The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Federated Causally Invariant Feature Learning

Xianjie Guo[1,2], **Kui Yu**[1,2*], **Lizhen Cui**[3], **Han Yu**[4], **Xiaoxiao Li**[4,5,6]
[1]School of Computer Science and Information Engineering, Hefei University of Technology, China
[2]Key Laboratory of Knowledge Engineering with Big Data of Ministry of Education, China
[3]School of Software, Shandong University, China
[4]College of Computing and Data Science, Nanyang Technological University, Singapore
[5]Department of Electrical and Computer Engineering, The University of British Columbia, Canada
[6]Vector Institute, Canada
xianjieguo@mail.hfut.edu.cn, yukui@hfut.edu.cn, clz@sdu.edu.cn, han.yu@ntu.edu.sg, xiaoxiao.li@ece.ubc.ca

Abstract

Federated feature selection (FFS) is a promising field for selecting informative features while preserving data privacy in federated learning (FL) settings. Existing FFS methods focus on capturing the correlations between features and labels. They struggle to achieve satisfactory performance in the face of data distribution heterogeneity among FL clients, and cannot address the out-of-distribution (OOD) problem that arises when a significant portion of clients do not actively participate in FL training. To address these limitations, we propose Federated Causally Invariant Feature Learning (`FedCIFL`), a novel approach for learning causally invariant features in a privacy-preserving manner. We design a sample reweighting strategy to eliminate spurious correlations introduced by selection bias and iteratively estimate the federated causal effect between each feature and the labels (with the remaining features initially treated as confounders). By iteratively refining the confounding feature set to identify the true confounders, `FedCIFL` mitigates the impact of limited local data on the accuracy of federated causal effect estimation. Theoretical analysis proves the correctness of `FedCIFL` under reasonable assumptions. Extensive experiments on synthetic and real-world datasets demonstrate the superiority of `FedCIFL` against eight state-of-the-art baselines, beating the best-performing approach by 3.19%, 9.07% and 2.65% in terms of average test Accuracy, RMSE and F1 score, respectively. It is a first-of-its-kind FFS approach capable of handling Non-IID and OOD data simultaneously. The source code is available at https://github.com/Xianjie-Guo/FedCIFL.

1 Introduction

Background In recent years, feature selection has become an increasingly important research topic due to its ability to improve model performance, reduce computational complexity, and enhance interpretability (Khaire and Dhanalakshmi 2022; Guo et al. 2022b; Xiao et al. 2022). Under federated learning (FL) settings (Yang et al. 2019; Yang, Fan, and Yu 2020; Kairouz et al. 2021; Guo et al. 2021; Li et al. 2024; Ren et al. 2024; Guo et al. 2024b), data are often distributed across multiple FL clients, making it challenging to perform feature selection across the entire dataset. This has led to the emergence of federated feature selection (FFS), which aims to select informative features while preserving data privacy (Banerjee, Elmroth, and Bhuyan 2021).

Challenges Recently, the FFS problem (Banerjee, Elmroth, and Bhuyan 2021; Cassará, Gotta, and Valerio 2022; Hu et al. 2022, 2023; Zhang et al. 2023; Hermo, Bolón-Canedo, and Ladra 2024; Banerjee et al. 2024) has been explored by considering scenarios where data are either Independent and Identically Distributed (IID) (Hu et al. 2023) or Non-Independent and Identically Distributed (Non-IID) (Banerjee, Elmroth, and Bhuyan 2021) across FL clients. A more detailed treatment of related work can be found in ***Appendix B.1***. However, practical FL often involves a vast number of clients with diverse data distributions. Furthermore, a significant proportion of these clients might not actively participate in the FL training process. As a result, the discrepancy in data distributions between the participating and non-participating (i.e., unseen) FL clients can cause the suboptimal performance of the trained models when applied to the unseen clients' data, a challenge commonly referred to as the out-of-distribution (OOD) problem (Yuan et al. 2022). This issue poses a critical challenge in FL, as the models trained on the participating clients' data might not generalize well to the unseen clients, thus limiting their applicability and effectiveness.

Motivation Existing FFS methods primarily exploit the correlation between labels and features. They cannot address the selection bias (Huang and Wu 2024) present in the data, resulting in the inability to learn feature subsets with strong generalization ability. Although some works attempt to address the OOD problem (Guo et al. 2023) and domain adaptation (Sun et al. 2021) in FL settings, they focus on learning generalizable representations in the representation space for classification tasks. While these methods can achieve satisfactory performance, they have poor interpretability as it is difficult to determine which original features have truly invariant relationships with the labels.

Contributions In this paper, we focus on learning causally invariant features in the original feature space to jointly address the challenges of Non-IID and OOD in FL settings. By leveraging the invariant property of causal features, we propose Federated Causally Invariant Features Learning (`FedCIFL`), a method for learning causally invariant features in a privacy-preserving manner to address the complex

[*]Corresponding Author

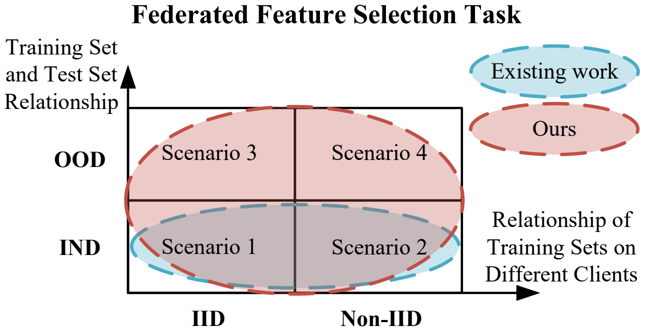
Federated Feature Selection Task
Training Set and Test Set Relationship
Existing work
Ours
OOD
IND
Scenario 3
Scenario 4
Scenario 1
Scenario 2
Relationship of Training Sets on Different Clients
IID
Non-IID

Figure 1: `FedCIFL` vs. existing works.

scenarios 3 and 4 as illustrated in Figure 1.

Specifically, `FedCIFL` first reweights samples on each client's local data with the aim of eliminating spurious correlations introduced by selection bias and learning the true causal relationships between labels and invariant features. Each FL client then computes the causal effect between each feature and the labels (by treating the remaining features as confounders) to obtain its local irrelevant feature subset. These subsets are sent to the FL server for alignment to produce the optimal irrelevant feature subset. However, the limited local data on each FL client might preclude the sample reweighting strategy from effectively learning the causal effect between labels and each feature when the confounder set is large. To address this issue, `FedCIFL` iteratively removes selected irrelevant features from the confounder set, and then repeats the aforementioned steps. This iterative process continues until no more irrelevant feature subsets are learned. The remaining features make up the invariant causal feature subset. By gradually reducing the size of the confounder set, `FedCIFL` mitigates the impact of limited local data on the accuracy of sample reweighting and, consequently, on causal effect estimation. Learning invariant features having a causal relationship with the labels enables strong generalization ability and interpretability.

To the best of our knowledge, `FedCIFL` is the first approach designed to perform FFS under Non-IID and OOD settings. Under reasonable assumptions, we theoretically prove the correctness of `FedCIFL`. Extensive experiments on both synthetic and real-world datasets demonstrate the superiority of `FedCIFL` against eight state-of-the-art baselines, beating the best-performing approach by 3.19%, 9.07% and 2.65% in terms of average test Accuracy, RMSE and F1 score, respectively.

2 Preliminaries

Notations and Definitions. In this paper, we focus on the horizontal FL setting, consisting of an FL server and a set of m FL clients $\{c_k\}_{k\in\{1,2,\ldots,m\}}$ with the same feature space. Each client c_k owns a private labeled dataset $\{(\mathbf{X}^{c_k}, \mathbf{Y}^{c_k})\}_{k=1}^{m}$, where $\mathbf{X}^{c_k} = \{\mathbf{x}_i^{c_k}\}_{i=1}^{n_k}$ follows distribution $\mathcal{P}^{c_k}$ over the feature space $\mathcal{X} = \{X_1, X_2, \ldots, X_d\}$ (i.e., $\mathbf{x}_i^{c_k} \sim \mathcal{P}^{c_k}$). $\mathbf{Y}^{c_k} = \{y_i^{c_k}\}_{i=1}^{n_k}$ denotes the ground-truth labels of $\mathbf{X}^{c_k}$. The total number of samples across all clients is denoted by $n = \sum_{k=1}^{m} n_k$. This paper focuses on Scenario 3 and Scenario 4 as illustrated in Figure 1. **For Scenario 3**, the data on different clients are IID, but the training set and the test set are OOD (i.e., $\mathcal{P}^{c_{k_1}} = \mathcal{P}^{c_{k_2}} \wedge \mathcal{P}^{c_{k_1}} \neq \mathcal{P}_{test}$ for $\forall k_1 \neq k_2, k_1, k_2 \in \{1, 2, \ldots, m\}$). **For Scenario 4**, the data on different clients are Non-IID, and the training set and the test set are OOD (i.e., $\mathcal{P}^{c_{k_1}} \neq \mathcal{P}^{c_{k_2}} \wedge \mathcal{P}^{c_{k_1}} \neq \mathcal{P}_{test} \wedge \mathcal{P}^{c_{k_2}} \neq \mathcal{P}_{test}$ for $\forall k_1 \neq k_2, k_1, k_2 \in \{1, 2, \ldots, m\}$).

For each client, we assume that the feature space $\mathcal{X}$ can be partitioned into two disjoint subsets: $\mathbf{X} = \{\mathbf{C}, \mathbf{V}\}$. We define $\mathbf{C}$ as the set of invariant causal features, and refer to the remaining features $\mathbf{V} = \mathbf{X} \setminus \mathbf{C}$ as irrelevant features, where the following assumption characterizes their properties:

Assumption 1 ((Kuang et al. 2018)**).** *There exists a probability mass function $P(y|c)$ such that for all distributions $\mathcal{P} \in \{\mathcal{P}^{c_1}, \ldots, \mathcal{P}^{c_m}, \mathcal{P}_{test}\}$, $Pr(\mathbf{Y}^{\mathcal{P}} = y|\mathbf{C}^{\mathcal{P}} = c, \mathbf{V}^{\mathcal{P}} = v) = Pr(\mathbf{Y}^{\mathcal{P}} = y|\mathbf{C}^{\mathcal{P}} = c) = P(y|c)$.*

By learning a model that captures the invariant function $P(y|c)$ under Assumption 1, `FedCIFL` can learn invariant causal features across all clients, and achieve strong generalization ability and interpretability across Scenario 3 and Scenario 4. We also adopt the overlap assumption, which is commonly used in the literature on treatment effect estimation (Athey, Imbens, and Wager 2018):

Assumption 2 (Overlap)**.** *For each client c_k, when setting any feature $\mathbf{X}_{\cdot,j}^{c_k}$ as the treatment feature, it satisfies $0 < P(\mathbf{X}_{\cdot,j}^{c_k} = 1|\mathbf{X}_{\cdot,-j}^{c_k}) < 1$, $\forall j$, where $\mathbf{X}_{\cdot,j}^{c_k}$ denotes the j-th feature in $\mathbf{X}^{c_k}$, and $\mathbf{X}_{\cdot,-j}^{c_k} = \mathbf{X}^{c_k} \setminus \mathbf{X}_{\cdot,j}^{c_k}$ represents all other features obtained by removing the j-th feature from $\mathbf{X}^{c_k}$.*

Accurately estimating causal effects between features and labels requires identifying the appropriate set of confounders, which influence both the feature and the label (Definition 1). Failure to account for confounders leads to biased causal effect estimates. In FL scenarios with limited local samples, selecting a suitable confounder set is crucial for achieving sample balance between treatment and control groups, ensuring accurate causal effect estimation.

Definition 1 (Confounders (Cai et al. 2023)**).** *A variable Z is a confounder for the effect of feature X on label Y if: 1) Z is associated with X: $P(X|Z) \neq P(X)$, 2) Z is associated with Y conditional on X: $P(Y|X,Z) \neq P(Y|X)$, and 3) Z is not a descendant of X in the causal graph.*

Supervised AutoEncoder. An unsupervised autoencoder is a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer. The autoencoder framework consists of two phases: encoding and decoding. Specifically, given input data $\mathbf{X}^{c_k}$, the autoencoder first employs multiple nonlinear encoding processes to learn low-dimensional representations $\xi(\mathbf{X}^{c_k})$ of $\mathbf{X}^{c_k}$. Subsequently, the autoencoder decodes $\xi(\mathbf{X}^{c_k})$ to obtain the reconstructed output data $\hat{\mathbf{X}}^{c_k}$. The encoding and decoding processes can be formalized as:

$$\begin{aligned}
&\text{Encode}: \xi^{(t)} = \sigma(\xi^{(t-1)}\mathbf{U}_1^{(t)} + \mathbf{b}_1^{(t)}), t = 1, 2, \ldots, l, \\
&\text{Decode}: \boldsymbol{\psi}^{(t)} = \sigma(\boldsymbol{\psi}^{(t-1)}\mathbf{U}_2^{(t)} + \mathbf{b}_2^{(t)}), t = 1, 2, \ldots, l,
\end{aligned} \tag{1}$$

where σ is a nonlinear activation function (e.g., sigmoid function). l is the number of hidden layers. Here,

$\xi^{(0)} = \mathbf{X}^{c_k}$, and $\xi^{(l)}$, denoted by $\xi(\cdot)$, represents the low-dimensional representations of $\mathbf{X}^{c_k}$. In addition, $\psi^{(0)} = \xi^{(l)}$, and $\mathbf{U}_1^{(t)}$ and $\mathbf{U}_2^{(t)}$ are the weight matrices, while $\mathbf{b}_1^{(t)}$ and $\mathbf{b}_2^{(t)}$ are the bias vectors. The autoencoder optimizes $\xi(\mathbf{X}^{c_k})$ by minimizing the reconstruction error between $\mathbf{X}^{c_k}$ and $\hat{\mathbf{X}}^{c_k}$. To further improve the quality of low-dimensional representations $\xi(\mathbf{X}^{c_k})$, the supervised autoencoder uses the label information and incorporates a cross-entropy loss $\ell(\cdot)$ into the objective function. Thus, the objective function of a supervised autoencoder is formalized as follows:

$$\mathcal{L}_{sae}^{c_k} = \frac{1}{n_k} \left\| \mathbf{X}^{c_k} - \hat{\mathbf{X}}^{c_k} \right\|_2^2 + \lambda_1 \sum_{t=1}^{l} \sum_{a=1}^{2} \left(\left\| \mathbf{U}_a^{(t)} \right\|_2^2 + \left\| \mathbf{b}_a^{(t)} \right\|_2^2 \right) + \lambda_2 \ell(f(\xi(\mathbf{X}^{c_k})), \mathbf{Y}^{c_k}), \tag{2}$$

where $f(\cdot)$ is a classifier, and λ_1 and λ_2 are the balancing parameters.

3 The Proposed FedCIFL Method

As illustrated in Figure 2, the proposed FedCIFL method consists of four iterative steps. Sections 3.1, 3.2, 3.3 and 3.4 respectively describes each of them. The detailed pseudo-code of FedCIFL is provided in ***Appendix C***, while theoretical analysis about the privacy and communication overhead of FedCIFL are presented in ***Appendix D***.

3.1 Sample Weight Learning and Causal Effect Estimation

Sample Weight Learning. As introduced in Section 2, the key challenge in estimating causal effects from observational data is to remove the confounding bias (Rubin 1973) induced by confounders that affect both the treatment T and the label. To this end, a confounder balancing technique is designed. Specifically, given a treatment feature T, when estimating its causal effect on the label, we first need to identify confounders. However, in observational studies, prior knowledge of the causal structure is unknown, meaning we do not know which features are confounders. Therefore, initially, all remaining features are treated as potential confounders (i.e., the confounder set of T for each FL client is $\mathcal{X} \setminus \{T\}$). Samples are then divided into two groups based on their T values, with $T = 1$ indicating a treatment group, and $T = 0$ indicating a control group. The causal effect of T on the label can be estimated by comparing the average difference between the treatment and control groups.

However, in practice, FL clients not only have different sample spaces but also typically possess limited local data, potentially leading to widespread sample selection bias. Consequently, the distribution of the treatment group often differs from that of the control group. Moreover, to select causally invariant features, we need to estimate the causal effect of each feature on the label. However, learning a separate set of sample weights for each feature is impractical, especially in FL scenarios with a large number of clients and potentially high-dimensional data. Inspired by (Kuang et al. 2018; Yang et al. 2023), we propose to learn a single set of weights W from a global perspective to align the distributions of the treatment and control groups corresponding to each feature. Consequently, the loss function for optimizing the sample weight set W^{c_k} on client c_k is:

$$\mathcal{L}_{sw}^{c_k} = \sum_{j=1}^{d} \left\| \sum_{i=1}^{n_k} W_i^{c_k} \cdot \mathbf{x}_i^{c_k} \cdot T_i^j - \sum_{i=1}^{n_k} W_i^{c_k} \cdot \mathbf{x}_i^{c_k} \cdot (1 - T_i^j) \right\|_2^2 + \lambda_3 \left(\sum_{i=1}^{n_k} W_i^{c_k} - n_k \right)^2 + \lambda_4 \sum_{i=1}^{n_k} \left(W_i^{c_k} - 1 \right)^2, \tag{3}$$

where $W_i^{c_k}$ is the weight of $\mathbf{x}_i^{c_k}$. λ_3 and λ_4 are the balancing parameters. $T_i^j \in \{0, 1\}$ denotes the value of the j-th feature when it is considered as the treatment feature for the i-th sample in $\mathbf{X}^{c_k}$. $\sum_{i=1}^{n_k} W_i^{c_k} \cdot \mathbf{x}_i^{c_k} \cdot T_i^j$ and $\sum_{i=1}^{n_k} W_i^{c_k} \cdot \mathbf{x}_i^{c_k} \cdot (1 - T_i^j)$ are the first-order moments of the treatment and control groups, respectively, for feature T.

In practice, nonlinear relationships among features and noise in the data can easily disrupt the balance of the data distribution between the treatment and control groups, leading to suboptimal quality of the learned weights W^{c_k}. To address this issue, we designed a supervised autoencoder, which offers several advantages. Firstly, it reduces the dimensionality of the confounders, thereby reducing the required sample size for local data on each client. Secondly, it captures nonlinear relationships among features, enabling a more accurate representation of the data. Thirdly, it mitigates the impact of noise in the original data, enhancing the robustness of the learned weights. Once the supervised autoencoder model is learned using Eq. (2) with input data $\mathbf{X}^{c_k}$ and the label $\mathbf{Y}^{c_k}$, the low-dimensional representations of the treatment and control groups can be obtained. Consequently, Eq. (3) can be rewritten as:

$$\mathcal{L}_{sw2}^{c_k} = \lambda_3 \left(\sum_{i=1}^{n_k} W_i^{c_k} - n_k \right)^2 + \lambda_4 \sum_{i=1}^{n_k} \left(W_i^{c_k} - 1 \right)^2 + \sum_{j=1}^{d} \left\| \frac{\xi(\mathbf{X}_{\cdot,-j}^{c_k})^T \cdot (W^{c_k} \odot \mathbf{X}_{\cdot,j}^{c_k})}{(W^{c_k})^T \cdot \mathbf{X}_{\cdot,j}^{c_k}} - \frac{\xi(\mathbf{X}_{\cdot,-j}^{c_k})^T \cdot (W^{c_k} \odot (1 - \mathbf{X}_{\cdot,j}^{c_k}))}{(W^{c_k})^T \cdot (1 - \mathbf{X}_{\cdot,j}^{c_k})} \right\|_2^2, \tag{4}$$

where $\odot$ is the Hadamard product. To improve the convergence speed of the reweighting loss function $\mathcal{L}_{sw2}^{c_k}$, we discretize the values of each representation in $\xi(\mathbf{X}_{\cdot,-j}^{c_k})$ into $(\omega+1)$ evenly spaced constants in the range of $[0, 1]$. Specifically, for $\forall i, q$, $[\xi(\mathbf{X}_{\cdot,-j}^{c_k})]_{i,q} \in \{0, \frac{1}{\omega}, \frac{2}{\omega}, \ldots, 1\}$, where $q \in \{1, 2, \ldots, p\}$ and $p = dim(\xi(\mathbf{X}_{\cdot,-j}^{c_k}))$. Since $\xi(\mathbf{X}_{\cdot,-j}^{c_k})$ is a low-dimensional representation of $\mathbf{X}_{\cdot,-j}^{c_k}$, we extend Assumption 2 from the binary original feature space to the multi-valued low-dimensional representation space and propose the following reasonable assumption:

Assumption 3. *For each FL client c_k, when setting any feature $\mathbf{X}_{\cdot,j}^{c_k}$ as the treatment feature, it satisfies $0 < P(\mathbf{X}_{\cdot,j}^{c_k} = 1 | \xi(\mathbf{X}_{\cdot,-j}^{c_k})) < 1$, $\forall j$.*

Then, we have Lemma 1 and Theorem 1 (proofs can be found in ***Appendix A.1*** and ***Appendix A.2***, respectively).

Lemma 1. *If for $\forall j$, $0 < P(\mathbf{X}_{\cdot,j}^{c_k} = 1 | \xi(\mathbf{X}_{\cdot,-j}^{c_k})) < 1$, and $\mathbf{X}^{c_k}$ is binary, then for $\forall i$, $0 < P(([\xi(\mathbf{X}_{\cdot,-j}^{c_k})]_{i,\cdot}, \mathbf{X}_{i,j}^{c_k}) = x) < 1$, where $([\xi(\mathbf{X}_{\cdot,-j}^{c_k})]_{i,\cdot}, \mathbf{X}_{i,j}^{c_k})$ is a sample of length $(p + 1)$, formed by concatenating the i-th row of the low-dimensional representation space $[\xi(\mathbf{X}_{\cdot,-j}^{c_k})]_{i,\cdot}$ with $\mathbf{X}_{i,j}^{c_k}$.*

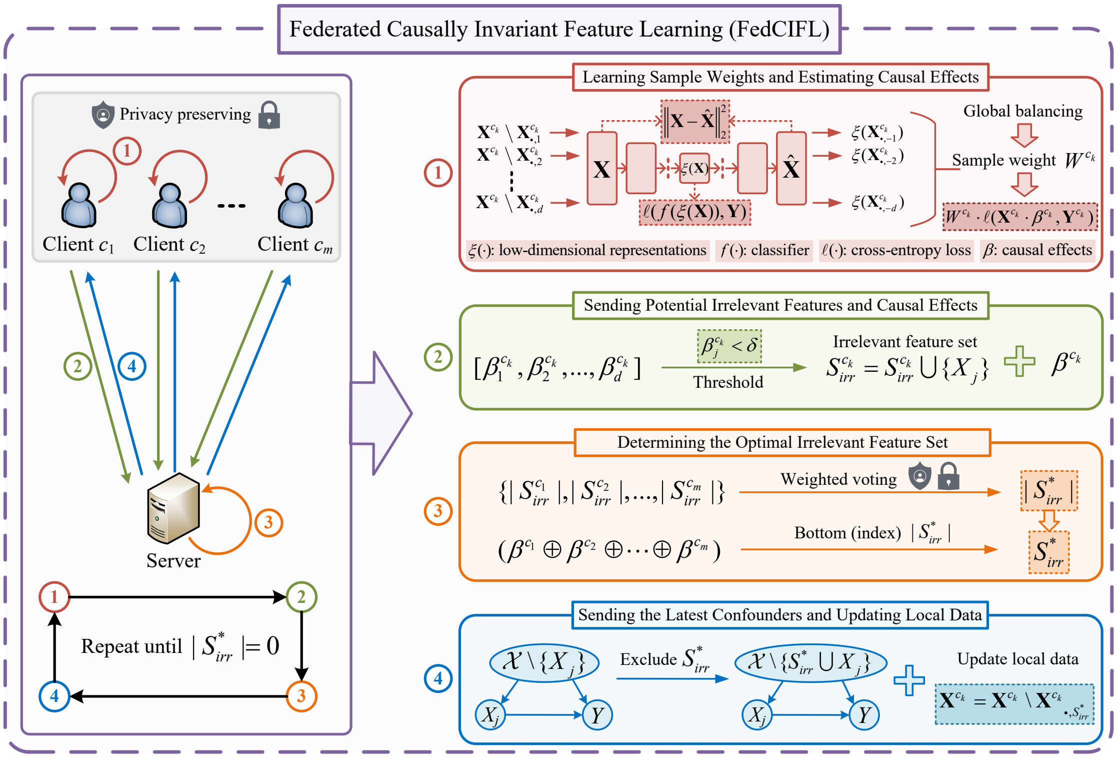
Federated Causally Invariant Feature Learning (FedCIFL)
Privacy preserving
1
Client c_1
Client c_2
Client c_m
2
4
Server
3
Repeat until $|S_{irr}^*| = 0$
Learning Sample Weights and Estimating Causal Effects
$\mathbf{X}^{c_k} \setminus \mathbf{X}^{c_k}_{\cdot,1}$
$\mathbf{X}^{c_k} \setminus \mathbf{X}^{c_k}_{\cdot,2}$
$\mathbf{X}^{c_k} \setminus \mathbf{X}^{c_k}_{\cdot,d}$
$\mathbf{X}$
$\|\mathbf{X}-\hat{\mathbf{X}}\|_2^2$
$\xi(\mathbf{X})$
$\ell(f(\xi(\mathbf{X})),\mathbf{Y})$
$\hat{\mathbf{X}}$
$\xi(\mathbf{X}^{c_k}_{\cdot,-1})$
$\xi(\mathbf{X}^{c_k}_{\cdot,-2})$
$\xi(\mathbf{X}^{c_k}_{\cdot,-d})$
Global balancing
Sample weight W^{c_k}
$W^{c_k} \cdot \ell(\mathbf{X}^{c_k} \cdot \beta^{c_k}, \mathbf{Y}^{c_k})$
$\xi(\cdot)$: low-dimensional representations
$f(\cdot)$: classifier
$\ell(\cdot)$: cross-entropy loss
β: causal effects
Sending Potential Irrelevant Features and Causal Effects
$[\beta_1^{c_k}, \beta_2^{c_k}, ..., \beta_d^{c_k}]$
$\beta_j^{c_k} < \delta$
Threshold
Irrelevant feature set
$S_{irr}^{c_k} = S_{irr}^{c_k} \bigcup \{X_j\}$
β^{c_k}
Determining the Optimal Irrelevant Feature Set
$\{|S_{irr}^{c_1}|, |S_{irr}^{c_2}|, ..., |S_{irr}^{c_m}|\}$
Weighted voting
$|S_{irr}^*|$
$(\beta^{c_1} \oplus \beta^{c_2} \oplus \cdots \oplus \beta^{c_m})$
Bottom (index) $|S_{irr}^*|$
S_{irr}^*
Sending the Latest Confounders and Updating Local Data
$\mathcal{X} \setminus \{X_j\}$
X_j
Y
Exclude S_{irr}^*
$\mathcal{X} \setminus \{S_{irr}^* \bigcup X_j\}$
Update local data
$\mathbf{X}^{c_k} = \mathbf{X}^{c_k} \setminus \mathbf{X}^{c_k}_{\cdot, S_{irr}^*}$

Figure 2: An overview of the proposed FedCIFL method.

Theorem 1. *Under Lemma 1, if the dimension p of the low-dimensional representation space $\xi(\mathbf{X}^{c_k}_{\cdot,-j})$ is finite, then ∃ a W^{c_k} such that $P(\lim_{n_k\to\infty} \sum_{j=1}^{d} \| \frac{\xi(\mathbf{X}^{c_k}_{\cdot,-j})^T \cdot (W^{c_k} \odot \mathbf{X}^{c_k}_{\cdot,j})}{(W^{c_k})^T \cdot \mathbf{X}^{c_k}_{\cdot,j}} - \frac{\xi(\mathbf{X}^{c_k}_{\cdot,-j})^T \cdot (W^{c_k} \odot (1-\mathbf{X}^{c_k}_{\cdot,j}))}{(W^{c_k})^T \cdot (1-\mathbf{X}^{c_k}_{\cdot,j})} \|_2^2 = 0) = 1$. In particular, a W^{c_k} solution that satisfies the above equation is $\hat{W}_i^{c_k} = 1/P(([\xi(\mathbf{X}^{c_k}_{\cdot,-j})]_{i,\cdot}, \mathbf{X}^{c_k}_{i,j}) = x)$.*

Therefore, based on Theorem 1 and Eq. (4), we can theoretically learn the optimal sample weights on the local data of each FL client under certain conditions.

Causal Effect Estimation. By learning the sample weights W^{c_k} at each FL client c_k, the confounding bias can be eliminated. It can be demonstrated that once the confounding bias is removed, the correlation between a given feature T and the label represents the causal effect (Kuang et al. 2017). Inspired by this, we design a weighted cross-entropy loss function, which is to be minimized, to estimate the causal effect of each feature on the label at client c_k as:

$$\mathcal{L}_{wce}^{c_k} = -\sum_{i=1}^{n_k} W_i^{c_k} \cdot (y_i^{c_k} \cdot \log \frac{1}{1+exp(-\mathbf{x}_i^{c_k} \cdot \beta^{c_k})} + (1-y_i^{c_k}) \cdot \log(1 - \frac{1}{1+exp(-\mathbf{x}_i^{c_k} \cdot \beta^{c_k})})) + \lambda_5 \|\beta^{c_k}\|_1, \tag{5}$$

where $y_i^{c_k}$ is the label of $\mathbf{x}_i^{c_k}$, $\beta_j^{c_k}$ is the causal effect between the j-th feature and the label at client c_k, and λ_5 is the balancing parameter.

3.2 Transmission of Potentially Irrelevant Features and Causal Effects

Based on Eq. (5), at client c_k, we can learn the causal effect values $\beta^{c_k} = [\beta_1^{c_k}, \beta_2^{c_k}, \ldots, \beta_d^{c_k}]^T$ of each feature on the label. Due to the diverse sample spaces across FL clients, the learned $\{\beta^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ might vary significantly. This step aims to determine the potentially irrelevant feature sets learned on each client based on β^{c_k}, and send them to the server to determine the optimal irrelevant feature set in Section 3.3. Specifically, given a fixed threshold $\delta > 0$, if $|\beta_j^{c_k}| \geq \delta$, the j-th feature at c_k is considered a causally invariant feature; otherwise, it is deemed as an irrelevant feature. Let $S_{irr}^{c_k}$ denote the irrelevant feature set learned by c_k. We have: $S_{irr}^{c_k} = S_{irr}^{c_k} \cup \{X_j\}$ if $|\beta_j^{c_k}| < \delta$. Finally, the irrelevant feature sets of all FL clients $\{S_{irr}^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ can be obtained.

3.3 Optimization of the Irrelevant Feature Set

To learn the optimal irrelevant feature set S_{irr}^* across all clients, the first step is to determine the optimal number of elements in the irrelevant feature set (i.e., $|S_{irr}^*|$). A common approach is perform majority voting using the learned $|S_{irr}^{c_k}|_{k\in\{1,2,\ldots,m\}}$ from all clients to determine the mode $|S_{irr}^*|$. However, in practice, FL systems often face conflicts arising from multiple modes. Traditional approaches resolve such conflicts by assuming knowledge of each client's sample size and performing weighted decision-making based on this information (Yang et al. 2019). Here, the sample size of clients is often considered a form of privacy as well (Guo et al. 2024a). To achieve a higher degree of privacy protection, we propose a novel strategy that assumes the sample size of each client is unknown. Our approach is based on a key observation: when the autoencoder model is sufficiently expressive and the sample size is large enough, the supervised autoencoder loss for each client will primarily be determined by the weight regularization term. This implies that clients with larger sample sizes are generally better at minimizing their local loss function, as they can more effectively learn the underlying data distribution.

Following this observation, we can reasonably conclude that when the weight regularization terms are comparable across different clients during the training of autoencoder models, clients with larger sample sizes tend to achieve better optimization of their local loss function. This suggests that such clients should be given more weight in their contribution to the global model.

Building on this theoretical foundation, we propose a stronger privacy-preserving strategy to handle conflicts arising from multiple modes. We introduce a vector Δ that represents the weighted ranking of each client. This ranking is calculated based on $\mathcal{L}_{sae}^{c_k}$ achieved by training a supervised autoencoder using Eq. (2) on each client. According to the previous analysis, we assign higher weight rankings to clients with lower $\mathcal{L}_{sae}^{c_k}$ values. Thus, Δ is defined as:

$$\Delta = \overset{\circ}{Rank}([\mathcal{L}_{sae}^{c_1}, \mathcal{L}_{sae}^{c_2}, \dots, \mathcal{L}_{sae}^{c_m}]). \tag{6}$$

$\overset{\circ}{Rank}(\cdot)$ takes a vector as input and returns a new vector of the same size, where each element in the output vector represents the rank of the corresponding element in the input vector sorted in ascending order. For example, $\Delta(k) = 3$ indicates that client c_k has the third highest weight ranking among all clients. If there exist multiple modes $M_h \in \{M_1, M_2, \dots\}$ in $\{|S_{irr}^{c_k}|\}_{k\in\{1,2,\dots,m\}}$, $|S_{irr}^*|$ can be calculated as:

$$\begin{aligned} |S_{irr}^*| =& \arg\min\nolimits_{M_h \in \{M_1, M_2, \dots\}} (\sum_{k=1}^{m} \Delta(k) \\ & subject\ \ to\ \ |S_{irr}^{c_k}| = M_h). \end{aligned} \tag{7}$$

By adopting this strategy, `FedCIFL` can determine the optimal size of the irrelevant feature set without requiring knowledge of individual clients' sample sizes, thereby providing stronger privacy protection. Subsequently, we rank the total causal effect of each feature on the label learned from all clients in descending order, and select the bottom $|S_{irr}^*|$ elements as the optimal set of irrelevant features S_{irr}^*:

$$S_{irr}^* = \overset{\circ}{Bottom}_{|S_{irr}^*|}(\beta^{c_1} \oplus \beta^{c_2} \oplus \cdots \oplus \beta^{c_m}). \tag{8}$$

$\overset{\circ}{Bottom}_{|S_{irr}^*|}$ is used to obtain the feature index corresponding to the bottom $|S_{irr}^*|$ elements in a vector based on the order of their values.

3.4 Latest Confounder Transmission and Local Data Updates

In Section 3.1, we initially regarded all features in $\mathcal{X}$ (except for the treatment feature T) as potential confounders. However, in FL scenarios with limited sample sizes in local datasets, identifying a set of confounders much larger than the true set makes it difficult to achieve sample balance between treatment and control groups within each local dataset. In addition, if irrelevant features are mistakenly considered confounders, both treatment and control group data will include these irrelevant features, disrupting the balance of true positive confounders. Consequently, the learned sample weight set W^{c_k} might be inaccurate. Recent research indicates that failing to adjust for confounders properly can lead to incorrect conclusions (Shi, Blei, and Veitch 2019; Yao et al. 2021). In other words, if confounders are not well-balanced, the causal effect estimation will be flawed, resulting in low-quality causal feature sets. Therefore, removing irrelevant features from the confounder set is crucial to achieving a more accurate causal effect estimation. According to Definition 1, irrelevant features are definitively not confounders. Thus, we remove the learned optimal set of irrelevant features from the original feature space $\mathcal{X}$, and update the original dataset $\mathbf{X}^{c_k}$ to enable more accurate causal effect estimation as:

$$\mathbf{X}^{c_k} = \mathbf{X}^{c_k} \setminus \mathbf{X}^{c_k}_{\cdot, S_{irr}^*}. \tag{9}$$

Finally, as illustrated in Figure 2, `FedCIFL` naturally converges by iteratively executing Steps 1 to 4 until $|S_{irr}^*| = 0$.

4 Experimental Evaluation

4.1 Experiment Settings

Datasets. The datasets used in the experiments include the following two types.

• **Synthetic data.** Firstly, we generate the features $\mathbf{X} = \{\mathbf{C}, \mathbf{V}\} = \{C_1, \dots, C_{d_c}, V_1, \dots, V_{d_v}\} \sim \mathcal{N}(0, 1)$ from an independent Gaussian distribution, where $d_c + d_v = d$. To make $\mathbf{X}$ binary, we set $\mathbf{X}_{i,j} = 1$ when $\mathbf{X}_{i,j} > 0$; otherwise, $\mathbf{X}_{i,j} = 0$. To simulate complex causal relationships, we separate the invariant causal features into a linear part $\mathbf{C}_l$ and a non-linear part $\mathbf{C}_n$. Then, we generate the label data $\mathbf{Y}$ using the following function (Kuang et al. 2018):

$$\begin{aligned} \mathbf{Y} =& 1/(1 + exp(-\sum_{\mathbf{X}_{\cdot,j_1} \in \mathbf{C}_l} \alpha_{j_1} \cdot \mathbf{X}_{\cdot,j_1} - \\ & \sum_{\mathbf{X}_{\cdot,j_2} \in \mathbf{C}_n} \beta_{j_2} \cdot \mathbf{X}_{\cdot,j_2} \cdot \mathbf{X}_{\cdot,(j_2+1)})) + \mathcal{N}(0, 0.2). \end{aligned} \tag{10}$$

$\alpha_{j_1} = (-1)^{j_1} \cdot (j_1\%3 + 1) \cdot d/3$ and $\beta_{j_2} = d/2$. To generate different data distributions that simulate the complex scenarios as in Figure 1, we create a set of distributions $\{\mathcal{P}^{c_1}, \dots, \mathcal{P}^{c_m}, \mathcal{P}_{test}\}$ by varying $P(\mathbf{Y}|\mathbf{V})$ with a bias rate $r \in [0, 1]$. Specifically, to emulate Scenario 3 (i.e., $\mathcal{P}^{c_{k_1}} = \mathcal{P}^{c_{k_2}} \wedge \mathcal{P}^{c_{k_1}} \neq \mathcal{P}_{test}$ for $\forall k_1 \neq k_2, k_1, k_2 \in \{1, 2, \dots, m\}$), we set $r^{c_k} = 0.4$ and $r_{test} = 0.9$. To emulate Scenario 4 (i.e., $\mathcal{P}^{c_{k_1}} \neq \mathcal{P}^{c_{k_2}} \wedge \mathcal{P}^{c_{k_1}} \neq \mathcal{P}_{test} \wedge \mathcal{P}^{c_{k_2}} \neq \mathcal{P}_{test}$ for $\forall k_1 \neq k_2, k_1, k_2 \in \{1, 2, \dots, m\}$), we set $r_{test} = 0.9$ and then uniformly assign different bias rates to each client within the interval $[0.1, 0.7]$ using the following equation:

$$r^{c_k} = 0.1 + (k-1) * \frac{0.7 - 0.1}{m - 1}, \quad k \in \{1, 2, \dots, m\}. \tag{11}$$

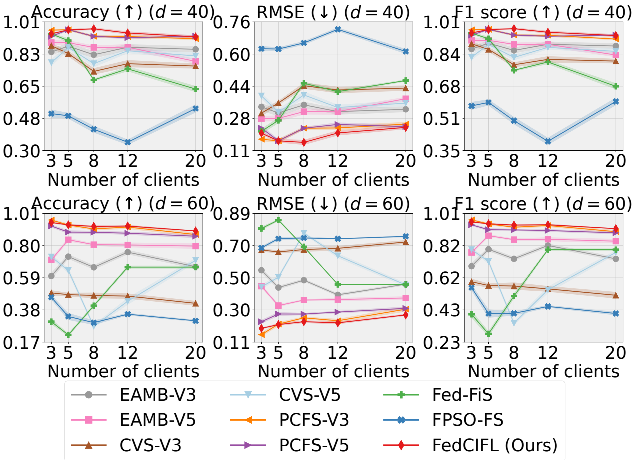

Figure 3: Results on synthetic datasets where data is IID across clients but OOD for the test set. A total of 6,000 samples are *unevenly* distributed among {3, 5, 8, 12, 20} clients.

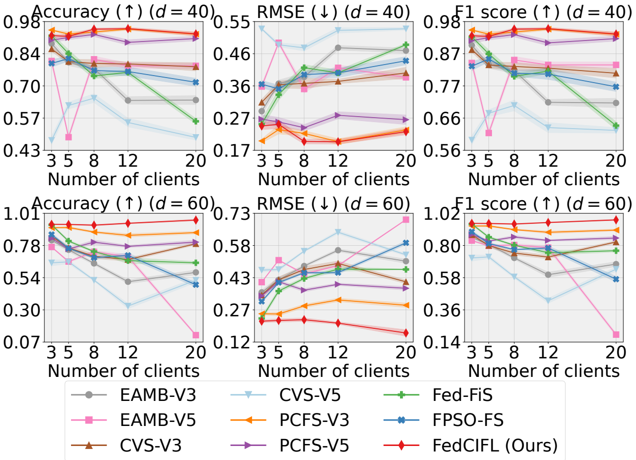

Figure 4: Results on synthetic datasets where data is Non-IID across clients and OOD for the test set.

In addition, to further simulate practical FL scenarios, the local datasets at different clients are set to different sample sizes in our experiments. Let $n = \sum_{k=1}^{m} n_k$ be the sum of sample sizes owned by the m clients, the sample size of each local dataset is set as:

$$n_1 = \lfloor \frac{n}{2m} \rfloor, \quad n_k = n_1 + \lfloor \frac{2(n - mn_{c_1})}{m(m-1)} \rfloor (k-1), \tag{12}$$

where $k \in \{2, 3, \ldots, m\}$.

- **Real-world data.** We also compare FedCIFL with the baselines on the Amazon Review dataset. Amazon Review is a cross-domain sentiment classification dataset of product reviews collected from four types of products: *Books* (B), *DVDS* (D), *Electronics* (E) and *Kitchen appliances* (K), each of which contains about 1,000 positive and 1,000 negative reviews. In our experiments, we use the preprocessed version of the Amazon Review dataset reported in (Wang et al. 2018), and construct four tasks: 1) DEK→B, 2) BEK→D, 3) BDK→E and 4) BDE→K, where "DEK→B" indicates that the D, E and K domain datasets are used as the FL training data, and the B domain dataset is used as the testing data.

Comparison Baselines. We compare FedCIFL with two state-of-the-art FFS methods: 1) Fed-FiS (Banerjee, Elmroth, and Bhuyan 2021) and 2) FPSO-FS (Hu et al. 2023). Since FedCIFL focuses on capturing causal features, we also include three state-of-the-art causal feature selection methods for a more comprehensive comparison: EAMB (Guo et al. 2022a), CVS (Kuang et al. 2023) and PCFS (Yang et al. 2023). Since existing causal feature selection methods have not yet considered FL scenarios, we implement six additional FL variants of these methods: 3) EAMB-V3, 4) EAMB-V5, 5) CVS-V3, 6) CVS-V5, 7) PCFS-V3 and 8) PCFS-V5. In these new baselines, "-V3" and "-V5" denote the use of 30% and 50% thresholds, respectively, when voting on the causal feature subsets learned from different clients to obtain the optimal causal feature subset. For more detailed discussions on these related works on causal feature selection methods, please refer to ***Appendix B.2***. Implementation details of the FedCIFL algorithm and the baselines are provided in ***Appendix E***.

Evaluation Metrics. Based on the selected features, we establish an FL system to train logistic regression (LR) and multilayer perceptron (MLP) classifiers separately. These classifiers are employed to perform classification tasks in an FL setting, where the training data is distributed across multiple participating clients. We then evaluate the quality of the selected features using Test Accuracy, Root Mean Square Error (RMSE), and F1 score (Guo et al. 2022a; Xiao et al. 2024). Comprehensive experimental results of various metrics on the LR classifier can be found in ***Appendix F***.

4.2 Results and Discussion (Synthetic Data)

We emulate Scenario 3 (i.e., the data on different clients are IID, but the training set and the test set are OOD) and Scenario 4 (i.e., the data on different clients are Non-IID, and the training set and the test set are OOD) from Figure 1 on synthetic data. The experimental results are presented in Figures 3 and 4, respectively. In Figure 3, it can be observed that FedCIFL achieves the best performance on all metrics in most cases. Moreover, compared to other baselines, the performance of our method remains stable as the number of clients and the data dimension d increase. This demonstrates that FedCIFL indeed captures causally invariant features, leading to satisfactory generalization performance. Existing FFS algorithms (i.e., Fed-FiS and FPSO-FS) focus on capturing the correlation between features and labels, resulting in suboptimal performance and large fluctuations in metrics in this OOD scenario. Although existing causal feature selection algorithms aim to capture causal features, they lack reasonable and effective federated aggregation strategies, leading to the loss of some causally invariant features or the inclusion of additional irrelevant features. As a result, their performance is inferior to FedCIFL.

From Figure 4 which depicts a more complex federated training scenario, it can be seen that the performance gaps between FedCIFL and existing FFS and causal feature se-

Metrics	Tasks	EAMB-V3	EAMB-V5	CVS-V3	CVS-V5	PCFS-V3	PCFS-V5	Fed-FiS	FPSO-FS	FedCIFL (Ours)
Accuracy (↑)	DEK→B	**73.40±0.54**	67.40±1.86	62.45±3.27	59.45±1.73	67.65±1.62	65.40±2.27	70.05±1.57	65.80±1.68	73.05±1.44
	BEK→D	75.89±1.46	69.52±1.13	67.62±1.88	60.35±1.16	69.47±1.86	66.02±2.78	72.48±1.50	76.29±0.83	**79.85±2.02**
	BDK→E	77.24±2.40	72.33±1.92	73.28±2.33	59.80±1.85	70.33±1.66	67.62±2.54	78.40±1.29	77.39±2.00	**83.61±2.57**
	BDE→K	79.20±2.14	72.28±1.89	74.94±1.05	58.55±2.35	75.84±2.50	69.12±2.19	82.56±1.77	79.80±2.60	**84.01±2.32**
RMSE (↓)	DEK→B	**0.436±0.00**	0.457±0.01	0.533±0.02	0.503±0.01	0.485±0.01	0.469±0.01	0.507±0.01	0.545±0.01	0.482±0.01
	BEK→D	0.415±0.01	0.442±0.01	0.496±0.01	0.493±0.00	0.467±0.01	0.466±0.01	0.471±0.02	0.434±0.01	**0.412±0.02**
	BDK→E	0.393±0.02	0.426±0.01	0.451±0.02	0.489±0.01	0.450±0.01	0.460±0.01	0.407±0.01	0.414±0.02	**0.362±0.03**
	BDE→K	0.377±0.01	0.425±0.00	0.431±0.00	0.487±0.00	0.412±0.02	0.457±0.01	0.361±0.01	0.398±0.03	**0.358±0.02**
F1 (↑)	DEK→B	**74.49±0.81**	69.03±1.07	62.05±3.83	52.94±2.96	68.64±1.73	64.86±1.63	65.00±2.59	57.49±3.31	69.36±2.38
	BEK→D	75.85±1.53	70.14±1.43	67.21±3.14	53.82±3.12	68.34±3.01	65.22±2.77	73.18±1.93	75.74±0.84	**79.78±2.31**
	BDK→E	76.92±2.70	73.11±2.47	72.82±3.18	63.57±2.22	70.24±2.05	62.37±3.11	77.32±1.91	77.08±2.60	**83.01±2.86**
	BDE→K	80.04±2.32	73.56±1.86	75.75±1.34	61.41±1.47	76.60±2.69	66.65±1.54	83.11±1.51	80.08±2.36	**84.48±2.21**

Table 1: Accuracy (%), RMSE, and F1 score (%) of the 4 cross-domain tasks on the Amazon Review dataset.

lection methods widen further, becoming more pronounced as the number of clients and data dimensions d increase. The stable performance exhibited by FedCIFL further demonstrates that it accurately estimates the causal effects between features and labels even in complex FL scenarios with limited samples, enabling the selection of causally invariant features and achieving strong generalization.

4.3 Results and Discussion (Real-World Data)

The experimental results on the Amazon Review dataset using the MLP classifier, as presented in Table 1, demonstrate the superiority of FedCIFL in learning causally invariant features for improved cross-domain generalization. It outperforms all baselines on most cross-domain tasks, including state-of-the-art FFS methods and causal feature selection methods. The superior performance of FedCIFL can be attributed to its ability to effectively capture the underlying causal relationships between features and labels, while mitigating the impact of data heterogeneity and distribution shift. The satisfactory performance of FedCIFL across different cross-domain tasks highlights its robustness to various domain adaptation scenarios in real-world applications.

4.4 Ablation Study

To validate the effectiveness of each module in FedCIFL, we conduct extensive ablation experiments. Specifically, we develop three variants of FedCIFL: “FedCIFL w/o iter”, “FedCIFL w/o SAE” and “FedCIFL w/o weighting”. “FedCIFL w/o iter” represents a variant of FedCIFL that does not employ the iterative strategy to optimize the confounder set and instead executes Steps 1 to 4 of FedCIFL only once. “FedCIFL w/o SAE” denotes a variant of FedCIFL which does not utilize the supervised autoencoder to learn a low-dimensional representation space for balancing the sample distribution between the treatment and control groups, but instead directly balances the sample distribution in the original feature space. “FedCIFL w/o weighting” refers to a variant of FedCIFL which does not employ the highly privacy-preserving weighted voting strategy based on Eq. (7) to resolve conflicts arising from the presence of multiple modes.

We then compare FedCIFL with these three variants under the synthetic Non-IID+OOD scenario (i.e., Scenario 4 in Figure 1). The results are presented in Figure 5. It can be observed that FedCIFL outperforms these three variants for all metrics across different data dimensions. This finding indicates that each key module in FedCIFL is effective and necessary for the FFS task.

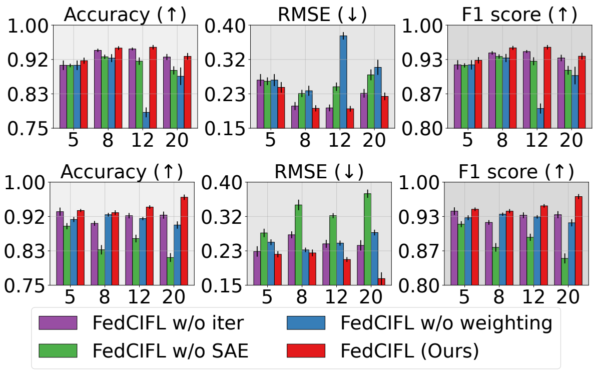

Figure 5: Experimental results of ablation experiments. The figure shows results for dimensions $d = 40$ and $d = 60$ (from top to bottom), with client counts on the x-axis.

5 Conclusions and Future Work

In this paper, we proposed FedCIFL, a novel federated causally invariant feature learning approach that addresses the challenges of data heterogeneity and OOD generalization in FL settings. At its core are a sample reweighting strategy and iterative refinement of the confounding feature set to identify true confounders, mitigating the impact of limited local data on the accuracy of federated causal effect estimation. Extensive experiments on synthetic and real-world datasets demonstrate the superiority of FedCIFL against state-of-the-art baselines, outperforming them in most cases in terms of average test accuracy, RMSE, and F1 score across various FL scenarios. To the best of our knowledge, FedCIFL is the first federated feature selection approach capable of handling Non-IID and OOD data simultaneously, achieving strong generalization ability and interoperability. In future work, we plan to extend FedCIFL to handle multi-label classification tasks, thereby broadening its applicability to a wider range of real-world scenarios.

Acknowledgments

X. Guo and K. Yu are supported by the National Science and Technology Major Project of China (Grant No. 2021ZD0111801) and the National Natural Science Foundation of China (Grant No. 62376087). L. Cui is supported by NSFC No. 92367202. H. Yu and X. Li are supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1. The research is also supported, in part, by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award No. I2301E0026) administered by A*STAR, as well as supported by Alibaba Group and Nanyang Technological University Singapore through the Alibaba-NTU Global e-Sustainability CorpLab (ANGEL); and the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No. AISG2-RP-2020-019).

References

Athey, S.; Imbens, G. W.; and Wager, S. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4): 597–623.

Banerjee, S.; Bhuyan, D.; Elmroth, E.; and Bhuyan, M. 2024. Cost-Efficient Feature Selection for Horizontal Federated Learning. *IEEE Transactions on Artificial Intelligence*.

Banerjee, S.; Elmroth, E.; and Bhuyan, M. 2021. Fed-FiS: A novel information-theoretic federated feature selection for learning stability. In *International Conference on Neural Information Processing*, 480–487. Springer.

Cai, R.; Huang, Z.; Chen, W.; Hao, Z.; and Zhang, K. 2023. Causal discovery with latent confounders based on higher-order cumulants. In *International conference on machine learning*, 3380–3407. PMLR.

Cassará, P.; Gotta, A.; and Valerio, L. 2022. Federated feature selection for cyber-physical systems of systems. *IEEE Transactions on Vehicular Technology*, 71(9): 9937–9950.

Guo, S.; Zhang, T.; Yu, H.; Xie, X.; Ma, L.; Xiang, T.; and Liu, Y. 2021. Byzantine-resilient decentralized stochastic gradient descent. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 4096–4106.

Guo, X.; Yu, K.; Cao, F.; Li, P.; and Wang, H. 2022a. Error-aware Markov blanket learning for causal feature selection. *Information Sciences*, 589: 849–877.

Guo, X.; Yu, K.; Liu, L.; Cao, F.; and Li, J. 2022b. Causal feature selection with dual correction. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1): 938–951.

Guo, X.; Yu, K.; Liu, L.; and Li, J. 2024a. FedCSL: A Scalable and Accurate Approach to Federated Causal Structure Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12235–12243.

Guo, X.; Yu, K.; Wang, H.; Cui, L.; Yu, H.; and Li, X. 2024b. Sample Quality Heterogeneity-aware Federated Causal Discovery through Adaptive Variable Space Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 4071–4079. ijcai.org.

Guo, Y.; Guo, K.; Cao, X.; Wu, T.; and Chang, Y. 2023. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, 11905–11933. PMLR.

Hermo, J.; Bolón-Canedo, V.; and Ladra, S. 2024. Fed-mRMR: A lossless federated feature selection method. *Information Sciences*, 669: 120609.

Hu, Y.; Zhang, Y.; Gao, X.; Gong, D.; Song, X.; Guo, Y.; and Wang, J. 2023. A federated feature selection algorithm based on particle swarm optimization under privacy protection. *Knowledge-Based Systems*, 260: 110122.

Hu, Y.; Zhang, Y.; Gong, D.; and Sun, X. 2022. Multi-participant federated feature selection algorithm with particle swarm optimization for imbalanced data under privacy protection. *IEEE Transactions on Artificial Intelligence*.

Huang, W.; and Wu, X. 2024. Robustly Improving Bandit Algorithms with Confounded and Selection Biased Offline Data: A Causal Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18, 20438–20446.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Khaire, U. M.; and Dhanalakshmi, R. 2022. Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, 34(4): 1060–1073.

Kuang, K.; Cui, P.; Athey, S.; Xiong, R.; and Li, B. 2018. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1617–1626.

Kuang, K.; Cui, P.; Li, B.; Jiang, M.; and Yang, S. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 265–274.

Kuang, K.; Wang, H.; Liu, Y.; Xiong, R.; Wu, R.; Lu, W.; Zhuang, Y.; Wu, F.; Cui, P.; and Li, B. 2023. Stable Prediction With Leveraging Seed Variable. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6392–6404.

Li, Z.; Wu, X.; Pan, W.; Ding, Y.; Wu, Z.; Tan, S.; Xu, Q.; Yang, Q.; and Ming, Z. 2024. FedCORE: Federated Learning for Cross-Organization Recommendation Ecosystem. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 3817–3831.

Ren, C.; Yu, H.; Peng, H.; Tang, X.; Li, A.; Gao, Y.; Tan, A. Z.; Zhao, B.; Li, X.; Li, Z.; et al. 2024. Advances and open challenges in federated learning with foundation models. *arXiv preprint arXiv:2404.15381*.

Rubin, D. B. 1973. Matching to remove bias in observational studies. *Biometrics*, 159–183.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32.

Sun, B.; Huo, H.; Yang, Y.; and Bai, B. 2021. Partialfed: Cross-domain personalized federated learning via partial initialization. *Advances in Neural Information Processing Systems*, 34: 23309–23320.

Wang, J.; Feng, W.; Chen, Y.; Yu, H.; Huang, M.; and Yu, P. S. 2018. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, 402–410.

Xiao, L.; Wu, X.; Xu, J.; Li, W.; Jin, C.; and He, L. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, 106: 102304.

Xiao, L.; Zhou, E.; Wu, X.; Yang, S.; Ma, T.; and He, L. 2022. Adaptive multi-feature extraction graph convolutional networks for multimodal target sentiment analysis. In *2022 IEEE International Conference on Multimedia and Expo*, 1–6. IEEE.

Yang, Q.; Fan, L.; and Yu, H. 2020. *Federated learning: Privacy and incentive*, volume 12500. Springer Nature.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2): 1–19.

Yang, S.; Guo, X.; Yu, K.; Huang, X.; Jiang, T.; He, J.; and Gu, L. 2023. Causal feature selection in the presence of sample selection bias. *ACM Transactions on Intelligent Systems and Technology*, 14(5): 1–18.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–46.

Yuan, H.; Morningstar, W. R.; Ning, L.; and Singhal, K. 2022. What Do We Mean by Generalization in Federated Learning? In *International Conference on Learning Representations*.

Zhang, X.; Mavromatis, A.; Vafeas, A.; Nejabati, R.; and Simeonidou, D. 2023. Federated feature selection for horizontal federated learning in IoT networks. *IEEE Internet of Things Journal*, 10(11): 10095–10112.