Vertical Federated Knowledge Transfer via Representation Distillation for Healthcare Collaboration Networks

Chung-ju Huang
chongruhuang.pku@gmail.com
Key Laboratory of High Confidence
Software Technologies (Peking
University), Ministry of Education
Beijing, China
School of Computer Science, Peking
University
Beijing, China

Leye Wang
leyewang@pku.edu.cn
Key Laboratory of High Confidence
Software Technologies (Peking
University), Ministry of Education
Beijing, China
School of Computer Science, Peking
University
Beijing, China

Xiao Han*
xiaohan@mail.shufe.edu.cn
School of Information Management
and Engineering, Shanghai University
of Finance and Economics
Shanghai, China

ABSTRACT

Collaboration between healthcare institutions can significantly lessen the imbalance in medical resources across various geographic areas. However, directly sharing diagnostic information between institutions is typically not permitted due to the protection of patients' highly sensitive privacy. As a novel privacy-preserving machine learning paradigm, federated learning (FL) makes it possible to maximize the data utility among multiple medical institutions. These feature-enrichment FL techniques are referred to as vertical FL (VFL). Traditional VFL can only benefit multi-parties' shared samples, which strongly restricts its application scope. In order to improve the information-sharing capability and innovation of various healthcare-related institutions, and then to establish a next-generation open medical collaboration network, we propose a unified framework for **v**ertical **fed**erated knowledge **trans**fer mechanism (VFedTrans) based on a novel cross-hospital representation distillation component. Specifically, our framework includes three steps. First, shared samples' *federated representations* are extracted by collaboratively modeling multi-parties' joint features with current efficient vertical federated representation learning methods. Second, for each hospital, we learn a *local-representation-distilled module*, which can transfer the knowledge from shared samples' federated representations to enrich local samples' representations. Finally, each hospital can leverage local samples' representations enriched by the distillation module to boost arbitrary downstream machine learning tasks. The experiments on real-life medical datasets verify the knowledge transfer effectiveness of our framework.

CCS CONCEPTS

• **Computing methodologies → Knowledge representation and reasoning**; • **Applied computing → Health care information systems**.

---

*Corresponding author

KEYWORDS

vertical federated learning, healthcare collaboration network, knowledge distillation, representation learning

1 INTRODUCTION

Currently, the disparity in healthcare resources [43] continues to be a significant challenge for both developed and developing countries. For a myriad of reasons, there are huge differences in the healthcare resources accessible to distinct areas, classes, and ethnicities even within the same country [2, 39]. After Covid-19 officially became a pandemic, it overwhelmed healthcare facilities in less developed areas due to the severity of clinical symptoms and the unpredictability of post-recovery sequelae [34]. The supply-demand conflict of unbalanced healthcare resources is thus rapidly increasing, which will affect the sustainability of the healthcare system [10] and the viability of health policy reform [11]. In order to promote social equality and social justice, it is crucial from a strategic perspective to address the disparity in healthcare resources [24, 55]. In this paper, we will focus on the secure utilization of adequate medical data from developed regions to make up for inadequate and incomplete hospital data from lagging regions.

Unlike other fields, medical data contains many of the most private details of patients' personal lives, psychological conditions, social relationships, and financial situations, making it particularly sensitive to privacy [1, 26, 32, 35]. The disclosure of such identifiable privacy about individuals can greatly damage the level of public trust in healthcare institutions. Therefore, no institution will be permitted to provide patient information to another institution directly. With the enactment of the EU General Data Protection Regulation (GDPR)[1] and US Health Insurance Portability and Accountability Act (HIPAA)[2], access to and use of private data has been further restricted. How to conduct machine learning and data mining in a privacy-preserving and law-regulated way has attracted much interest in both academia and industry. Federated learning (FL) [36]

---

[1] https://gdpr-info.eu/
[2] https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

has thus become a promising solution [28, 54]. In general, FL does not need different parties to exchange their raw data; instead, every party runs local computation and training on their own data and then uploads the intermediate results (e.g., gradients) to a server. By integrating these intermediate results from all the parties, a federated global model can be learned. Especially, such a federated model can achieve similar prediction performance as a centralized model directly trained on all the parties' data [51].

In general, there are two main types of FL algorithms, *horizontal* and *vertical*. The first FL algorithm proposed by Google is horizontal [36]; the setting is that different parties (often devices) hold different samples with the same features or data formats. A representative application of horizontal FL is the mobile phone keyboard next-word prediction, where a global next-word prediction model can be learned without collecting users' raw keyboard inputs [53]. In contrast, vertical FL's (VFL) setting is that different parties (often organizations) hold different features of the same set of samples. This work focuses on the vertical setting.

The successful adoption of current VFL methods is highly dependent on how many overlapped samples exist between parties. Hence, most VFL collaborations are conducted by involving at least one giant data holder with abundant data. For instance, FDN (federated data network) [29] includes anonymous data from one of the largest social network service providers in China and thus can cover most user samples from other data holders (e.g., customers of banks). However, this makes giant data holders occupy a dominant position over other small data holders in VFL, which could lead to unfair trades and data monopoly in the digital economy.[3] Collaborative healthcare network [45] is composed of hospitals with multiple locations and different medical resources. In this scenario, the characteristics and attributes of smaller hospitals' own data are often overlooked when larger hospitals with more data dominate the collaboration. The capacity of the hospitals receiving assistance to use the local data's specificity to give patients more individualized treatment is severely hampered by this. For patients, they may go to different hospitals for the same disease. Differences in the level of care at the hospital will affect the analysis and diagnosis of the disease. High level hospitals tend to detect more hidden symptoms, i.e., have richer sample features. However, these medical records can only be kept in multiple locations. This leads to the fact that in the traditional VFL, patients can only get better joint services if they have been to multiple hospitals. Patients who are limited to a few or even one hospital due to region, race, etc. do not have access to equitable medical resources. Attention and protection for this group are crucial and necessary. To alleviate this pitfall and expand application scenarios, *a VFL-based collaborative framework that can benefit various ordinary hospitals and vulnerable populations is urgently needed.*

As a pioneering attempt in this direction, this paper proposed a novel vertical-federated-knowledge-transfer (VFedTrans) unified framework that can transfer the medical knowledge from (a limited number of) collaborative healthcare networks' shared samples to each hospital's local (non-shared) samples. The key challenge is *how to fill the gap between hospital's local samples (with only this*

hospital's features) and cross-hospital shared samples (with multiple hospitals' features). To address this issue, we propose a novel local-representation-distilled module that can distill the knowledge from shared samples' federated representations to enrich local samples' representations. More specifically, shared samples' federated representations are first learned by some federated latent representation extraction methods (e.g., federated singular vector decomposition [3]); then, the small hospital can leverage shared samples' federated representation as the guidance to enrich its local samples' feature representation via a knowledge distilling strategy [18]. Especially, our knowledge transfer mechanism has the following characteristics.

- *Knowledge transfer to local samples.* Different from most VFL algorithms focusing on shared samples, our mechanism aims to improve the learning performance on different parties' local samples via vertical knowledge transfer. In this way, a set of hospitals with only a limited number of shared samples can still benefit from our VFL process.
- *Task-independent transfer.* Our knowledge transfer process is task-independent. That is, each hospital can leverage its enriched local samples' representations for an arbitrary (new) medical task.
- *Scalable to multiple hospitals.* The complexity of our mechanism is linearly proportional to the number of involved hospitals. More importantly, our mechanism can be learned in an online manner. That is, when a new hospital comes, existing hospitals can efficiently update their local sample representations by just learning with the new hospitals.

In summary, this work makes the following contributions:

(1) To the best of our knowledge, this work is the first one to explore how to enable vertical knowledge transfer from shared samples to each hospital's local samples in a task-independent manner.
(2) We propose a novel *federated-representation-distilled* framework, *VFedTrans*, to transfer medical knowledge from shared samples to local samples. VFedTrans includes the following steps. First, a federated representation learning method is applied to extract shared samples' representations. Second, each hospital can enrich its local feature representation-distilled module by knowledge distilling on the shared samples' federated representations. The module can then be leveraged to enrich local samples' feature representations.
(3) Experiments on four real-life medical datasets have verified the effectiveness of our mechanism for knowledge transfer and the generalizability of our enriched feature representations of local samples. This demonstrates that VFedTrans enables hospitals with scarce medical resources to provide better medical services through VFL collaboration.

The source code for this work is available at: https://doi.org/10.5281/zenodo.7623519.

2 PROBLEM FORMULATION

In this section, we clarify the definitions of key concepts used in this paper. Afterward, we formulate our research problem. Appendix A.1 lists the notations used throughout this paper.

---

[3]https://www.theguardian.com/technology/2015/apr/19/google-dominates-search-real-problem-monopoly-data

2.1 Concepts

Our approach enables all hospitals to benefit from the collaboration. Without loss of generality, we classify them into two types of roles: the *task hospital* and the *data hospital*.

Task Hospital. A task hospital t has a set of samples with features X_t and a task label Y_t to predict. The sample IDs of the task hospital are denoted as I_t.

Data Hospital. A data hospital d has a set of samples with features X_d. The data hospital's sample IDs are denoted as I_d.

Remark. A hospital may play both task and data roles simultaneously in a VFL campaign (i.e., a hospital contributes its features to other hospitals and also benefits from other hospitals' features). Our mechanism can be efficiently applied to this case.

2.2 Research Problem

Local-Sample Vertical Federated Knowledge Transfer Problem. Given a task hospital t and n data hospitals $d_i (i = 1, 2, ..., n)$, t has certain shared samples with any data hospital d_i ($I_t \cap I_{d_i} \neq \phi$), the objective is to design a federated knowledge transfer algorithm to predict the task label Y_t of t's (non-shared) local samples as accurately as possible.

Remark. Traditional VFL problems often require that $I_t = I_{d_i}$. However, our vertical federated knowledge transfer setting only needs that $I_t \cap I_{d_i} \neq \phi$. Without loss of generality, we solely validate the impact of knowledge transfer on a task hospital to check that VFedTrans has strong service support for healthcare-related institutions with limited knowledge. Briefly, the objective of our proposed VFedTrans is to improve the task performance of t's local samples ($I_t \setminus I_{d_i}$) by transferring the knowledge from shared samples ($I_t \cap I_{d_i}$). This expands the practical application range of FL by complementing traditional VFL.

3 FRAMEWORK DESIGN

3.1 Overview

We demonstrate the overall process of VFedTrans (Fig. 1). Note that before our mechanism runs, we suppose that shared samples between the task hospital t and any data hospital d_i are known, which can be learned by PSI (private set intersection) methods [22]. Our framework can be simplified into three main steps (Fig. 2).

- **Step 1. Federated Representation Learning (FRL)**. First, the task hospital and data hospital collaboratively learn federated latent representations for shared samples using secure VFL techniques. In brief, these federated latent representations would incorporate the hidden knowledge among multiple parties while not leaking these parties' raw features.

- **Step 2. Local Representation Distillation (LRD)**. Second, the task hospital trains a *federated-representation-distilled module* that can distill the knowledge from shared samples' federated latent representations to enrich local samples' representations.

- **Step 3. Learning on Enriched Representations**. After Step 2, the FRD module is distilled and ready for local feature enrichment. Then, given an arbitrary label y_t to predict,

the task hospital can use local samples' enriched representations (i.e., task hospital's local features + enriched representations) to conduct training and inference with state-of-the-art (SOTA) machine learning algorithms.

Step 3 generally follows traditional supervised learning methods to train a task-specific medical prediction model, where various machine learning algorithms can be applied, such as random forest and neural networks. Next, we illustrate more details about Step 1 and 2. For brevity, we first assume that only one data hospital d exists. At the end of Sec. 3.3, we will discuss how to deal with multiple data hospitals $\{d_1, d_2, ..., d_n\}$.

3.2 Federated Representation Learning

The purpose of Step 1 is to extract latent representations of shared samples by considering both task and data hospitals' features. In general, we can adopt various vertical federated representation learning methods for this step. In this work, we adopt a matrix decomposition-based federated representation method, as literature has shown that matrix decomposition is effective for extracting meaningful latent representations for machine learning tasks [25]. Here, we introduce how to leverage two state-of-the-art federated matrix decomposition methods, i.e., FedSVD [3] and VFedPCA [8], to learn shared samples' federated representations by considering both task and data hospitals' features.

3.2.1 FedSVD. In FedSVD [3], all hospitals use two random orthogonal matrices to transform the local samples into local masked samples. This maintains the invariance of the decomposition results despite the masking transformation of the local samples. The masked samples are then uploaded to a third-party server, which applies the SVD algorithm to the samples from all hospitals. Finally, the task hospital can reconstruct the federated latent representation based on the decomposition results.

Suppose the task hospital holds the shared samples' feature matrix $S_t \in \mathbb{R}^{|I_s| \times |X_t|}$, and the data hospital d holds the shared samples' feature matrix $S_d \in \mathbb{R}^{|I_s| \times |X_d|}$ ($I_s = I_t \cap I_d$ is the shared sample ID set). Denote $S = [S_t | S_d]$ (combination of both task and data hospitals' feature matrices), we want to leverage $S = U\Sigma V^T$ (SVD) to learn the latent representations U, Inspired by FedSVD [3], we use a randomized masking method to learn U as,

(1) A trusted key generator generates two randomized orthogonal matrices $A \in \mathbb{R}^{|I_s| \times |I_s|}$ and $B \in \mathbb{R}^{|X_{td}| \times |X_{td}|}$ ($|X_{td}| = |X_t| + |X_d|$). B is further partitioned to two parts $B_t \in \mathbb{R}^{|X_t| \times |X_{td}|}$ and $B_d \in \mathbb{R}^{|X_d| \times |X_{td}|}$, i.e., $B^T = [B_t^T | B_d^T]$.

(2) A and B_t are sent to the task hospital; A and B_d are sent to the data hospital. Each hospital does a local computation by masking its own feature matrices with the received matrices:

$$\hat{S}_k = A S_k B_k, \forall k \in \{t, d\} \tag{1}$$

(3) Both task and data hospitals send $\hat{S}_t$ and $\hat{S}_d$ to a third-party server[4] and the third-party server runs SVD on the combined data matrix $\hat{S} = \hat{U}\Sigma\hat{V}^T$, where $\hat{S} = [\hat{S}_t | \hat{S}_d]$. $\hat{U}$ is then sent to the task hospital.

---

[4]The third-party server needs to be semi-honest. Note that in FL, such a security configuration (i.e., the information aggregation server is semi-honest) is widely accepted [51].
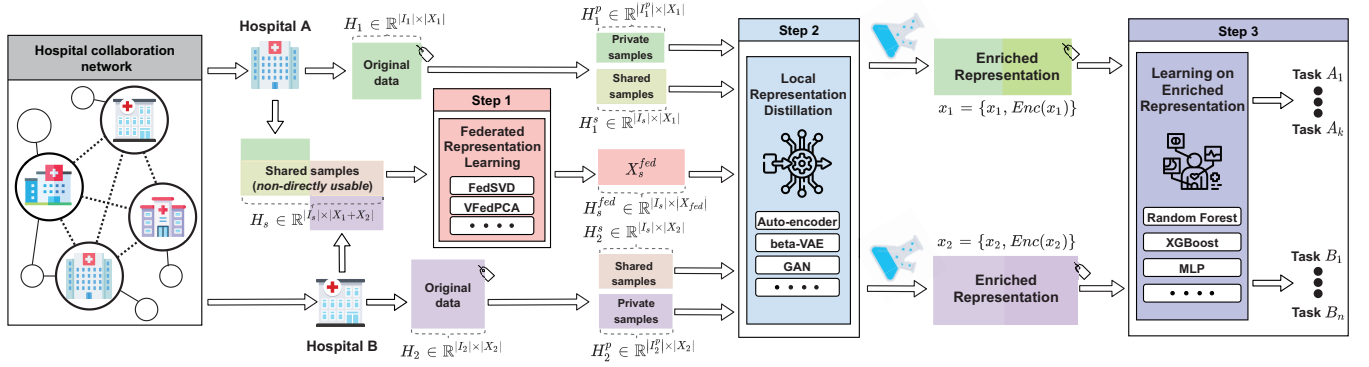
Figure 1: Flowchart of VFedTrans for two hospitals. Each hospital can be either the task or data party. For task $A_1, ..., A_k$, *Hospital A* is the task party and *Hospital B* is the data party; for task $B_1, ..., B_k$, *Hospital B* is the task party and *Hospital A* is the data party. In Step 1, we use the VFL techniques for federated representation learning on non-directly usable shared samples to get a federated latent representation x_s^{fed}. In Step 2, both hospitals are able to train a local-representation-distilled module for knowledge transfer on their own. For the shared samples, the loss function in knowledge distillation not only contains the reconstruction loss but also adds a new extra distillation loss term that we designed. Then we align the learned representations with the original data to obtain new local enriched representations. In Step 3, both hospitals are able to use the learned enriched representations to complete their respective downstream healthcare-related machine-learning tasks.
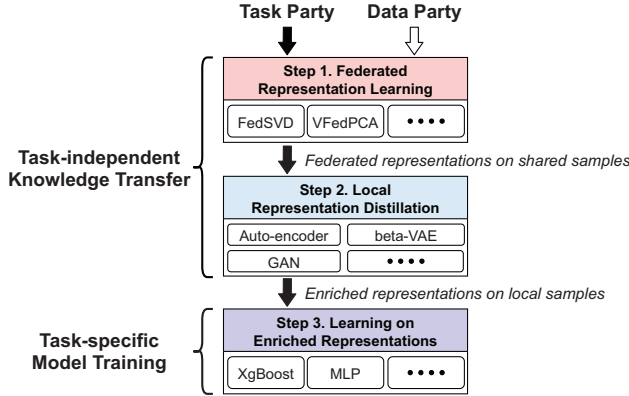

Figure 2: Overview of VFedTrans.

(4) The task hospital can recover the federated latent representation of shared samples, denoted as x_s^{fed}, by

$$x_s^{fed} = U = A^T \hat{U} \qquad (2)$$

Compared to the original FedSVD which aims to recover both U and V [3], we only need to recover U. Hence, in VFedTrans, only $\hat{U}$ is transmitted to the task hospital to reduce the communication cost. The correctness of the above process depends on the fact that S and $\hat{S}$ (multiplying S by two orthogonal matrices) must hold the same singular value Σ [3].

3.2.2 VFedPCA. To enhance the generality of VFedTrans, we also use vertical federated principal component analysis (VFedPCA) [8] to extract latent representations. Under VFedPCA's setting, each hospital makes its own federated eigenvector u converge to global eigenvector u_G without needing to know the mutual data of all hospitals. Each hospital is able to train the local eigenvector using local power iteration [42]. Then the eigenvectors from each hospital are merged into the federated eigenvector u. Finally, task hospital can use u to reconstruct the original data to obtain the federated latent representation.

Suppose the task hospital holds the shared samples' feature matrix $S_t \in \mathbb{R}^{|I_s| \times |X_t|}$, and the data hospital holds the shared

samples' feature matrix $S_d \in \mathbb{R}^{|I_s| \times |X_d|}$. Denote $S = [S_t | S_d], S \in \mathbb{R}^{|I_s| \times |X_t + X_d|}$.

(1) For each hospital $i \in \{t, d\}$, we calculate the largest eigenvalue $A_i = \frac{1}{|X_i|} S_i^T S_i$ and a non-zero vector a_i corresponding to the eigenvector $\alpha_i (A_i a_i = \alpha_i a_i)$. The number of local iterations is L, each hospital will compute locally until convergence as follows:

$$a_i^l = \frac{A_i a_i^{l-1}}{||A_i a_i^{l-1}||}, \quad \alpha_i^l = \frac{A_i (a_i^l)^T a_i^l}{(a_i^l)^T a_i^l} \qquad (3)$$

where $l = 1, 2, \cdots, L$.

(2) Then each hospital upload the eigenvector a_i^L and the eigenvalue α_i^L to third-party server. The server aggregates the results and generates the federated eigenvalue:

$$u = w_t a_t^L + w_d a_d^L, \quad w_i = \frac{\alpha_i^L}{\sum_{i \in \{t, d\}} \alpha_i^L} \qquad (4)$$

(3) Task hospital t can use the federated eigenvalue u to reach the federated latent representation:

$$x_s^{fed} = S_t \frac{MM^T}{||MM^T||}, M = S_t^T u \qquad (5)$$

3.3 Local Representation Distillation

After obtaining x_s^{fed} for shared samples, Step 2 aims to enrich the task hospital's local sample representations. We thus design a novel local feature extracting strategy, which is combined with knowledge distilling from shared samples' x_s^{fed}. Specifically, for a certain unsupervised local representation learner, we enhance it by adding a new loss function, i.e., making the shared samples I_s's learned representations be close to x_s^{fed}, thus enabling the knowledge distillation effect.

In our mechanism implementation, we consider several representative unsupervised representation extraction methods, i.e., auto-encoder (AE) [17], beta-VAE [16], and GAN [14]. Especially, if the input features are from a shared sample, we add a knowledge
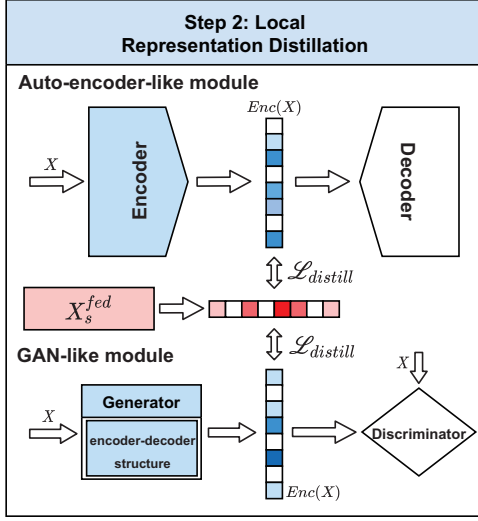
Figure 3: The structure of distillation module in Sec. 3.3. In VFedTrans, the generator of the GAN-like module uses the encoder-decoder structure [38, 44].

distillation loss function by comparing the encoder's output to the shared sample's federated representation (learned from Step 1),

$$\mathcal{L}_{distill}(x_s^t) = |Enc(x_s^t) - x_s^{fed}| \qquad (6)$$

where x_s^t is the shared samples' local features in the task hospital. Fig. 3 shows the structure of distillation module for various representation extraction methods.

Hence, the complete loss function of the distilled module is,

$$loss = \begin{cases} \mathcal{L}_{recons}(x_i^t) + \theta \mathcal{L}_{distill}(x_i^t) & i \in I_s \\ \mathcal{L}_{recons}(x_i^t) & i \in I_t \setminus I_s \end{cases} \qquad (7)$$

That is, for the task hospital's (non-shared) local samples, the loss function is the same as the original distillation module. For the shared samples, a new knowledge distillation loss is added to the original reconstruction loss; θ is the weight parameter to balance two loss function parts.

After training the federated-representation-distilled module until convergence, the distillation module's output can be a feature enrichment function for the task hospital's local samples. That is, for $i \in I_t \setminus I_s$, $Enc(x_i^t)$ can be used to enrich the original local feature x_i^t. In other words, the enriched local samples' representations $x_i^* = \langle x_i^t, Enc(x_i^t) \rangle$ are given to Step 3 for training a machine learning model for a medical task.

Extension to multiple data hospitals. When there are n data hospitals, the task hospital can repeat the aforementioned Step 1 and 2 with each data hospital. Specifically, for each data hospital d_i, the task hospital can learn a local feature enrichment function Enc_i. Then, by aggregating n local feature enrichment functions learned from n data hospitals, the local samples' final enriched representations become,

$$x_i^* = \langle x_i^t, Enc_1(x_i^t), Enc_2(x_i^t), ..., Enc_n(x_i^t) \rangle \qquad (8)$$

Appendix A.2 summarizes the pseudo-code of VFedTrans.

3.4 Security and Privacy

Security and privacy are key factors to consider in FL mechanism design. While VFedTrans is a knowledge transfer framework that incorporates existing VFL algorithms, the security and privacy protection levels are mainly dependent on the included VFL algorithm. In particular, the FRL module (Sec. 3.2) is the key part to determine the overall security and privacy levels of VFedTrans, as cross-party communications and computations are only conducted in this step. Currently, we implement the FRL module with SOTA VFL representation learning methods including, FedSVD [3] and VFedPCA [8].FedSVD uses two random orthogonal matrices to mask the original data. The third-party server can only use the masked data of each party to obtain SVD result. The third-party server of VFedPCA only needs to use the eigenvectors and eigenvalues of each party's data for weighted summation. All of these methods protect privacy by preventing direct use of data by non-holders. Due to the page limitation, readers may refer to the original papers [3, 8] for specific security and privacy analysis.

3.5 Updating

In general, VFedTrans is efficient to update without the need to completely re-running three steps for all the hospitals.

Local Incremental Learning - *New task hospital samples*. Note that the representation distillation is conducted locally at the task hospital. Then, if the task hospital t has a number of new local samples, t can locally re-conduct the representation distillation to learn an updated local feature enrichment function. The task hospital t does not need to communicate with any data hospitals for this updating, which is very efficient and convenient.[5]

Task Independence - *New tasks*. Similar to new samples, if the task hospital t has a new task label to predict, t also does not need to communicate with other hospitals. t only needs to repeat Step 3 with the new task label.

Knowledge Extensibility - *New data hospitals*. The task hospital can learn a new local feature enrichment function Enc' from the new data hospital (repeat Steps 1 and 2 with the new data hospital), and then enrich the local feature representation as $x_i^* = \langle x_i^*, Enc'(x_i^t) \rangle$.

4 EVALUATION

In this section, we empirically verify the effectiveness of our mechanism with four real-life medical datasets. Our experiments were performed on the workstation using NVIDIA RTX 3090, Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz, 160GB RAM, PyTorch 1.10.0, Python 3.8 and CUDA 11.3.

4.1 Datasets

We evaluate our mechanism on the four medical-related datasets: MIMIC-III [20], RNA-Seq [46], HAPT [41], and Breast [21]. Appendix A.3 details these four datasets and shows the default data split to different hospitals in the experiments. We suppose that there exists one task hospital and one data hospital by default. Due to the page limitation, for most experiments, we show the results on MIMIC-III and HAPT datasets.

4.2 Baselines

To verify the effectiveness of our mechanism, we compare it with four baselines:

---

[5]In practice, for new samples, the existing local feature enrichment function (without updating) can still be used. Our evaluation would test this setting (Sec. 4.7).

- *LOCAL*: This baseline leverages only the task hospital's local features for training the task-specific model.
- *FTL [30]*: FTL is an end-to-end FL method for transferring knowledge to local samples. Specifically, based on shared samples, FTL maps different parties' raw features to a common feature space to achieve knowledge transfer.
- *IAVFL [40]*: IAVFL first learns a federated model on shared samples and then learns a local model (for local samples) by considering both ground-truth labels and soft labels produced by the federated model.
- *FedSimLoc*: FedSim [49] is originally designed for fuzzy linking of samples between VFL parties when samples' exact identifies are unavailable. We modify it to our scenario, denoted as *FedSimLoc*, by linking a task party's local (non-shared) sample with the top-K similar shared samples. Then, this local sample can be predicted considering its similar shared samples' federated features from other data parties.

While FTL, IAVFL, and FedSimLoc can be used to assist local samples' learning in VFL, they do not hold some key characteristics of VFedTrans, such as *task-independence* (Sec. 3.5). Moreover, these baselines are all designed purely using deep learning models (i.e., neural networks), whereas VFedTrans can do prediction tasks using any machine learning model (e.g., random forest and XGBoost). Note that in many medical tasks, traditional machine learning models still perform very efficiently and effectively [37] (we also run a set of experiments on our datasets to verify this in Sec. 4.6); VFedTrans is thus more suitable for such medical tasks.

4.3 Training Configurations

In our experiments, we use the random forest (RF) as the default machine learning algorithm, FedSVD as the default FRL method, and AE as the default LRD method. Details of the remaining training configurations are in Appendix A.4.

4.4 Main Results

We first report the results when there is one data hospital. Fig. 4 depicts the prediction performance on MIMIC-III and HAPT by varying the number of features in the task hospital. Results show that our framework can consistently outperform LOCAL and FL baselines. The former means that VFedTrans can effectively improve the diagnosis accuracy when the task hospital's features are insufficient; the latter shows that our knowledge distillation mechanism can achieve better performance compared to other FL knowledge transfer mechanisms.

Fig. 5 shows the prediction performance on MIMIC-III and HAPT by varying the number of features in the data hospital. The accuracy of VFedTrans gradually goes up as the number of features in the data hospital increases. More interestingly, the accuracy increasing speed of VFedTrans is more significant than baselines. This indicates that VFedTrans can transfer knowledge from rich features of the data hospital much more efficiently than baselines.

Fig. 6 shows how our mechanism performs by changing the number of shared samples between the task hospital and the data hospital. We observe that the performance gets better as there are more shared samples. This also validates the effectiveness of

Method	HAPT	RNA-Seq	MIMIC-III	Breast
VFedTrans (ADA)	.9002 ± .0048	.9556 ± .0034	.6848 ± .0049	.9233 ± .0054
LOCAL (ADA)	.8825 ± .0068	.9333 ± .0052	.6769 ± .0048	.9067 ± .0048
VFedTrans (NN)	**.9530** ± .0089	.9600 ± .0190	.7687 ± .0119	.8436 ± .0035
LOCAL (NN)	.9502 ± .0125	.9583 ± .0172	.7643 ± .0080	.8250 ± .0038
VFedTrans (KNN)	.9203 ± .0079	.9578 ± .0158	.6839 ± .0141	.8583 ± .0049
LOCAL (KNN)	.9122 ± .0070	.9512 ± .0102	.6761 ± .0167	.8300 ± .0058
VFedTrans (XGB)	.9519 ± .0066	.9625 ± .0140	**.8042** ± .0107	.9233 ± .0043
LOCAL (XGB)	.9495 ± .0057	.9548 ± .0089	.7940 ± .0072	.9116 ± .0087
VFedTrans (RF)	.9341 ± .0054	**.9635** ± .0036	.7910 ± .0040	**.9253** ± .0068
LOCAL (RF)	.9267 ± .0062	.9524 ± .0042	.7715 ± .0084	.9100 ± .0042
FTL	.9288 ± .0046	.9413 ± .0075	.7810 ± .0087	.8628 ± .0133
IAVFL	.9295 ± .0083	.9512 ± .0065	.7735 ± .0086	.9085 ± .0066
FedSimLoc	.9301 ± .0065	.9468 ± .0082	.7805 ± .0079	.8786 ± .0084

Table 1: Prediction accuracy on four datasets under different downstream classification models (ADA: AdaBoost, NN: neural networks, KNN: K nearest neighbours, XGB: XGBoost, RF: random forest).

Method	FRL	LRD	Accuracy MIMIC-III	HAPT
VFedTrans	FedSVD	AE	**.7910** ± .0040	.9341 ± .0054
	FedSVD	beta-VAE	.7895 ± .0078	**.9345** ± .0039
	FedSVD	GAN	.7889 ± .0058	.9325 ± .0058
	VFedPCA	AE	.7886 ± .0045	.9330 ± .0068
	VFedPCA	beta-VAE	.7875 ± .0061	.9321 ± .0034
	VFedPCA	GAN	.7868 ± .0088	.9332 ± .0077
LOCAL	-	-	.7715 ± .0084	.9267 ± .0062
FTL	-	-	.7810 ± .0087	.9288 ± .0046
IAVFL	-	-	.7735 ± .0086	.9295 ± .0083
FedSimLoc	-	-	.7805 ± .0079	.9301 ± .0065

Table 2: Prediction accuracy by changing FRL and LRD modules.

VFedTrans: with more knowledge sources (i.e., shared samples), our transfer can always be better.

4.5 Few-shot Results

Real-world data holders mostly keep unlabeled data and have access to only few samples of labeled data. We thus consider a test situation where the task hospital does not have enough labeled samples available. We reduce the samples used for training the downstream model in Step 3 to 10% of the original size and keep the test part the same. Fig. 7 explores our mechanism's performance under this few-shot setting. When the number of data hospitals is 1 (the same setting as in Sec. 4.4), our mechanism can still outperform baselines; meanwhile, all the methods have some degree of performance loss due to the limited number of training samples.

Additionally, building a VFL mechanism by involving multiple data parties is realistically advantageous, particularly for few-shot cases. To validate the effectiveness of VFedTrans under this scenario, we increase the number of data hospitals involved in collaboration. Note that the baselines FTL and IAVFL do not consider multiple data parties in their original design. For these two methods, we run the two-party collaboration between the task hospital and every data hospital, and finally output the ensemble prediction with averaging. With the increase in data hospitals, the performance of VFedTrans grows obviously and outperforms baselines consistently. This means that, with VFedTrans, the task hospital can obtain effective information from multiple data hospitals to compensate for the insufficiency of local data volume and features.
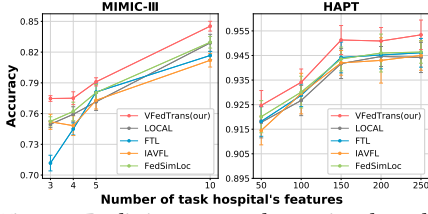
Figure 4: Prediction accuracy by varying the task hospital's feature number.
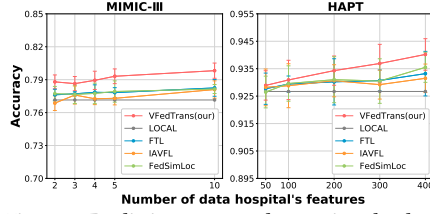

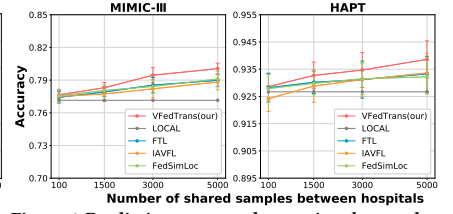Figure 5: Prediction accuracy by varying the data hospital's feature number.

Figure 6: Prediction accuracy by varying the number of shared samples.
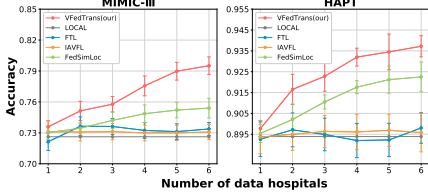

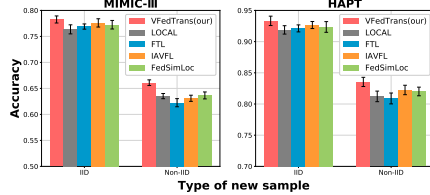Figure 7: Prediction accuracy by varying the number of data hospitals.

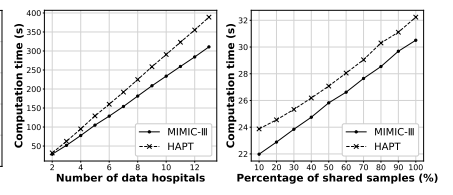Figure 8: Prediction accuracy of new samples.

Figure 9: Computation time by varying the number of data parties and the number of shared samples.

4.6 Robustness Check

Previous experiments were conducted with the default configurations of VFedTrans and two datasets, MIMIC-III and HAPT. Here, we check the robustness of VFedTrans by modifying its configurations (classification, FRL, and LRD modules) on more datasets.

Table 1 shows the results when VFedTrans uses different machine learning models for training the task classifier on four datasets. Our VFedTrans consistently outperforms LOCAL and other baselines, verifying the generalized effectiveness of our knowledge transfer method in various datasets and classifier models. We also find that no single classification model dominates across all the datasets. Then, the flexibility of VFedTrans to incorporate any classification model turns out to be a significant benefit in reality, as we can customize the classifier according to the target dataset.

Besides, we change the methods in FRL and LRD of VFedTrans. Table 2 illustrates the prediction accuracy when we modify the modules in VFedTrans. We can see that the resultant accuracy is robust to such modifications.

4.7 Inductive Learning Results

Moreover, we check how VFedTrans can facilitate inductive learning for new samples of the task hospital (i.e., not used in training the federated-representation-distilled module in Sec. 3.3). In reality, new samples may have a different feature or label distribution from old samples since many factors may change with time, leading to a non-IID case. We thus run inductive learning on both IID and non-IID cases.[6]

Fig. 8 demonstrates that VFedTrans can achieve better performance than other baselines for both IID and non-IID new samples. This reveals the good generalizability of our mechanism's enriched representations. Specifically, while the prediction accuracy decreases for all methods when the experiment setting changes from IID to non-IID, the loss of accuracy is much smaller for our method. For hospitals, the non-IID sample is a completely different case from the original local sample. Models using only local

knowledge cannot fit well with a large number of new non-IID samples. Our framework enables these local hospitals to benefit from collaborative federated medical knowledge learning and to maintain a more solid and trustworthy medical diagnosis in the face of unknown cases and more complicated clinical situations.

4.8 Analysis on $\mathcal{L}_{distill}$

To further verify the effectiveness of VFedTrans on knowledge transfer, we compare the changes in prediction accuracy before and after using our proposed knowledge distillation loss function $\mathcal{L}_{distill} = |Enc(x_s^t) - x_s^{fed}|$ (Sec. 3.3). As shown in Fig. 10, the prediction accuracy is significantly improved with the use of the novel loss component for knowledge transfer. Similarly, when the features of the task hospital change, our mechanism always performs better than the mechanism without knowledge transfer. It is worth noting that our proposed loss function can bring higher accuracy improvement when the task hospital has fewer features. This indicates that VFedTrans enables hospitals with insufficient features to significantly benefit from cross-institution collaboration.

Overall, the results show that the proposed loss is effective for carrying out knowledge transfer. That is, adding $|Enc(x_s^t) - x_s^{fed}|$ into the loss function of representation learning can achieve sensible knowledge transfer. x_s^{fed} can be regarded as the teacher and $Enc(x_s^t)$ as the student. Students use both their own data and the teacher's shared x_s^{fed} to conduct knowledge distillation. The generated representations thus benefit from the teacher's knowledge.

4.9 Computation Time

The computation efficiency of VFedTrans generally depends on the FRL module, as only this step requires collaboration between data and task hospitals. It is worth noting that our implemented FRL algorithm, such as FedSVD, is highly efficient and can be applied to a billion-scale feature matrix [3]. This fundamentally supports the high computation efficiency of VFedTrans.

Besides, we vary the problem scale to check how computation time changes. Fig. 9 records the computation time of VFedTrans by varying the number of data hospitals and the number of shared

---

[6]For the non-IID case, the label distribution of new samples is different from training samples. Details are in Appendix A.5.
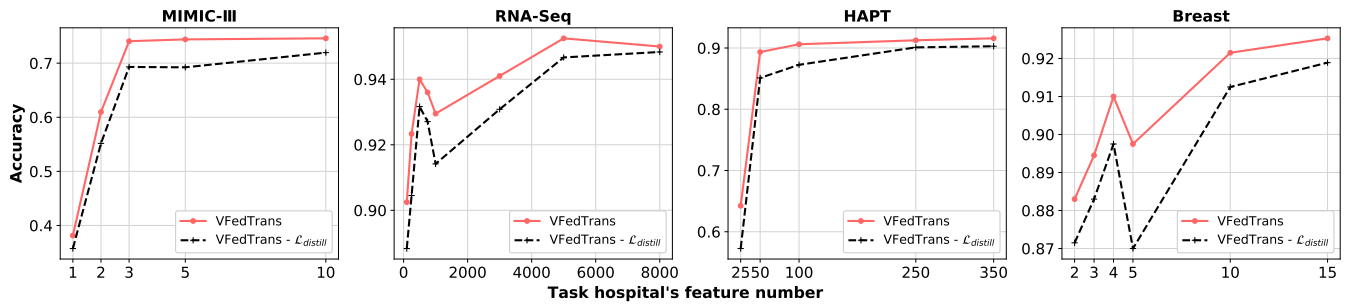
Figure 10: Prediction accuracy by varying the task hospital's feature number with or without the knowledge distillation loss $\mathcal{L}_{distill}$.

samples, respectively. In general, the computation time of our mechanism is linearly proportional to the number of data hospitals and the number of shared samples. This linear relationship further indicates the good scalability of our mechanism.

5 RELATED WORK

In this section, we introduce the related work from two perspectives, VFL methods and FL in healthcare.

5.1 Vertical Federated Learning

VFL focuses on cross-organizational collaborative learning. The common setting [51] is that different organizations hold different features of the same set of samples. Researchers have proposed diverse mechanisms, such as tree-based models [5, 7, 48] and neural networks [12, 19, 23] for VFL collaborations. Compared to these VFL algorithms, the key difference of our mechanism is the application scope. Existing VFL algorithms focus on improving the prediction performance on shared samples. In contrast, our mechanism aims to improve the prediction performance of each party's local (non-shared) samples by transferring the knowledge from shared samples. This study validates the effectiveness of our framework for collaborative healthcare learning. We believe that our proposed VFedTrans can be a good complement to existing VFL algorithms, thus boosting the practicability of FL in reality.

A prior study close to our research is the FTL (federated transfer learning) framework [30]. FTL first trains a specific neural network model according to the task, and maps the heterogeneous feature space of both parties to a common latent subspace by aligned samples[31]. The task hospital then trains the local network model in this subspace. However, our knowledge transfer process is *task-independent*, which means that the distilled representation of the task hospital's samples (i.e., learned from the distilled encoder) can benefit an arbitrary machine learning task and flexibly select any classification model for local tasks; in comparison, FTL is a neural network-based end-to-end training framework that lacks modules for directly training intermediate layers. This makes it challenging to use non-neural network classifiers in FTL. Another recent work on local samples' learning for VFL is proposed by Ren et al. [40], which transfers the knowledge from the shared samples' federated model to the local model by distilling the soft labels generated by the federated model. Like FTL, it also works in a task-dependent manner based on neural networks, which is different from our VFedTrans.

5.2 Federated Learning in Healthcare

FL is a distributed AI paradigm that has been recognized as a promising solution in the field of intelligent healthcare [9, 47, 50]. Fed-Health [6] is designed for wearable health monitoring with smartphone collaboration. Actually, the three steps of this method — local training, model sharing, and server aggregation — are carried out under HFL. Then transfer learning is used in the phase of model personalization. FGTF [33] investigates enhancing a tensor factorization-based collaborative model to handle sensitive health data. FGTF is more concerned with ensuring model convergence and quality reliability while reducing uplink communication costs. Flop [52] is an application of HFL in the field of medical image classification. In Flop, the client only needs to share a partial model with the server for federated averaging; the remaining few layers of neural network can remain private. These existing federated healthcare frameworks rarely consider how to address the imbalance, insufficiency, and heterogeneity of health data among healthcare institutions from a VFL perspective, which however is the objective of our research.

6 CONCLUSION

In this work, we propose a vertical-federated-knowledge-transfer unified framework (VFedTrans) to transfer the knowledge from cross-institutional shared samples to each hospital's local samples. VFedTrans can significantly improve the application scenarios of VFL in healthcare collaborative learning, as it is complementary to the traditional VFL solutions that work only for shared samples. Extensive experiments on medical datasets verify the effectiveness of VFedTrans. Future work may focus on incorporating various SOTA FL techniques [13, 27] into VFedTrans to enrich the framework and considering a new blockchain-based peer-to-peer collaborative learning paradigm [15, 45, 56] to remove the reliance on the third-party server for higher privacy protection.

ACKNOWLEDGMENTS

REFERENCES

[1] Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn M. Eskofier. 2022. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Transactions on Intelligent Systems and Technology* 13, 4 (2022), 54:1–54:23. https://doi.org/10.1145/3501813
[2] Donald A Barr. 2019. *Health Disparities in the United States: Social Class, Race, Ethnicity, and the Social Determinants of Health.* JHU Press.
[3] Di Chai, Leye Wang, Junxue Zhang, Liu Yang, Shuowei Cai, Kai Chen, and Qiang Yang. 2022. Practical Lossless Federated Singular Vector Decomposition over

Billion-Scale Data. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*. ACM, 46–55. https://doi.org/10.1145/3534678.3539402

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Weijing Chen, Guoqiang Ma, Tao Fan, Yan Kang, Qian Xu, and Qiang Yang. 2021. SecureBoost+ : A High Performance Gradient Boosting Tree Framework for Large Scale Vertical Federated Learning. *CoRR* abs/2110.10927 (2021). arXiv:2110.10927 https://arxiv.org/abs/2110.10927

[6] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. 2020. FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. *IEEE Intelligent Systems* 35, 4 (2020), 83–93. https://doi.org/10.1109/MIS.2020.2988604

[7] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. 2021. SecureBoost: A Lossless Federated Learning Framework. *IEEE Intelligent Systems* 36, 6 (2021), 87–98. https://doi.org/10.1109/MIS.2021.3082561

[8] Yiu-ming Cheung, Juyong Jiang, Feng Yu, and Jian Lou. 2022. Vertical Federated Principal Component Analysis and Its Kernel Extension on Feature-wise Distributed Data. *CoRR* abs/2203.01752 (2022). arXiv:2203.01752 https://doi.org/10.48550/arXiv.2203.01752

[9] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. 2021. Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nature medicine* 27, 10 (2021), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3

[10] Mohammad Fattahi, Esmaeil Keyvanshokooh, Devika Kannan, and Kannan Govindan. 2023. Resource planning strategies for healthcare systems during a pandemic. *European Journal of Operational Research* 304, 1 (2023), 192–206. https://doi.org/10.1016/j.ejor.2022.01.023

[11] Yujie Feng, Jiangtao Wang, Yasha Wang, and Sumi Helal. 2021. Completing Missing Prevalence Rates for Multiple Chronic Diseases by Jointly Leveraging Both Intra- and Inter-Disease Population Health Data Correlations. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 183–193. https://doi.org/10.1145/3442381.3449811

[12] Fangcheng Fu, Xupeng Miao, Jiawei Jiang, Huanran Xue, and Bin Cui. 2022. Towards Communication-efficient Vertical Federated Learning Training via Cache-enabled Local Update. *Proceedings of the VLDB Endowment* 15, 10 (2022), 2111–2120. https://www.vldb.org/pvldb/vol15/p2111-fu.pdf

[13] Fangcheng Fu, Huanran Xue, Yong Cheng, Yangyu Tao, and Bin Cui. 2022. BlindFL: Vertical Federated Machine Learning without Peeking into Your Data. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*. ACM, 1316–1330. https://doi.org/10.1145/3514221.3526127

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144. https://doi.org/10.1145/3422622

[15] Jialiang Han, Yun Ma, and Yudong Han. 2022. Demystifying Swarm Learning: A New Paradigm of Blockchain-based Decentralized Federated Learning. *CoRR* abs/2201.05286 (2022). arXiv:2201.05286 https://arxiv.org/abs/2201.05286

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Sy2fzU9gl

[17] Geoffrey E. Hinton and Ruslan Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507. https://www.science.org/doi/abs/10.1126/science.1127647

[18] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015). arXiv:1503.02531 http://arxiv.org/abs/1503.02531

[19] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. 2019. FDML: A Collaborative Machine Learning Framework for Distributed Features. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 2232–2240. https://doi.org/10.1145/3292500.3330765

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, A Freely Accessible Critical Care Database. *Scientific data* 3, 1 (2016), 1–9. https://doi.org/10.1038/sdata.2016.35

[21] Kaggle. 2022. *Breast Cancer Wisconsin (Diagnostic) Data Set*. Retrieved 2022-06-08 from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[22] Seny Kamara, Payman Mohassel, Mariana Raykova, and Seyed Saeed Sadeghian. 2014. Scaling Private Set Intersection to Billion-Element Sets. In *Financial Cryptography and Data Security - 18th International Conference, FC 2014, Christ Church,*

[23] Barbados, March 3-7, 2014, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 8437). Springer, 195–215. https://doi.org/10.1007/978-3-662-45472-5_13

[23] Yan Kang, Yang Liu, and Xinle Liang. 2022. FedCVT: Semi-supervised Vertical Federated Learning with Cross-view Training. *ACM Transactions on Intelligent Systems and Technology* 13, 4 (2022), 64:1–64:16. https://doi.org/10.1145/3510031

[24] Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James R. Foulds. 2021. Equitable Allocation of Healthcare Resources with Fair Survival Models. In *Proceedings of the 2021 SIAM International Conference on Data Mining, SDM 2021, Virtual Event, April 29 - May 1, 2021*. SIAM, 190–198. https://doi.org/10.1137/1.9781611976700.22

[25] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110

[26] Taisa Kushner and Amit Sharma. 2020. Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. ACM / IW3C2, 2906–2912. https://doi.org/10.1145/3366423.3380056

[27] Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-Contrastive Federated Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 10713–10722. https://openaccess.thecvf.com/content/CVPR2021/html/Li_Model-Contrastive_Federated_Learning_CVPR_2021_paper.html

[28] Bingyan Liu, Yao Guo, and Xiangqun Chen. 2021. PFA: Privacy-preserving Federated Adaptation for Effective Model Personalization. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 923–934. https://doi.org/10.1145/3442381.3449847

[29] Yang Liu, Tao Fan, Tianjian Chen, Qian Xu, and Qiang Yang. 2021. FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection. *Journal of Machine Learning Research* 22 (2021), 226:1–226:6. http://jmlr.org/papers/v22/20-815.html

[30] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. 2020. A Secure Federated Transfer Learning Framework. *IEEE Intelligent Systems* 35, 4 (2020), 70–82. https://doi.org/10.1109/MIS.2020.2988525

[31] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. 2022. Vertical Federated Learning. *CoRR* abs/2211.12814 (2022). https://doi.org/10.48550/arXiv.2211.12814 arXiv:2211.12814

[32] Zelei Liu, Yuanyuan Chen, Yansong Zhao, Han Yu, Yang Liu, Renyi Bao, Jinpeng Jiang, Zaiqing Nie, Qian Xu, and Qiang Yang. 2022. Contribution-Aware Federated Learning for Smart Healthcare. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 12396–12404. https://ojs.aaai.org/index.php/AAAI/article/view/21505

[33] Jing Ma, Qiuchen Zhang, Jian Lou, Li Xiong, and Joyce C. Ho. 2021. Communication Efficient Federated Generalized Tensor Factorization for Collaborative Health Data Analytics. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 171–182. https://doi.org/10.1145/3442381.3449832

[34] Liantao Ma, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. 2021. Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 3558–3568. https://doi.org/10.1145/3442381.3449855

[35] Francesca Marazzi, Andrea Piano Mortari, Federico Belotti, Giuseppe Carrà, Ciro Cattuto, Joanna Aleksandra Kopinska, Daniela Paolotti, and Vincenzo Atella. 2022. Staying Strong, But For How Long? Mental Health During COVID-19 in Italy. *Mental Health During COVID-19 in Italy (April 26, 2022). CEIS Working Paper* 541 (2022). http://dx.doi.org/10.2139/ssrn.4094108

[36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

[37] James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, and Terry Lyons. 2019. The Signature-Based Model for Early Detection of Sepsis From Electronic Health Records in the Intensive Care Unit. In *2019 Computing in Cardiology (CinC)*. IEEE.

[38] Meng Pang, Binghui Wang, Yiu-ming Cheung, Yiran Chen, and Bihan Wen. 2021. VD-GAN: A Unified Framework for Joint Prototype and Representation Learning From Contaminated Single Sample per Person. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2246–2259. https://doi.org/10.1109/TIFS.2021.3050055

[39] Agha Ali Raza, Mustafa Naseem, Namoos Hayat Qasmi, Shan Randhawa, Fizzah Malik, Behzad Taimur, Sacha St-Onge Ahmad, Sarojini Hirshleifer, Arman Rezaee, and Aditya Vashistha. 2022. Fostering Engagement of Underserved Communities with Credible Health Information on Social Media. In *WWW '22: The ACM Web*

Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022. ACM, 3718–3727. https://doi.org/10.1145/3485447.3512267

[40] Zhenghang Ren, Liu Yang, and Kai Chen. 2022. Improving Availability of Vertical Federated Learning: Relaxing Inference on Non-overlapping Data. *ACM Transactions on Intelligent Systems and Technology* 13, 4 (2022), 58:1–58:20. https://doi.org/10.1145/3501817

[41] Jorge Luis Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing* 171 (2016), 754–767. https://doi.org/10.1016/j.neucom.2015.07.085

[42] Yousef Saad. 2011. *Numerical methods for large eigenvalue problems: revised edition.* SIAM. https://epubs.siam.org/doi/pdf/10.1137/1.9781611970739.bm

[43] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021.* ACM / IW3C2, 194–205. https://doi.org/10.1145/3442381.3450097

[44] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 1283–1292. https://doi.org/10.1109/CVPR.2017.141

[45] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. 2021. Swarm Learning for Decentralized and Confidential Clinical Machine Learning. *Nature* 594, 7862 (2021), 265–270. https://doi.org/10.1038/s41586-021-03583-3

[46] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* 45, 10 (2013), 1113–1120. https://doi.org/10.1038/ng.2764

[47] Qiong Wu, Xu Chen, Zhi Zhou, and Junshan Zhang. 2022. FedHome: Cloud-Edge Based Personalized Federated Learning for In-Home Health Monitoring. *IEEE Transactions on Mobile Computing* 21, 8 (2022), 2818–2832. https://doi.org/10.1109/TMC.2020.3045266

[48] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. 2020. Privacy Preserving Vertical Federated Learning for Tree-based Models. *Proceedings of the VLDB Endowment* 13, 11 (2020), 2090–2103. http://www.vldb.org/pvldb/vol13/p2090-wu.pdf

[49] Zhaomin Wu, Qinbin Li, and Bingsheng He. 2022. A Coupled Design of Exploiting Record Similarity for Practical Vertical Federated Learning. In *Advances in neural information processing systems.* https://nips.cc/Conferences/2022/Schedule?showEvent=55343

[50] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter B. Walker, Jiang Bian, and Fei Wang. 2021. Federated Learning for Healthcare Informatics. *Journal of Healthcare Informatics Research* 5, 1 (2021), 1–19. https://doi.org/10.1007/s41666-020-00082-4

[51] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 12:1–12:19. https://doi.org/10.1145/3298981

[52] Qian Yang, Jianyi Zhang, Weituo Hao, Gregory P. Spell, and Lawrence Carin. 2021. FLOP: Federated Learning on Medical Datasets using Partial Networks. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* ACM, 3845–3853. https://doi.org/10.1145/3447548.3467185

[53] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied Federated Learning: Improving Google Keyboard Query Suggestions. *CoRR* abs/1812.02903 (2018). arXiv:1812.02903 http://arxiv.org/abs/1812.02903

[54] Xue Yang, Yan Feng, Weijun Fang, Jun Shao, Xiaohu Tang, Shu-Tao Xia, and Rongxing Lu. 2022. An Accuracy-Lossless Perturbation Method for Defending Privacy Attacks in Federated Learning. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022.* ACM, 732–742. https://doi.org/10.1145/3485447.3512233

[55] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021.* ACM / IW3C2, 1397–1409. https://doi.org/10.1145/3442381.3449860

[56] Liang Yuan, Qiang He, Siyu Tan, Bo Li, Jiangshan Yu, Feifei Chen, Hai Jin, and Yun Yang. 2021. CoopEdge: A Decentralized Blockchain-based Platform for Cooperative Edge Computing. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021.* ACM / IW3C2, 2245–2257. https://doi.org/10.1145/3442381.3449994

A APPENDIX

A.1 Notation

The used notation can be found in Table 3 (as referred to Sec. 2.)

Notation	Description
t, d	Task hospital and Data hospital.
d_i	The i-th data hospital.
X_t, X_d	Features of task hospital and data hospital.
I_t, I_d	Samples of task hospital and data hospital.
Y_t, Y_d	Labels of task hospital and data hospital.
H_i	The original local data of hospital i.
H_i^p, H_i^s	Hospital i's private samples and shared samples.
H_s	Non-directly usable shared samples between two hospitals.
x_s^{fed}	Federated latent representations (Sec. 3.2).
$Enc(x)$	Encoder's output in LRD module (Sec. 3.3).
x^*	Enriched loacal samples' representations.
$\mathcal{L}_{recons}$	Reconstruction loss (Equa. 7).
$\mathcal{L}_{distill}$	Novel distillation loss (Equa. 7).
θ	Weight parameter (Equa. 7).

Table 3: List of used notions.

A.2 Algorithm

The procedure steps of VFedTrans can be found in Algorithm 1 (as referred to Sec. 3).

Algorithm 1 VFedTrans algorithm

Require: Hospital series: $\{H_1, H_2, \cdots, H_n\}$
  Choose two hospitals H_i and H_j
  Get the private samples H_i^p and H_j^p
  Get the shared samples H_i^s and H_j^s
  Get the non-directly usable samples H_s
  $X_s^{fed} \leftarrow FRL(H_s)$ ▷ Details are in Sec. 3.2
  for $H_k \in \{H_i, H_j\}$ **do**
    $Enc(X_k) \leftarrow LRD(H_k^p, H_k^s, X_s^{fed})$ ▷ Details are in Sec. 3.3
    $X_k^* \leftarrow <X_k, Enc(X_k)>$
  end for
  for $H_k \in \{H_i, H_j\}$ **do**
    $MedicalTask(X_k^*)$
  end for

A.3 Dataset

This is a supplement to Sec. 4.1.

A.3.1 Details of the medical datasets.

- *Medical Information Mart for Intensive Care (MIMIC-III)* dataset provides de-identified health-related data for 58976 patients from 2001 to 2012. The dimension is $\mathbb{R}^{58976 \times 15}$. Length of stays (LOS) is the target of prediction and varies between 1 and 4.
- *Human Activities and Postural Transitions (HAPT)* is an activity recognition dataset based on smartphone sensor readings.
- *Gene Expression Cancer RNA-Seq (RNA-Seq)* dataset includes gene expressions in patients with different types of tumor. The dimension is $\mathbb{R}^{801 \times 20531}$ and there are 5 tumor types

to predict. The dataset dimension is $\mathbb{R}^{10929 \times 561}$. The task label is the activity type (12 types).

- *Breast* is calculated from the digital image of the fine needle aspirate of the breast lumps. It has 569 samples and 31 features. The diagnosis result is a binary classification (M = malignant, B = benign).

A.3.2 Data split.

The data held by hospitals under different task settings is shown in Table 4. We shuffle the data before dividing to prevent interference from the label distribution. Assuming that the task hospital t has the fewer resources and its data is insufficient, i.e., t holds a smaller number of samples of features than a data hospital d with more abundant medical resources. This setting is effective in experiments to verify that the knowledge transfer in VFedTrans can better help the weaker party. All experiments except Sec. 4.5 and the first experiment in Sec. 4.9 are single-party tasks (one data hospital). In the multi-party task, the number of samples and features of each data hospital are randomly generated within a given interval. The sample size of the shared H_{s_i} is also generated in this manner while the number of features is $X_t + X_{d_i}$. In addition, we set the dimension of the federated latent representation x_s^{fed} in Sec. 3.2 to be the same as H_t, i.e., both $\mathbb{R}^{I_t \times X_t}$.

Task	Hospital's data		Dimension			
			MIMIC-III	*HAPT*	*RNA-Seq*	*Breast*
Common	H_t	I_t	5000	4000	600	300
		X_t	5	100	6000	15
Single party	H_d	I_d	20000	8000	600	400
		X_d	10	250	8000	15
	H_s	I_s	4000	3000	500	200
		X_s	15	350	16000	30
Multi-party	H_{d_i}	I_{d_i}	$8000 \sim 25000$	$5000 \sim 10000$	-	-
		X_{d_i}	$5 \sim 10$	$200 \sim 400$	-	-
	H_{s_i}	I_{s_i}	$2000 \sim 4000$	$2000 \sim 4000$	-	-
		X_{s_i}	$X_t + X_{d_i}$	$X_t + X_{d_i}$	-	-

Table 4: Default samples and features held by each hospital.

A.4 Training configuration

This is a supplement to Sec. 4.3.

A.4.1 FRL techniques.

In FRL, we use two VFL techniques, FedSVD and VFedPCA, with the former being the default. We list the key parameters of two methods in Table 5.

VFL	Parameter	Default	Description
FedSVD	*num_party*	2	The number of participants.
	block_size	100	Build fix-size block in orthogonal matrix generation.
VFedPCA	*party_num*	2	The number of participants.
	iter_num	100	The number of local power iteration.
	period_num	10	The number of communication period.
	warm_start	True	Use the previous global aggregation vector.

Table 5: Default key parameters in FRL.

A.4.2 LRD modules. We choose Adam optimizer for training distillation module, with learning rate=0.001, batch size=100, epoch=500. AE is the default LRD method. Simultaneously, we also carry out experiments under different LRD modules, such as beta-VAE and GAN, to verify the framework's robustness. The key parameters of the three distillation modules are shown in Table 6.

LRD	Parameter	Default	Description
AE	*depth*	6	The depth of encoder and decoder.
	activation	Sigmoid	The activation function of hidden layers.
	θ	0.001	The weight parameter of $\mathcal{L}_{ditill}$.
beta-VAE	β	4	Balance $\mathcal{L}_{recons}$ and $\mathcal{L}_{KL}$.
	kld_weight	0.00025	The weight of $\mathcal{L}_{KL}$.
	θ	0.00001	The weight parameter of $\mathcal{L}_{ditill}$.
GAN	*d_depth*	4	The depth of discriminator.
	g_depth	4	The depth of generator.
	activation	LeakyReLU	The depth of generator and discriminator.
	negative_slope	0.2	The angle of the negative slope.
	θ	0.00001	The weight parameter of $\mathcal{L}_{ditill}$.

Table 6: Default key parameters in LRD.

A.4.3 Task-specific medical models. For all the datasets, when training the task-specific medical model, we choose 80% of the data as the training set and 20% as the test set. In order to prevent the interference of random seeds, we carry out experiments under 10 different random seeds and compute the average results. Note that we can leverage various machine learning algorithms to train the task-specific model. In our experiments, we use the random forest (RF) as the default machine learning algorithm. We also test the

other popular algorithms, including AdaBoost, KNN, XGBoost [4], and neural network (NN) for robustness checks. The parameters of the downstream model are summarized in Table 7.

Model	Parameter	Default	Description
RF	*n_estimators*	200	The number of the trees.
	max_depth	10	The maximum depth of the tree.
AdaBoost	*max_depth*	3	DecisionTreeClassifier's maximum depth.
	n_estimators	100	The maximum number of estimators.
	learning_rate	0.5	Each classifier's weight at each iteration.
NN	*hidden_layer_sizes*	$(100, 100, 50)$	The number of units in hidden layers.
	α	0.01	Weight of the L2 regularization term.
	max_iter	400	Maximum of iterations.
	activation	relu	Activation function for the hidden layer.
KNN	*n_neighbors*	8	Number of neighbors.
XGBoost	*max_depth*	7	The maximum depth of a tree.
	learning_rate	0.01	Weight at each iteration.

Table 7: Default key parameters in downstream medical models.

A.5 Inductive learning: non-IID setting

This is a supplement to Sec. 4.7. We purposely choose half of the labels and their corresponding samples. Then we randomly select 40% of this portion as new non-IID samples. The remaining 60% of the samples and the other half of label's samples are used for the generation of hospital data.