega), \quad (6)$$

where s_k and v_k represent a prototype example and representation for the k^{th} client. After $\mathcal{V} = \{v_k, 1 \leq k \leq K\}$ is distributed to the corresponding client in secret, we train the local model of the k^{th} client to estimate the prototypical target representation v_k . We employ the \mathcal{KL} -divergence to minimize the discrepancy between two distributions v_k and either r_k or \hat{r}_k . Therefore, we invent the prototypical distillation loss as follows:

$$\mathcal{L}_{pd}^\theta(\theta_k) = \frac{1}{N} \sum_{i=1}^N [\mathcal{KL}(v_k \| r_{k,i}) + \mathcal{KL}(v_k \| \hat{r}_{k,i})], \quad (7)$$

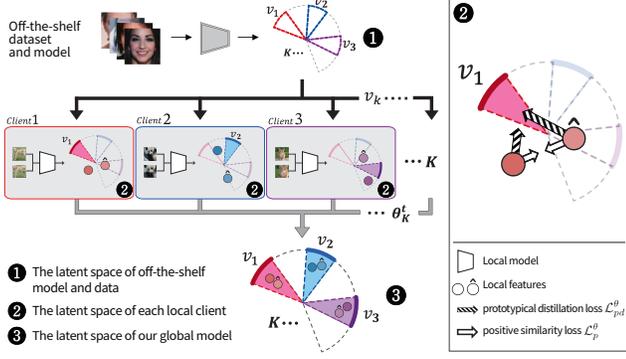


Figure 3: Illustration of learning a global model with prototypical representation distillation on distributed devices without sharing private data among participating clients.

where $\mathcal{KL}(v_k \| r_{k,i})$ and $\mathcal{KL}(v_k \| \hat{r}_{k,i})$ are equally contributed to our distillation loss. Figure 3 describes the details of Eq. 5 and Eq. 7, and the overall objective of the first phase (phase 1) is defined as follows:

$$\mathcal{L}_\phi(\theta_k) = (1 - \alpha) \times \mathcal{L}_{pd}^\theta(\theta_k) + \alpha \times \mathcal{L}_p^\theta(\theta_k), \quad (8)$$

where α is a balancing weight. In our experiment, α is empirically set to 0.1.

4.4. Local one-class classifier via Normalizing Flows (OC-NF)

After unsupervised federated representation learning for the global model, we leverage a general flow-based model for training the local one-class classifier in each client as shown in Fig. 2. To estimate log-likelihood of transformed latent variables r_k and \hat{r}_k , we exploit Eq. 3 to learn the probabilistic normalizing flows \mathfrak{t} through the local dataset $D_{k,N}$ and the local model $\phi(\cdot; \theta_k)$ on the k^{th} client as follows:

$$\begin{aligned} & \mathcal{L}_{mle}^\psi(\psi_k) \\ &= \frac{1}{N} \sum_{i=1}^N \left[\frac{\|\mathfrak{t}^{-1}(r_{k,i}; \psi_k)\|_2^2}{2} - \log |\det j_{k,i}| \right] + cont \\ &\approx \frac{1}{N} \sum_{i=1}^N \left[\frac{\|z_{k,i}\|_2^2}{2} - \log |\det j_{k,i}| \right] + cont, \end{aligned} \quad (9)$$

where $z_{k,i}$ is the estimated distribution as it passes through $\mathfrak{t}^{-1}(r_{k,i}; \psi_k)$. $r_{k,i}$ is a latent variable from $\phi(x_{k,i}; \theta)$. When optimizing the flow-based model, we ignore constants, det and *cont*. Since $r_{k,i}$ and $\hat{r}_{k,i}$ are close each other in a latent space, $z_{k,i}$ and $\hat{z}_{k,i}$ also predict the same distribution so that we introduce the following objective as a regularizer:

$$\mathcal{L}_{reg}^\psi(\psi_k) = \frac{1}{N} \sum_{i=1}^N [1 - d(z_{k,i}, \hat{z}_{k,i})], \quad (10)$$

where $z_{k,i}$ and $\hat{z}_{k,i}$ are the normal distribution estimated by $\mathfrak{t}^{-1}(\phi(x_{k,i}; \theta_k); \psi_k)$ and $\mathfrak{t}^{-1}(\phi(\hat{x}_{k,i}; \theta_k); \psi_k)$ respectively. The overall objective of the second phase (phase 2) is composed of Eq. 9 and Eq. 10 as follows:

$$\mathcal{L}_\mathfrak{t}(\psi_k) = \mathcal{L}_{mle}^\psi(\psi_k) + \lambda \times \mathcal{L}_{reg}^\psi(\psi_k), \quad (11)$$

where λ is a hyper-parameter for effectiveness of regularization. For our experiment, we empirically set λ to 0.01.

5. Experiments

In this section, we demonstrate the proposed approach on image and tabular benchmarks. We evaluate the performance of our proposed ProtoFL and compare our model with the other FL methods and one-class detectors. And we analyze the impact of new client participation in the context of OCC on FL, as well as ablation studies on each proposed component.

5.1. Datasets

We thoroughly evaluate our proposed method on several benchmark datasets, including the widely used MNIST [23], CIFAR-10 [21], CIFAR-100 [21], ImageNet-30 [16], and Keystroke-Dynamics [19], to tackle the demanding task of one-class classification. Our evaluation scrutinizes the efficacy, robustness, and scalability of our method, highlighting its strengths in handling real-world scenarios. In case of CIFAR-100, we adopt 20 super-class labels, denoted by CIFAR-100[‡], as suggested in [10] to our experiments. The numbers of categories in the benchmarks are 10, 10, 20, 30, and 51, which are the numbers of participating clients on our federated learning. The vision and tabular based benchmarks for OCC follow a one-vs-rest protocol [10, 42, 43]. In the protocol, a set of samples from one class indicates a target class for one client, whereas a set of samples from the remaining classes represents a non-target class for the remaining clients.

5.2. Experiment Setting

Off-the-shelf model and dataset In the central mediator server, the off-the-shelf dataset and model generate prototypical representation when a new client participates in our unsupervised federated learning. For the off-the-shelf model and dataset, we exploit ArcFace [7] based ResNet-50 backbone [14] and MS-Celeb-1M [12], respectively.

Architectures In this experiment, we employ ResNet-18/32 backbones [14] to learn the local and global models on our proposed approach. Since batch normalization is detrimental to the performance in federated learning [17] both ResNet-18/32 backbones replace batch- with group-normalization. In the Keystroke Dynamics dataset, we employ multi-layer perceptions as the local and global

Algorithm 1 Procedure of our two-phase learning

Phase 1 (ProtoFL) : Start phase 1 for T rounds.

Initialize and broadcast a global model θ_g^0 to all joined K clients. Send an acknowledgement (ACK) with prototypical representation v_k to the k^{th} client.

for $t \leftarrow 0$ to $T - 1$ **do**

$\theta_g^{t+1} \leftarrow \text{GLOBAL_TRAINING}(t, \theta_g^t)$

end for

Update the global model θ_g with θ_g^T .

function GLOBAL_TRAINING(t, θ_g^t)

Randomly select k number of clients from $\{C_k\}_{k=1}^K$.

for $k \leftarrow 1$ to K **do**

Broadcast the global model θ_g^t to C_k .

$\theta_k^t, |D_{k,N}| \leftarrow \text{CLIENT_TRAINING}(t, k, \theta_g^t)$

end for

$\theta_g^{t+1} \leftarrow \sum_{k=1}^K \frac{|D_{k,N}|}{\sum_{k=1}^K |D_{k,N}|} \theta_k^t$

return θ_g^{t+1}

end function

function CLIENT_TRAINING(t, k, θ_g^t)

Download the global model θ_g^t , and assign it to θ_k^t .

for $x_{k,i} \in D_{k,N}$ **do**

$\mathcal{L}_\phi(\theta_k^t) = (1 - \alpha) \times \mathcal{L}_{pd}^\theta(\theta_k^t) + \alpha \times \mathcal{L}_p^\theta(\theta_k^t)$

$\theta_k^t \leftarrow \theta_k^t - \eta \nabla \mathcal{L}_\phi(\theta_k^t)$

end for

Calculate the number of the k^{th} client data $D_{k,N}$.

return θ_k^t and $|D_{k,N}|$

end function

Phase 2 (OC-NF) : All joined clients start Algorithm 2 for OCC.

Algorithm 2 Procedure of our OC-NF for OCC

Download the global model θ_g , and assign it to the k^{th} client model θ_k . Freeze the global representation model $\phi(\bullet; \theta_k)$. Initialize the flow-based model of the k^{th} client \mathfrak{t}_{ψ_k} .

for $x_{k,i} \in D_{k,N}$ **do**

$\mathcal{L}_t(\psi_k) = \mathcal{L}_{mle}^\psi(\psi_k) + \lambda \times \mathcal{L}_{reg}^\psi(\psi_k)$

$\psi_k \leftarrow \psi_k - \eta \nabla \mathcal{L}_t(\psi_k)$

end for

model, consisting of three linear layers, ReLU, and group-normalization. And we utilize NFs as a local classifier model composed of 8 coupling layers [8].

Implementation details In the first phase as described in Algorithm 1, we train our expressive global model based on the FedAvg [30] method with 1 local epoch and 900 communication rounds, using randomly chosen clients from all participating clients. We utilize the RAdam optimizer [26] with betas 0.94 and 0.98, a weight decay 1e-3 and a learning rate 1e-6 to learn local models. In the second phase as described in Algorithm 2, we use the SGD optimizer [35] with a learning rate 5e-3 and 5 epochs to learn our one-class classifier. Data augmentations include the set of random processes (crop, horizontal-flip, and gaussian-blur) and color

jitter. To evaluate the performance of various methods, we use the area under curve of ROC curve (AUROC) and equal error rate (EER).

Method	Network	MNIST		CIFAR-10	
		AUROC	Round	AUROC	Round
FedAwS [45]	ResNet-32	99.6	10K	94.1	100K
FedUV [17]	ResNet-32	99.7	20K	87.2	20K
FedMetric [34]	ResNet-32	99.6	10K	94.2	100K
Ours	ResNet-32	99.9	0.9K	95.2	0.9K

Table 1: Performance comparison FL-based methods with our proposed approach on MNIST and CIFAR-10 benchmarks. Our approach outperforms both benchmarks while reducing the communication round cost.

Type	Method	Network	CIFAR-10	CIFAR-100 [‡]	ImageNet-30
			AUROC	AUROC	AUROC
Centralized Learning	AE [27]	-	-	-	56.0
	OC-SVM [41]	-	58.8	63.1	-
	Geom [10]	WRN-16-8	86.0	78.7	-
	Rot [16]	ResNet-18	89.8	77.7	77.9
	GOAD [3]	ResNet-18	85.1	74.5	-
	DROC [42]	ResNet-18	92.5	86.5	-
	CSI [43]	ResNet-18	94.3	86.6	91.6
	FLOW [†] [40]	ResNet-50	<u>95.2</u>	93.0	-
Federated Learning	FedUV [17]	ResNet-18	79.8 [*]	55.9 [*]	62.8 [*]
	FedRep [6]	ResNet-18	58.2	56.9	56.3
	Ours	ResNet-18	95.3	<u>89.9</u>	95.4

Table 2: AUROC of various centralized and decentralized detection methods on CIFAR-10/100[‡], and ImageNet-30 for OCC. [†] and ^{*} denote the usage of a pre-trained model learned on ImageNet-1M and the values from our re-implementation respectively, whereas **bold** and underline indicate the best results and the second results, respectively.

5.3. Experimental Results

Image and tabular benchmarks For the image benchmarks, we present the results on MNIST and CIFAR-10 datasets for OCC in **extreme** non-i.i.d. ($\alpha = 0.0$) data setting of FL. Table 1 shows the outstanding result of not only the significantly improved performance but also the effective communication cost on FL methods. We found that our proposed method exploits the off-the-shelf model and dataset to compensate the shortage of the local data. For updating a global model every rounds, naive FL methods [45, 17, 34] minimize the distance among the latent variables of the local data and map the distribution of the local data into one secure code designated. Those methods demand the inordinate communication cost and show the limited performance due to the deficiency of the local data on each device and the entangled secure code on latent space. Furthermore, we compare our approach with various centralized methods as depicted in Table 2. Our method outperforms ResNet-18 based methods on CIFAR-10 and ImageNet-30. In case of CIFAR-100[‡], our approach places the second-rank performance comparing with FLOW [40] leveraging the feature extraction of a pre-trained model learned on ImageNet-1M. Note that our proposed method is the first rank among ResNet-18 based methods having

Type	Method	Network	Plane	Car	Bird	Cat	Deer	Dog	Forg	Horse	Ship	Truck	Mean (std)
Centralized Learning	OC-SVM [41]	-	65.6	40.9	65.3	50.1	75.2	51.2	71.8	51.2	67.9	48.5	58.8 (± 11.6)
	DeepSVDD [37]	LeNet	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8 (± 7.17)
	Geom [10]	WRN-16-8	74.7	95.7	78.1	72.4	87.8	87.8	83.4	95.5	93.3	91.3	86.0 (± 8.52)
	Rot [16]	WRN-16-4	77.5	96.9	87.3	80.9	92.7	90.2	90.9	96.5	95.2	93.3	90.1 (± 6.52)
	GOAD [3]	WRN-16-4	77.2	96.7	83.3	77.7	87.8	87.8	90.0	96.1	93.8	92.0	88.2 (± 6.99)
	Rot [16]	ResNet-18	80.4	96.4	85.9	81.1	91.3	89.6	89.9	95.9	95.0	92.6	89.8 (± 5.75)
	GOAD [3]	ResNet-18	75.5	94.1	81.8	72.0	83.7	84.4	82.9	93.9	92.9	89.5	85.1 (± 7.61)
	DROC [42]	ResNet-18	90.9	98.9	88.0	83.1	89.9	90.3	93.5	98.2	96.5	95.2	92.5 (± 4.95)
	CSI [43]	ResNet-18	89.9	99.1	93.1	86.4	93.9	<u>93.2</u>	95.1	98.7	97.9	95.5	94.3 (± 3.97)
Federated Learning	Ours	ResNet-18	96.4	<u>97.9</u>	<u>92.7</u>	90.9	95.8	92.7	<u>96.4</u>	<u>96.4</u>	97.6	96.7	95.3 (± 2.38)

Table 3: AUROC of various centralized and decentralized detection methods on CIFAR-10 for one-class classification. We present the AUROC of each class and the mean and standard deviation (std) of AUROC for all classes. \dagger , **bold**, and underline denote the usage of a pre-trained model learned on ImageNet-1M, the best results, and the second results, respectively.

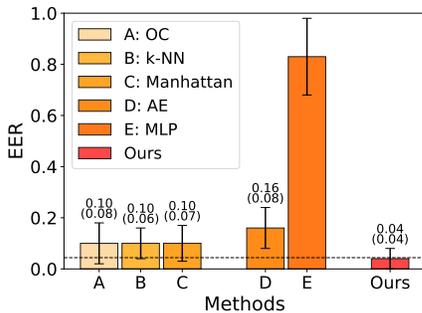


Figure 4: Comparison to previous methods of one-class classification on the Keystroke-Dynamics dataset. OC, Manhattan, AE, and MLP denote Outlier Count, Euclidean detector with Manhattan distance, Auto-Encoder, and 4-multi layer perception, respectively, from [19]. \mathbb{I} denotes a standard deviation.

no additional training datasets and limited computing resources. Among FL methods, our proposed method accomplishes the consistently improved performance on the image benchmarks.

For the tabular benchmark, we make the comparisons to the centralized methods on the Keystroke-Dynamics dataset for one-class classification as shown in Fig. 4. As the results, we found that our proposed method reduces EER more than two times comparing to prior methods. Training a foundation model as the global model is essential for the OCC task instead of the various architectures and distance metrics to achieve notably improved performance.

Centralized vs Federated learning As shown in Table 3, we compare our proposed method with the centralized learning based methods because there is no previous FL method studying sufficiently analysis results for CIFAR-10. Our proposed approach outperforms the previous methods in terms of the overall mean and standard deviation because we collaboratively train our global model to represent all classes by the distributively-learned local models, and then

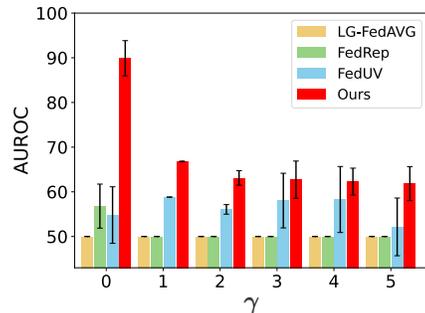


Figure 5: Analysis the scalability of our global model learned by ProtoFL through new clients on CIFAR-100 \ddagger dataset. Our approach modestly overcomes the other methods [6, 17, 24] in FL with **extreme** non-i.i.d. ($\alpha = 0.0$) data. γ and \mathbb{I} denote the number of new joined clients and a standard deviation each.

we apply our non-biased global model to transform raw data into latent spaces for learning a local one-class classifier in each client as shown in Fig. 5. In case of each class, this method shows the first and second-placed performance in 8 out of 10, whereas FLOW [40] and CSI [43] show the first and second-placed performance in 4 and 5 out of 10, which indicates that our collaboratively learned global model avoids over-fitting problem for each client.

Scalability of representation In this setup on CIFAR-100 \ddagger , let γ indicate the number of new joined clients for one-class classification, and $20 - \gamma$ represent the number of clients who participating in FL. To optimize the local classifier parameters of the new client, the new client downloads the global model as described in Algorithm 2. Note that we are the first to consider the effects of joining new clients. As presented in Fig. 5, we demonstrate the effectiveness for scalability of the learned representation as our global model through experiments and comparisons. Although new clients have increased, all federated learning methods suffer from degraded performances. However, our

Phase 1	Phase 2	CIFAR-10	CIFAR-100 [‡]	ImageNet-30	Mean
Ours (ProtoFL)	KDE	94.5	88.9	94.0	92.3
	GDE	94.3	88.6	92.8	91.9
	OC-SVM	97.1	86.9	90.8	91.6
	Ours (OC-NF)	95.3	89.9	95.4	93.5

Table 4: Ablation study of various classifiers on CIFAR-10, CIFAR-100[‡], and ImageNet-30 with our representative global model. The performance is evaluated by AUROC.

Method	Phase 1		Phase 2		CIFAR-100 [‡]	ImageNet-30
	L_{pd}^θ	L_p^θ	L_{mle}^ψ	L_{reg}^ψ	AUROC	AUROC
Ours w/o $L_{pd}^\theta, L_{reg}^\psi$		✓	✓		48.9	50.2
Ours w/o L_p^θ, L_{reg}^ψ	✓		✓		87.0	91.1
Ours w/o L_{reg}^ψ	✓	✓	✓		89.6	94.4
Ours	✓	✓	✓	✓	89.9	95.4

Table 5: Ablation study on CIFAR-100[‡] and ImageNet-30 to evaluate each component in the individual phase.

proposed method shows more scalability with respect to a new joiner than the others [17, 6, 24]. In addition, FedRep [6] and LG-FedAvg [24] are meaningless to support new clients. Thus, our method presents a novel property validating new joiners performance in FL.

5.4. Ablation Studies

We conducted ablation studies on CIFAR-10/100[‡] and ImageNet-30 (a) to verify the phase 1 and the phase 2, (b) to explore the contribution of each component in our method, and (c) to investigate the effectiveness of the off-the-shelf model in respect of the various off-the-shelf datasets.

Analyses of our global- and classifier- model With fixed our global model learned by joined clients in advance, we compared our proposed local one-class classifier via normalizing flows (OC-NF) with the prior one-class classifiers for the usability of the global model. As shown in Table 4, we observed the superiority of the global model and the local classifier either each or both comparing to the previous detectors [4, 41] non-parametric kernel density estimation, parametric gaussian density estimation, and one-class SVM (KDE, GDE, and OC-SVM).

Analyses of each component To analyze the effect of our proposed objective for each phase, we separated the representation objective of the phase 1 into distilling loss \mathcal{L}_{pd}^θ and similarity loss \mathcal{L}_p^θ , and the classification objective of the phase 2 into maximum likelihood \mathcal{L}_{mle}^ψ and regularizer \mathcal{L}_{reg}^ψ to show the importance in Table 5. Thus, we confirmed that each of them is the essential terms for the improving performance, and also observed the best results when all of them were used on CIFAR-100[‡] and ImageNet-30. In case of ImageNet-30, the improved performance indicates the effectiveness of this regularizer. As shown in Fig. 6, we found that our representative global model, trained without this distilling loss \mathcal{L}_{pd}^θ , leads the drastically impoverished representation.

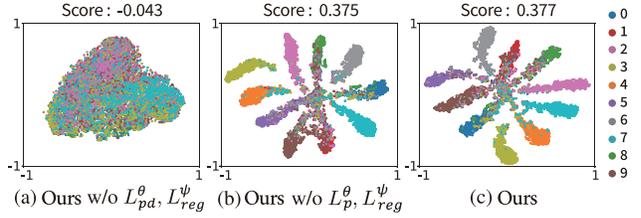


Figure 6: Visualization of latent-features using t-sne for our global model learned (*Phase 1*) with either each component or all. *Score* in each figure denote the silhouette values [36].

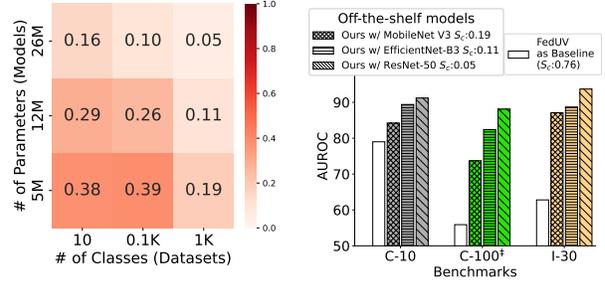


Figure 7: Left : Each score denotes cosine similarity \mathcal{S}_c between the various off-the-shelf datasets and models. Right : Performance comparison among the various global models whose off-the-shelf model is different. C and I denote CIFAR and ImageNet.

Analyses of distilling representation With the FL setting, we employ cosine similarity \mathcal{S}_c as a criterion for selecting the off-the-shelf models as shown in left of Fig. 7, and evaluated the performance based on various \mathcal{S}_c on all the benchmarks as shown in right of Fig. 7. Our proposed ProtoFL achieves the significantly improved performance comparing to FedUV since the ProtoFL is robust to the increased clients for one-class classification, and the results show that lower \mathcal{S}_c indicate more discriminative features. Note, distributing the prototypical representation in secure is important to overcome the shortage of training data, limited round communications, and greater number of clients.

6. Conclusion

We proposed ProtoFL, a method for achieving effective round cost and scalability representation in extreme non-i.i.d. data based FL. By utilizing an off-the-shelf model and dataset to distribute prototypical independent representations, we were able to learn a global model with all joined clients and optimize the flow-based classifier of each client. Our proposed method outperforms FL methods with efficient communication cost and presents a novel property of scalability of representation, validating new joiners' performance in FL. We believe that our method and experimental details could be adapted to handle further problems in both federated and centralized learning based one-class classification.

7. Acknowledgements

This work was supported by KakaoBank Corp., and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions). In particular, we would like to thank the designer Eunha Lim (*milkyway-mind@naver.com*) for well-presented Figures 1, 2, and 3.

References

- [1] Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129:108703, 2022. 2, 3
- [2] Divyansh Aggarwal, Jiayu Zhou, and Anil K. Jain. Fedface: Collaborative learning of face recognition model. In *IJCB*, 2021. 1
- [3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020. 6, 7
- [4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, number 2. ACM, 2000. 8
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4
- [6] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021. 3, 6, 7, 8
- [7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, June 2019. 5
- [8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. 3, 6
- [9] Xin Dong, Sai Qian Zhang, Ang Li, and H. T. Kung. Sphered: Hyperspherical federated learning. In *ECCV*, 2022. 3
- [10] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018. 5, 6, 7
- [11] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, pages 98–107, 2022. 2, 3
- [12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102. Springer, 2016. 5
- [13] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. FedX: Unsupervised Federated Learning with Cross Knowledge Distillation. In *ECCV*, 2022. 3
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5
- [15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019. 1
- [16] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 5, 6, 7
- [17] Hossein Hosseini, Hyunsin Park, Sungrack Yun, Christos Louizos, Joseph Soriaga, and Max Welling. Federated learning of user verification models without sharing embeddings. In *International Conference on Machine Learning*, pages 4328–4336. PMLR, 2021. 1, 2, 5, 6, 7, 8
- [18] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. 2019. 3
- [19] Kevin S Killourhy and Roy A Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 125–134. IEEE, 2009. 5, 7
- [20] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20578–20589. Curran Associates, Inc., 2020. 2, 3
- [21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [22] Liu L, Zhang Y, Gao H, and et al. Fedfv: federated face verification via equivalent class embeddings. 1
- [23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. att labs, 2010. 5
- [24] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 3, 7, 8
- [25] Chih-Ting Liu, Chien-Yi Wang, Shao-Yi Chien, and Shang-Hong Lai. Fedfr: Joint optimization federated framework for generic and personalized face recognition. abs/2112.12496. 1
- [26] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 6
- [27] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus Robert Muller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021. 2, 6
- [28] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Klaus Robert Muller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. In *TMLR*, 2022. 1

- [29] Ekdeep Singh Lubana, Chi Ian Tang, Fahim Kawsar, Robert P. Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering, 2022. [3](#)
- [30] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. [3](#), [6](#)
- [31] M M Moya, M W Koch, and L D Hostetler. One-class classifier networks for target recognition applications. 1 1993. [1](#), [2](#)
- [32] Poojan Oza and Vishal M. Patel. Federated learning-based active authentication on mobile devices. 2021. [1](#), [2](#)
- [33] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. In *JMLR*, 2021. [3](#)
- [34] Hyunsin Park, Hossein Hosseini, and Sungrack Yun. Federated learning with metric loss. [2](#), [3](#), [6](#)
- [35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. [6](#)
- [36] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. [8](#)
- [37] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. [7](#)
- [38] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020. [1](#)
- [39] Artem Ryzhikov, Maxim Borisyak, Andrey Ustyuzhanin, and Denis Derkach. Nfad: fixing anomaly detection using normalizing flows. *PeerJ Computer Science*, 2021. [1](#)
- [40] Robert Schmier, Ullrich Köthe, and Christoph-Nikolas Straehle. Anomaly detection using contrastive normalizing flows. *arXiv preprint arXiv:2208.14024*, 2022. [1](#), [2](#), [3](#), [6](#), [7](#)
- [41] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999. [6](#), [7](#), [8](#)
- [42] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minh Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*, 2020. [1](#), [2](#), [5](#), [6](#), [7](#)
- [43] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. [1](#), [2](#), [5](#), [6](#), [7](#)
- [44] Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Computers & Security*, 110:102402, 2021. [1](#)
- [45] Felix Yu, Ankit Singh Rawat, Aditya Menon, and Sanjiv Kumar. Federated learning with only positive labels. In *International Conference on Machine Learning*, pages 10946–10956. PMLR, 2020. [2](#), [6](#)