

# TOWARDS CAUSAL FEDERATED LEARNING FOR ENHANCED ROBUSTNESS AND PRIVACY

Sreya Francis<sup>1,2</sup>, Irene Tenison<sup>1,2</sup>, Irina Rish<sup>1,2</sup>  
MILA<sup>1</sup>, University of Montreal<sup>2</sup>

## ABSTRACT

Federated Learning is an emerging privacy-preserving distributed machine learning approach to building a shared model by performing distributed training locally on participating devices (clients) and aggregating the local models into a global one. As this approach prevents data collection and aggregation, it helps in reducing associated privacy risks to a great extent. However, the data samples across all participating clients are usually not independent and identically distributed (non-i.i.d.), and Out of Distribution (OOD) generalization for the learned models can be poor. Besides this challenge, federated learning also remains vulnerable to various attacks on security wherein a few malicious participating entities work towards inserting backdoors, degrading the generated aggregated model as well as inferring the data owned by participating entities. In this paper, we propose an approach for learning invariant (causal) features common to all participating clients in a federated learning setup and analyse empirically how it enhances the Out of Distribution (OOD) accuracy as well as the privacy of the final learned model.

## 1 EXISTING THREATS

### 1.1 DOMAIN SHIFT ISSUES

While federated learning promises better privacy and efficiency, most of the existing methods ignore the fact that the data on each client node are collected in a non-i.i.d manner, leading to data distribution shift issues between nodes (Quionero-Candela et al., 2019). For example, one device may take photos mostly indoors, while another mostly outdoors. Let  $A$  and  $B$  be two clients, and let  $P_A$  and  $P_B$  be their associated data distributions, respectively, in a Federated learning setup. In many real-world scenarios,  $P_A \neq P_B$ . The participating clients can have varying marginal distributions,  $\mathcal{P}_A(x)$ , although  $\mathcal{P}(y | x)$  remains the same, resulting into the so-called covariate shift; another example of shift is when marginal distributions of the class label,  $\mathcal{P}_A(y)$ , may vary across clients, even if  $\mathcal{P}(x | y)$  is the same, and so on (Kairouz et al., 2019).

### 1.2 PRIVACY THREATS

Although the Federated Learning process makes considerable efforts to keep the user’s data private, an attacker can analyze the weights of the sent updates to make conclusions about the data of users (Geyer et al., 2017). Certain machine learning algorithms such as Neural Networks and Recurrent Language models are known to memorize data labelling and patterns. In such cases, a user’s data may risk losing its privacy since they are represented in the model (McMahan et al., 2017). While this might sound unlikely if not done on purpose, there have been experiments that show it is possible to reconstruct some data points (Fredrikson et al., 2015). FL algorithms are vulnerable to some attacks, namely membership inference (Salem et al., 2018) (Shokri et al., 2017), model inversion (Fredrikson et al., 2018) and model extraction (Tramèr et al., 2016). Membership Inference typically determine whether a point is in the training dataset or not. (Shokri et al., 2017) propose a shadow training technique for this attack involving training  $k$  shadow models to mimic the behavior of target model initially, then accordingly train an attack (membership inference) model. Model Inversion attacks try to use black-box access to estimate the feature values from training dataset. (Fredrikson et al., 2018) explored model inversion attacks in two settings: decision trees and neural networks.

Model Extraction attacks try to duplicate the parameters of target model. (Tramèr et al., 2016) propose effective attack methods to logistic regression, neural networks and decision trees.

## 2 HOW CAN CAUSAL LEARNING ENHANCE FEDERATED LEARNING

Generalization to out-of-distribution (OOD) data in participating clients is still a challenging aspect for federated learning. This is because most statistical learning algorithms used in federated learning strongly rely on the i.i.d. assumption on client data, while in practice domain shift among participating client domains is common. As compared to associational models that are being used in federated learning, models that are learnt with respect to causal features always exhibit better generalization to non-iid data i.e. data from different distributions. As far as privacy is concerned, one of the main attacks posed to federated learning is membership inference attacks wherein only the model predictions can be observed by the attacker (Yeom et al., 2018), (Nasr et al., 2018b). In (Nasr et al., 2018b), it has been proved that the distribution of the training data as well as the generalizability of the model significantly contribute to the membership leakage. Particularly, they show that overfitted models are more susceptible to membership inference attacks than generalized models. Hence it can be inferred that such inference attacks can be nullified to a greater extent with learning networks that exhibit better generalization. In (Tople et al., 2020), the generalization property of causal learning has been proven wherein they establish a theoretical link between causality and privacy. It is shown that models learnt using causal features generalize better to unseen data, especially on data from different distributions than the train distribution. It was also proved that causal models provide better differential privacy guarantees as compared to the current associational models that we use. With our approach, we explore how causal learning can enhance the out of distribution robustness as well as the impact it can have on privacy enhancement in a federated learning setup.

## 3 PROPOSED APPROACH - CAUSALFED

### 3.1 IMPLEMENTATION WORKFLOW

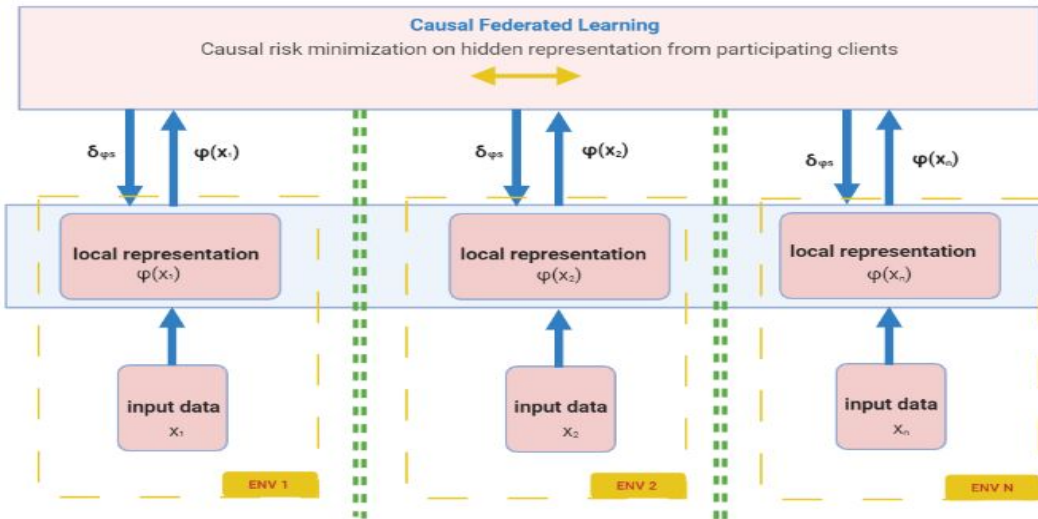


Figure 1: Causal Federated Learning

Keeping the data private, we propose an approach to collaboratively learn causal features common to all the participating clients in a federated learning setup. In our federated causal learning framework, the client layer (local) is the one where in each of the participating client entities do the local training for extracting features from their respective input data and outputs the respective features in the form of numerical vectors. Consider client data  $\mathcal{D}_C = (x_i^C, y_i^C)_{i=1}^{N_C}$  where  $x_i^C$  is  $i^{th}$  input and  $y_i^C$  is  $i^{th}$  label for client C. The hidden representation of each participating client is produced by

neural network as

$$h_i^C = \phi^C(x_i^C)$$

where  $h^C \in \mathcal{R}^{N_C \times d}$ ,  $d$  is the dimension of hidden representation layer. The global server layer is for the participating clients to exchange intermediate training components and train the federated model in collaboration by minimizing the empirical average loss as well as regularizing the model by the gradient norm of the loss for all the participating entities/environments as:

$$\sum_{C,i}^{S_i, N_C} \mathcal{L}_d(w \circ h_i, y_i) + \lambda \sum_C^S \left\| \nabla_{w|w=1.0} \sum_i^{N_C} \mathcal{L}_d(w \circ h_i, y_i) \right\|^2$$

where  $S$  equals set of clients/ source domains,  $N_C$  equals number of samples per client  $C$ ,  $\mathcal{L}_d$  equals classification loss, and  $h, y$  to represent the hidden representation and its corresponding true class label and  $\lambda$  is hyperparameter. With Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) we attempt to learn invariant predictors in a federated learning setup that can attain an optimal empirical risk on all the participating client domains. A.5 lists an alternative approach called CausalFedGSD for the same problem.

### 3.2 ALGORITHM

---

#### Algorithm 1 CausalFed

---

##### ServerCausalUpdate:

```

Initialize  $\mathbf{W}_0^s$ 
for each server epoch,  $t = 1, 2, \dots, k$  do
  Select random set of  $S$  clients
  Share initial model with the selected clients
  for each client  $k \in S$  do
     $(\phi(x_t^k), \mathbf{Y}^k) \leftarrow ClientRepresentation(k, \mathbf{W}_t^k)$ 
    Evaluate loss  $\mathcal{L}_k$ 
  end for
   $\mathcal{L}_s = \sum_k^S \mathcal{L}_k + \lambda \sum_k^S \left\| \nabla \mathcal{L}_k \right\|^2$ 
   $\mathbf{W}_{t+1}^s \leftarrow \mathbf{W}_t^s - \eta \nabla \mathcal{L}_s$ 
end for
 $\mathbf{W}_t^k \leftarrow ClientUpdate(\nabla \mathcal{L}_s)$ 
ClientRepresentation( $\mathbf{W}_t^k$ ):
  if  $k$  is first client to start training then
     $\mathbf{W}_t^k \leftarrow$  initial weights from server
  else
     $\mathbf{W}_t^k \leftarrow \mathbf{W}_{t-1}^{k-1}$  from the previous  $ClientUpdate(\nabla \mathcal{L}_s)$ 
  end if
  for each local client epoch,  $i = 1, 2, \dots, k$  do
    Calculate hidden representation  $\phi(x_t^k)$ 
  end for
  return  $\phi(x_t^k)$  and  $\mathbf{Y}^k$  to server
ClientUpdate:
  for each client  $k \in S$  do
     $\mathbf{W}_{t+1}^k \leftarrow \mathbf{W}_t^k - \eta \nabla \mathcal{L}_s$ 
  end for
  return  $\mathbf{W}_{t+1}^k$  to server

```

---

## 4 DATASET DETAILS

**Colored MNIST:** Unlike the MNIST dataset which consists of digits 0-9 in grayscale, the colored MNIST dataset consists of input images with digits 0-4 colored red and labelled 0 while digits 5-9

are colored green with label with shape of the digit as the causal feature. In our causal federated learning setup, we split the dataset to two environments, each corresponding to a participant/client. We sample 2000 data points per client/server domain. Within the client environments, 80 - 90 % of inputs have their color correlated to the digit whereas within the central server test environment has just 10% color-digit correlation which helps in testing the robustness despite the spurious correlation within the inputs.

**Rotated MNIST:** This dataset consist of original MNIST split to multiple client/participating environments by rotating each digit[0-9] with angles  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ , and  $60^\circ$ . We sample 1000 data points per client/server environment. The server side test domain consist of digits with angles  $75^\circ$  and  $90^\circ$

**Rotated Fashion MNIST:** Fashion-MNIST is a dataset of Zalando’s article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Here again we split the dataset to multiple client/participating environments by rotating each fashion item with angles  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ , and  $60^\circ$ . We sample 10000 data points per client/server environment. The server side test domain consist of fashion items with angles  $75^\circ$  and  $90^\circ$

## 5 RESULTS

In our experiments, we compare the performance of federated averaging (Fed-Avg) with the following approaches:

**Fed-ERM** Within the CausalFed setup, this approach minimizes the empirical average of loss over training data points and treats the data from different domains as i.i.d. ERM loss is given by:

$$\sum_{C,i}^{S,N_C} \mathcal{L}_s(w \circ h_i, y_i)$$

where S equals set of clients/ source domains,  $N_C$  equals number of samples per client  $C$ ,  $\mathcal{L}_s$  equals classification loss.

**CausalFed-RM** In this approach, we minimize the random match(RMatch) causal loss (Mahajan et al., 2020) within the CausalFed setup. RMatch loss is given by:

$$\sum_{C,i}^{S,N_C} \mathcal{L}_s(w \circ h_i, y_i) + \lambda * \sum_{\Omega(j,k)=1|j \sim N_C, k \sim N_{C'}} \text{Dist}(h_j, h_k)$$

where  $\Omega$  represents the match function used to randomly pair the data points across the different client domains.

**CausalFed-IRM** In this approach, we minimize the IRM loss (Arjovsky et al., 2019) within the CausalFed setup.

Table 1: Train Results

Dataset	Arch	Fed-Avg	Fed-ERM	CausalFed-RM	CausalFed-IRM
Colored MNIST	ResNet18	80.3%	<b>82.97</b> %	60.42 %	59.33 %
Rotated MNIST	ResNet18	85.2%	<b>86.5</b> %	79.8 %	80.2 %
Rotated FMNIST	LeNet	81.4%	<b>82.3</b> %	72.1 %	71.5 %

Table 2: Test Results

Dataset	Arch	Fed-Avg	Fed-ERM	CausalFed-RM	CausalFed-IRM
Colored MNIST	ResNet18	11%	10.2 %	<b>65.62</b> %	60.3 %
Rotated MNIST	ResNet18	82.7%	82.9 %	<b>90.2</b> %	89.1 %
Rotated FMNIST	LeNet	72%	71.6 %	<b>74.6</b> %	73.9 %

We observed that when clients have out of distribution data in a federated setup, FedAvg as well as FedERM does not fare well in the server side test data set though they give highly accurate results on train data(iid) whereas CausalFed-RM and CausalFed-IRM performs much better on test data(non iid).

**Privacy Leakage** In our experiments, within the CausalFed setup, we analyse the privacy leakage on 3 common attacks namely, Membership inference attack, Property inference attack and Backdoor attack. The privacy leakage on each of the attacks is measured by testing the accuracy of attack model. Details on each of the attacks are added in A.2 A.3.

Table 3: Leakage on inference attack

Dataset	Fed-Avg	Fed-ERM	CausalFed-RM	CausalFed-IRM
Colored MNIST	79.21 %	79.45 %	58.57 %	56.9 %
Rotated MNIST	84.4 %	85.24 %	68.3 %	64.4 %
Rotated FMNIST	76.61 %	78.23 %	57.55 %	55.7 %

We observe that in our setup with an out of distribution(OOD) test set, the membership inference attack accuracy of a federated causal client adversary model is much lesser as compared to a federated setup with associational client models. It was also observed that federated causal models provide better privacy guarantees against property inference attacks which could be owed to the fact that inversion based on learning correlations between attributes and final prediction, e.g., using color to predict the digit, can be eliminated by causal models, since a non-causal feature will not be included in our final causal federated model.

## 6 CONCLUSION

In this work, we show that CausalFed is more accurate than non-privacy-preserving federated learning approaches as well as superior to non-federated associational learning approaches in comparison to existing privacy enhancing approaches in federated setup which suffer from pretty high accuracy loss. We were able to experiment and confirm that causal feature learning can enhance out of distribution robustness in federated learning. Moving forward, we need to analyse the performance of this approach in real world datasets as well as compare various other causal learning approaches which can further enhance the out of distribution robustness and improve leakage protection in our current setup. We believe that CausalFed and CausalFedGSD serve as an initial approach to perform causal learning in a federated setting that offers several extensions for future work.

## REFERENCES

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.
- Melis C, Song E, De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. *IEEE Symposium on Security and Privacy*, 2019.
- Matt Fredrikson, Jha S, and Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. *ACM*, pp. pp. 1322–1333, 2015.
- Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. *arXiv:1806.01246*, 2018.
- Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv:1712.07557*, 2017.
- Peter Kairouz, Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, and Rachel Cummings. Advances and open problems in federated learning. *arxiv*, 2019.

- D Mahajan, S Tople, and A Sharma. Domain generalization using causal matching. *arXiv:2006.07500*, 2020.
- H Brendan McMahan, Ramage D.Talwar, and K.Zhang. Learning differentially private language models without losing accuracy. *arXiv:1710.06963*, 2017.
- Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *arXiv:1812.00910*, 2018a.
- Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. *arXiv:1812.00910*, 2018b.
- Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, , and Neil D Lawrence. Dataset shift in machine learning. *ISBN 0262170051*, 2019.
- Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Mleaks model and data independent membership inference attacks and defenses. *arXiv:1806.01246*, 2018.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *arXiv:1806.01245*, 2017.
- Shruti Tople, Amit Sharma, and Aditya V. Noris. Alleviating privacy attacks via causal learning. *arXiv:1909.12732*, 2020.
- Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, , and Thomas Ristenpart. Stealing machine learning models via prediction apis. *arXiv:1806.01246*, 2016.
- S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jh. Privacy risk in machine learning: Analyzing the connection to overfitting. *arXiv:1709.01604*, 2018.
- Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv:1806.00582*, 2018.

## A APPENDIX

### A.1 OBJECTIVE FUNCTIONS

**ERM** This approach minimizes the empirical average of loss over training data points and treats the data from different domains as i.i.d. ERM loss is given by:

$$\sum_{C,i}^{S,N_C} \mathcal{L}_s(w \circ h_i, y_i)$$

where S equals set of clients/ source domains,  $N_C$  equals number of samples per client  $C$ ,  $\mathcal{L}_s$  equals classification loss.

**RMatch** RMatch loss is given by:

$$\sum_{C,i}^{S,N_C} \mathcal{L}_s(w \circ h_i, y_i) + \lambda * \sum_{\Omega(j,k)=1|j \sim N_C, k \sim N_{C'}} \text{Dist}(h_j, h_k)$$

where  $\Omega$  represents the match function used to randomly pair the data points across the different client domains (Mahajan et al., 2020).

## A.2 INFERENCE ATTACK

### A.2.1 MEMBERSHIP INFERENCE

The main idea is that each training data point affects the gradients of the loss function such that the adversary can use Stochastic Gradient Descent algorithm (SGD) to extract information from other clients' data (Nasr et al., 2018a). The adversary can perform gradient ascent on a target data point before local parameter update. SGD reduces the gradient, in case the considered data point is part of a client's set resulting in a successful membership inference. Attack can come from both the client side and the server side. An adversarial client can observe the aggregated model updates and extract information about the union of the training dataset of all other participants by injecting adversarial model updates. For a server side attack, it can control the view of each target participant on the aggregated model updates and extract information from its dataset.

In our implementation, we sample 2,000 datapoints for Rotated-MNIST and 10,000 datapoints for Fashion-MNIST from the original train and test dataset to create the attack-train and attack-test dataset. We use pytorch code provided by (Shokri et al., 2017)(Nasr et al., 2018a)

### A.2.2 PROPERTY INFERENCE

The main idea behind this attack is that, at each round, each client's contribution is based on a batch of their local training data, so the attacker can infer properties that characterize the target dataset for which the adversary needs sample train data, which is labeled with the attribute to be inferred. (C et al., 2019) It is aimed at inferring properties of client data that are uncorrelated with the features that characterize the classes of the model. In our experiments we decided on client domain as the attribute which is to be inferred by the adversary. Another such attribute that is uncorrelated with the final prediction is the color of the input.

We observe that federated causal models provide better privacy guarantees against this attack which could be owed to the fact that inversion based on learning correlations between attributes and final prediction, e.g., using color to predict the digit, can be eliminated by causal models, since a non-causal feature will not be included in the our final causal federated model.

## A.3 BACKDOOR ATTACK

For an initial analysis, we experimented with two backdoor attacks:

- A single-pixel attack, where in the attacker changes the top-left pixel color of all the inputs, and mislabels them.
- A semantic backdoor where in the attacker selects certain features as the backdoors and misclassifies them. For example, the attacker classifies digits rotated  $15^\circ$  with label 7 as 0

In both the cases, CausalFed exhibited better resilience as compared to FedAvg.

## A.4 NETWORK ARCHITECTURE

Table 4: Network Architecture

Architecture	No of Layers	Kernel spec
LeNet	5	(5x5), (2x2)
AlexNet	8	(11x11), (5x5), (3x3)
ResNet18	18	(7x7), (3x3)

## A.5 CASUALFEDGSD - ALTERNATIVE APPROACH TO CAUSALFED

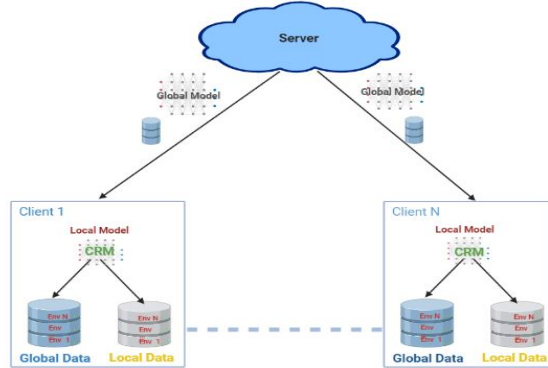


Figure 2: Causal Federated Learning

In (Zhao et al., 2018), it has been shown that globally shared data can reduce EMD(earth mover’s distance) between the data distribution on clients and the population distribution which can help in improved test accuracy. As this globally shared data is a separate dataset from that of the client, this approach is not privacy sensitive.

With the CausalFed approach, there can be privacy concerns regarding sharing the client data representation to the global server due to which depending on a global data set (with different environments) to enhance causal feature learning within a federated learning setup seems plausible. As we have no control on the clients’ data, we can distribute a small subset of global data containing a distribution over all the classes/environments from the server side to the clients during the initialization stage of federated learning.

The local model of each client is learned by minimizing the empirical average loss as well as regularizing the model by the gradient norm of the loss for both the shared data from server (Global Environment) and private data from each client (Local Environment). This enhances the learning of causal/invariant features common to both the client and global data environments without losing the privacy of client side data.

**Algorithm 2** CausalFedGSD**ServerUpdate:**


---

```

G ← distribution over all environments present in server
Initialize  $w_0$ 
Initialize random portion of G as  $G_0$ 
for each server epoch,  $t = 1, 2, \dots, k$  do
  Select random set of S clients
  Share  $G_0$  and initial model with the selected clients
  for each client  $k \in S$  do
     $w_{t+1}^k = \text{ClientUpdate}(k, w_t)$ 
  end for
   $w_{t+1} = \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$ 
end for

```

**ClientUpdate(w):**

```

 $\mathcal{E}_{tr} \in [\text{Client Env}] \cup [\text{Global Env}]$ 
for each local client epoch,  $t = 1, 2, \dots, k$  do
   $L_{\text{IRM}}(\Phi, w_t^k) = \sum_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi) + \lambda \cdot \mathbb{D}(w, \Phi, e)$ 
   $w_t^k = w_t^k - \eta \nabla L_{\text{IRM}}(w_t^k)$ 
end for
return w to server

```

---



Table 5: Train Results

<b>Dataset</b>	<b>Arch</b>	<b>Fed-Avg</b>	<b>Fed-ERM</b>	<b>CausalFedGSD-RM</b>	<b>CausalFedGSD-IRM</b>
Colored MNIST	ResNet18	80.3%	<b>82.97</b> %	57.42 %	55.32 %
Rotated MNIST	ResNet18	85.2%	<b>86.5</b> %	73.7 %	77.2 %
Rotated FMNIST	LeNet	81.4%	<b>82.3</b> %	69.2 %	68.6 %

Table 6: Test Results

<b>Dataset</b>	<b>Arch</b>	<b>Fed-Avg</b>	<b>Fed-ERM</b>	<b>CausalFedGSD-RM</b>	<b>CausalFedGSD-IRM</b>
Colored MNIST	ResNet18	11%	10.2 %	<b>55.62</b> %	52.3 %
Rotated MNIST	ResNet18	82.7%	82.9 %	<b>85.2</b> %	83.1 %
Rotated FMNIST	LeNet	72%	71.6 %	<b>71.9</b> %	70.2 %