Supplementary Materials

Liping Chen[1], Guangqing Bai[2], Zhuoyang Han[2], Jing Ren[1], Mujie Liu[3], Tao Tang[4(✉)], Shuo Yu[2], and Ivan Lee[4]

[1] RMIT University, Melbourne, Australia
[2] Dalian University of Technology, Dalian, China
[3] Federation University Australia, Ballarat, Australia
[4] University of South Australia, Adelaide, Australia
{lp.chen, zhuoyang.han, jing.ren}@ieee.org
{mujie.liu, tao.tang, shuo.yu}@ieee.org
baigq@mail.dlut.edu.cn, Ivan.Lee@unisa.edu.au

1 Preliminaries

1.1 Graph Contrastive Learning

Graph contrastive learning was proposed to pre-train graph representations in a self-supervised manner [9]. The process can be roughly introduced as follows.

Given an input graph $\mathcal{G}$, two correlated views $\mathcal{G}'_i$ and $\mathcal{G}'_j$ are generated via stochastic augmentation functions $q_i(\cdot|\mathcal{G})$ and $q_j(\cdot|\mathcal{G})$, forming a positive pair. Then, a graph neural network encoder Encoder$(\cdot)$ is used to extract graph-level representations $\mathbf{h}_i$ and $\mathbf{h}_j$ from the augmented graphs $\mathcal{G}'_i$ and $\mathcal{G}'_j$, respectively. It is noted that there are no architectural constraints on Encoder$(\cdot)$. Subsequently, the representation vectors are further transformed by a non-linear projection head $g(\cdot)$ to obtain the contrastive embeddings $\mathbf{z}_i = g(\mathbf{h}_i)$ and $\mathbf{z}_j = g(\mathbf{h}_j)$. Following common practice, we employ a two-layer MLP as the projection head.

Finally, a normalized temperature-scaled cross-entropy loss (NT-Xent) is applied to maximize agreement between positive pairs while contrasting them against negatives within the batch. For a minibatch of N graphs, each graph is augmented twice, resulting in $2N$ views. Let $\mathbf{z}_{n,i}, \mathbf{z}_{n,j}$ denote the two views of the n-th graph. The loss for graph n is given by:

$$\ell_n = -\log \frac{\exp\left(\mathrm{s}(\mathbf{z}_{n,i}, \mathbf{z}_{n,j})/\tau\right)}{\sum_{n'=1,\ n'\neq n}^{N} \exp\left(\mathrm{s}(\mathbf{z}_{n,i}, \mathbf{z}_{n',j})/\tau\right)}, \tag{1}$$

where $\mathrm{s}(\cdot,\cdot)$ is the cosine similarity function, and τ is a temperature parameter.

1.2 Federated Learning

Federated learning was proposed by Google, aims to address the privacy issues of data sharing and the data silo challenge. In the central server, the federated averaging (FedAvg) algorithm is commonly used to weight the updated parameters based on the varied scale of client-side data [5].

Specifically, in each communication round of the FedAvg process, the following steps are performed. Given K clients, a batch size B, and an initialized global model weight $\mathbf{w}^0$, the server selects a subset of clients $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$, where $k \leq K$ and each client $S_i \in \mathcal{S}$ holds a private dataset D_i.

The server then distributes the current global model $\mathbf{w}^t$ to all selected clients. Each selected client performs local training on its dataset and computes an updated local model $\mathbf{w}_i^t$. The local training process is formulated in Equation (2).

$$\mathbf{w}_i^t = \mathbf{w}^t - \eta \nabla \mathcal{L}(\mathbf{w}^t, D_i), \tag{2}$$

where $\nabla \mathcal{L}$ denotes the gradient of the loss function $\mathcal{L}$ with respect to model parameters, and η is the learning rate.

The server then aggregates the updated local models from selected clients to obtain the new global model $\mathbf{w}^{t+1}$ by Equation (3).

$$\mathbf{w}^{t+1} = \frac{1}{\sum_{j \in \mathcal{S}} |D_j|} \sum_{i \in \mathcal{S}} |D_i| \mathbf{w}_i^t, \tag{3}$$

where $\mathbf{w}_i^t$ denotes the local model of client S_i at round t, and $|D_i|$ is the number of local samples on client S_i. The global model $\mathbf{w}^{t+1}$ is updated as a weighted average of client models, where the contribution of each client is proportional to its data size.

2 Local Training Objectives

On each selected client $S_i \in \mathcal{S}_{\text{low}}$, the local training objective integrates three components: unsupervised contrastive loss, supervised contrastive loss, and classification loss. These losses are computed based on patient node embeddings obtained from the local bipartite graph.

(1) Unsupervised Contrastive Loss. We encourage consistency between representations derived from the original and augmented graphs. Let $\mathbf{h}_{v_i}$ denote the embedding of patient node $v_i \in \mathcal{V}_P$ from the original bipartite graph $\mathcal{G}$, and let $\mathbf{h}'_{v_q}$ be the corresponding embedding of patient node $v_q \in \mathcal{V}_P$ from the augmented graph $\mathcal{G}'$, which is generated by random edge dropout.

The unsupervised contrastive loss aims to align embeddings of the same patient while distinguishing them from others:

$$\mathcal{L}_{\text{ucl}} = -\frac{1}{|\mathcal{V}_P|} \sum_{v_i \in \mathcal{V}_P} \log \frac{\exp\left(s(\mathbf{h}_{v_i}, \mathbf{h}'_{v_i})/\tau\right)}{\sum_{v_q \in \mathcal{V}_P} \exp\left(s(\mathbf{h}_{v_i}, \mathbf{h}'_{v_q})/\tau\right)}, \tag{4}$$

where $s(\cdot, \cdot)$ denotes cosine similarity, and τ is the temperature parameter.

(2) Supervised Contrastive Loss. To promote label-aware clustering, we construct positive pairs from patient nodes that share the same ground-truth label. Let $\mathcal{P}_{\text{sup}} \subseteq \mathcal{V}_P \times \mathcal{V}_P$ denote the set of supervised positive pairs, where $(v_i, v_q) \in \mathcal{P}_{\text{sup}}$ if and only if $y_i = y_q$ and $v_i \neq v_q$.

For each anchor node v_i, we treat all other patient nodes with different labels (i.e., $y_j \neq y_i$) as negatives. The supervised contrastive loss is then defined by Equation (5).

$$\mathcal{L}_{\text{scl}} = -\frac{1}{|\mathcal{P}_{\text{sup}}|} \sum_{(v_i, v_q) \in \mathcal{P}_{\text{sup}}} \log \frac{\exp\big(s(\mathbf{h}_{v_i}, \mathbf{h}_{v_q})/\tau\big)}{\sum_{v_j} \exp\big(s(\mathbf{h}_{v_i}, \mathbf{h}_{v_j})/\tau\big)}. \tag{5}$$

It is noted that the denominator sums over all patient nodes v_j with $y_j \neq y_i$, i.e., nodes with labels different from that of anchor v_i.

(3) Classification Loss via Decoder. For downstream disease prediction, we use MLP as the decoder to map each patient embedding to a predicted label by Equation (6).

$$\widehat{y}_i = \text{Decoder}(\mathbf{h}_{v_i}). \tag{6}$$

The classification loss is computed as the average cross-entropy over labeled patient nodes by Equation (7).

$$\mathcal{L}_{\text{CE}} = \frac{1}{|\mathcal{V}_P|} \sum_{v_i \in \mathcal{V}_P} \text{CrossEntropy}(\widehat{y}_i, y_i), \tag{7}$$

where $\widehat{y}_i$ is the predicted label, and y_i is the ground-truth label of patient node v_i.

(4) Overall Objective. Each client S_i minimizes a local objective over its private dataset D_i, defined as Equation (8).

$$\begin{aligned} \mathcal{L}_i(\mathbf{w}) = \mathbb{E}_{(v_i, y_i) \sim D_i} \big[& \lambda_{\text{ucl}} \, \mathcal{L}_{\text{ucl}}(v_i) + \lambda_{\text{scl}} \, \mathcal{L}_{\text{scl}}(v_i) + \\ & \lambda_{\text{CE}} \, \mathcal{L}_{\text{CE}}(v_i, y_i) \big], \end{aligned} \tag{8}$$

where each term denotes the contrastive or classification loss computed for patient node v_i, and $\lambda_{\text{ucl}}, \lambda_{\text{scl}}, \lambda_{\text{CE}}$ are non-negative weighting coefficients.

3 Experiments

3.1 Datasets

Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] is a longitudinal database of elderly individuals, including cognitively normal controls, patients with mild cognitive impairment (MCI), and Alzheimer's disease (AD) patients. We use curated ADNI data from the TADPOLE challenge [3], covering structural MRI, FDG-PET, diffusion tensor imaging (DTI), CSF biomarkers, and clinical assessments. The TADPOLE challenge dataset contains 1,737 patients and their corresponding 12,741 visits.

MIMIC-IV [2] is a publicly available electronic health record (EHR) database collected from ICU and emergency department admissions. MIMIC-IV contains 180,000 patients and their corresponding 431,000 admissions. In our experiment, we take the modalities from demographics, diagnosis, procedure, medication, lab

tests, and clinical notes as the input. We extract structured data modalities including diagnoses (ICD codes), procedures, prescriptions, and laboratory tests. Following prior medical categorization [6], we select 111 lab items grouped into ten physiological categories such as hematologic, metabolic, and electrolyte panels. We retain ICU stays of adult patients (ages 18–89), excluding admissions exceeding 10 days or ending in in-hospital death.

In our experiment, we divided each dataset into training, validation, and test sets in a 70%:10%:20% ratio, respectively.

3.2 Task Settings and Evaluation Metrics

Task Settings. We focus on three common prediction tasks in healthcare: Both the mortality prediction task and the hospital readmission prediction task are based on MIMIC-IV data, and the Alzheimer's disease progression prediction task is based on ADNI data.

1) Mortality Prediction. This task aims to predict whether a patient will be *deceased* within 90 days after hospital discharge.

2) Readmission Prediction. This task aims to predict whether a patient will be *readmitted* to the hospital within 30 days after discharge, which is a key indicator of care quality and healthcare cost.

3) Alzheimer's disease progression (ADP) Prediction. The goal of this task is to classify a patient's clinical status at a given visit into one of three categories: normal control (NC), mild cognitive impairment (MCI), or Alzheimer's disease (AD).

Evaluation Metrics. We evaluate model performance using standard metrics. For the mortality prediction task, we report the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PRC). For the Alzheimer's disease progression task, we report Accuracy and AUC-ROC scores.

AUC-PRC is defined as the area under the precision-recall curve:

$$\text{AUC-PRC} = \int_0^1 \text{P}(R)\, dR, \tag{9}$$

where precision P and recall R are computed as:

$$R = \frac{TP}{TP + FN}, \quad \text{P} = \frac{TP}{TP + FP}. \tag{10}$$

AUC-ROC is defined as the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR):

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR)\, dFPR, \tag{11}$$

where $\text{TPR} = \frac{TP}{TP+FN}$ and $\text{FPR} = \frac{FP}{FP+TN}$.

Accuracy (ACC) measures the proportion of correctly predicted instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{12}$$

Here, TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3.3 Baselines

We compared the proposed model with the state-of-the-art baselines as follows.

- **Cafe** [4]: A federated imputation framework that leverages client-side missing data heterogeneity. CAFE clusters clients with similar missingness patterns and learns cluster-specific data imputers, which are then used to complete missing features before local model training. It enhances downstream performance by reducing the mismatch caused by diverse missing patterns.
- **PEARL** [7]: A personalized federated learning framework for disease prediction using Non-IID EHRs. PEARL leverages attention-based encoders to capture diagnosis- and admission-level semantics from each client's local data. It integrates self-supervised learning with federated training to learn a robust global model, followed by local fine-tuning for personalization. A differential privacy mechanism is further adopted to protect sensitive information during model aggregation.
- **M3Care** [10]: An end-to-end multimodal EHR analysis model designed to handle missing modalities by imputing task-relevant latent representations rather than reconstructing raw inputs. It leverages a task-guided, modality-adaptive similarity metric to retrieve semantically similar patients with shared observed modalities, enabling effective downstream prediction even under severe modality missingness. In our experiments, we adapt M3Care to the federated learning setting using the FedAvg protocol.

In addition to the main experiments, we also investigate the impact of differential privacy (DP) by applying DP-SGD during global updates to assess model robustness under stricter privacy constraints.

3.4 Additional Hyperparameter Analysis of θ.

Table 1 reports the AUC-PRC performance of each client on the ADP task under Case 2 settings with varying θ values. We observe that clients with lower missingness rates (e.g., Client-7 and Client-8 with $r = 0.6$) consistently achieve better performance across all thresholds. As missingness increases, the model's performance generally degrades, highlighting the challenge of high data incompleteness.

Interestingly, while the average AUC-PRC across all clients peaks at $\theta = 0.9$, the optimal θ varies slightly across individual clients. For example, Client-2 ($r = 0.9$) benefits from $\theta = 0.8$ or $\theta = 0.9$, while lower-missingness clients are relatively robust across thresholds.

Table 1: **AUC-PRC** results of MissPCL under varying threshold θ and missingness rates r by Case 2 clients in the ADP task.

Client	r	$\theta = 1.0$	$\theta = 0.9$	$\theta = 0.8$	$\theta = 0.7$
Client-1	0.9	0.4952	0.4745	0.4918	0.5006
Client-2	0.9	0.4661	0.5417	0.5419	0.5389
Client-3	0.8	0.6070	0.6111	0.6021	0.5891
Client-4	0.8	0.6345	0.6228	0.6024	0.6104
Client-5	0.7	0.6673	0.6480	0.6472	0.6431
Client-6	0.7	0.6202	0.6074	0.5961	0.6020
Client-7	0.6	0.7011	0.7103	0.7004	0.6917
Client-8	0.6	0.7498	0.7460	0.7417	0.7224
Average	–	0.6176	**0.6202**	0.6155	0.6123

Table 2: MissPCL performance under different privacy budgets (ϵ).

Privacy Budget	**Dataset**	**Task**	AUROC	AUPRC
$\epsilon = 1$	ADNI	ADP	0.5939	–
$\epsilon = 2$	ADNI	ADP	0.6216	–
$\epsilon = 3$	ADNI	ADP	0.6194	–
$\epsilon = 4$	ADNI	ADP	0.6288	–
$\epsilon = 1$	MIMIC-IV	Readmission	0.6446	0.3830
$\epsilon = 2$	MIMIC-IV	Readmission	0.6499	0.3892
$\epsilon = 3$	MIMIC-IV	Readmission	0.6514	0.3898
$\epsilon = 4$	MIMIC-IV	Readmission	0.6523	0.3903
$\epsilon = 1$	MIMIC-IV	Mortality	0.8206	0.3620
$\epsilon = 2$	MIMIC-IV	Mortality	0.8274	0.3708
$\epsilon = 3$	MIMIC-IV	Mortality	0.8279	0.3706
$\epsilon = 4$	MIMIC-IV	Mortality	0.8286	0.3701

3.5 Hyperparameter Analysis of Privacy Budget ϵ.

To address privacy concerns in federated learning, we further extend our framework with differential privacy, a widely adopted technique that provides formal guarantees against data leakage [8,7]. We apply DP-SGD during global model aggregation and perform ablation studies under varying privacy budgets $\epsilon \in \{1, 2, 3, 4\}$, following the Case 1 missingness setting. The results are summarized in Table 2. As expected, a smaller ϵ corresponds to stronger privacy protection. Notably, our model maintains competitive performance even under strong privacy constraints, demonstrating the robustness of MissPCL in privacy-sensitive federated healthcare scenarios.

References

1. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
2. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023)
3. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Golland, P., Klein, S., et al.: Tadpole challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data. In: Predictive Intelligence in Medicine: Second International Workshop. pp. 1–10. Springer (2019)
4. Min, S., Asif, H., Wang, X., Vaidya, J.: Cafe: Improved federated data imputation by leveraging missing data heterogeneity. IEEE Transactions on Knowledge and Data Engineering **37**(5), 2266–2281 (2025)
5. Nilsson, A., Smith, S., Ulm, G., Gustavsson, E., Jirstrand, M.: A performance evaluation of federated learning algorithms. In: Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning. pp. 1–8 (2018)
6. Sung, M., Hahn, S., Han, C.H., Lee, J.M., Lee, J., Yoo, J., Heo, J., Kim, Y.S., Chung, K.S.: Event prediction model considering time and input error using electronic medical records in the intensive care unit: Retrospective study. JMIR Medical Informatics **9**(11), e26426 (2021)
7. Tang, T., Han, Z., Cai, Z., Yu, S., Zhou, X., Oseni, T., Das, S.K.: Personalized federated graph learning on non-iid electronic health records. IEEE Transactions on Neural Networks and Learning Systems **35**(9), 11843–11856 (2024)
8. Wei, K., Li, J., Ma, C., Ding, M., Chen, W., Wu, J., Tao, M., Poor, H.V.: Personalized federated learning with differential privacy and convergence guarantee. IEEE Transactions on Information Forensics and Security **18**, 4488–4503 (2023)
9. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y.: Graph contrastive learning with augmentations. In: Advances in Neural Information Processing Systems. vol. 33, pp. 5812–5823 (2020)
10. Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., Zhao, J.: M3care: Learning with missing modalities in multimodal healthcare data. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 2418–2428 (2022)