# Incongruent Multimodal Federated Learning for Medical Vision and Language-based Multi-label Disease Detection

**Pramit Saha**[1], **Divyanshu Mishra**[1], **Felix Wagner**[1], **Konstantinos Kamnitsas**[1], **J. Alison Noble**[1]

[1] Department of Engineering Science, University of Oxford
{pramit.saha, divyanshu.mishra, felix.wagner, konstantinos.kamnitsas, alison.noble}@eng.ox.ac.uk

## Abstract

Federated Learning (FL) in healthcare ensures patient privacy by allowing hospitals to collaboratively train machine learning models while keeping sensitive medical data secure and localized. Most existing research in FL has concentrated on unimodal scenarios, where all healthcare institutes share the same type of data. However, in real-world healthcare situations, some clients may have access to multiple types of data pertaining to the same disease. Multimodal Federated Learning (MMFL) utilizes multiple modalities to build a more powerful FL model than its unimodal counterpart. However, the impact of missing modality in different clients, called modality incongruity, has been greatly overlooked. This paper, for the first time, analyses the impact of modality incongruity and reveals its connection with data heterogeneity across participating clients. We particularly inspect whether incongruent MMFL with unimodal and multimodal clients is more beneficial than unimodal FL. Furthermore, we examine three potential routes of addressing this issue. Firstly, we study the effectiveness of various self-attention mechanisms towards incongruity-agnostic information fusion in MMFL. Secondly, we introduce a modality imputation network (MIN) pre-trained in a multimodal client for modality translation in unimodal clients and investigate its potential towards mitigating the missing modality problem. Thirdly, we introduce several client-level and server-level regularization techniques including Modality-aware knowledge Distillation (MAD) and Leave-one-out teacher (LOOT) towards mitigating modality incongruity effects. Experiments are conducted with Chest X-Ray and radiology reports under several MMFL settings on two publicly available real-world datasets, MIMIC-CXR and Open-I.

## Introduction

Multimodal Learning (MML) (Xu, Zhu, and Clifton 2023; Bayoudh et al. 2021) has recently emerged as a pivotal area in machine learning research. Different modalities represent diverse features that are sourced from diverse domains but describe similar subjects, offering both shared and complementary information. The essence of MML lies in combining predictive insights from these different modalities to enhance model performance. Despite the effectiveness of MML, many existing methods are constrained by their reliance on complete modalities, which is often scarce in practice, particularly when dealing with numerous modalities. Real-world multimodal data often presents inherent challenges with missing or incomplete modalities posing significant hurdles in the learning process (Aguilar et al. 2019; Ma et al. 2021; Jaques et al. 2017; Pham et al. 2019; Parthasarathy and Sundaram 2020). The presence of missing modalities within the multimodal datasets introduces complexities that traditional models struggle to accommodate, demanding specialized techniques to ensure effectiveness.

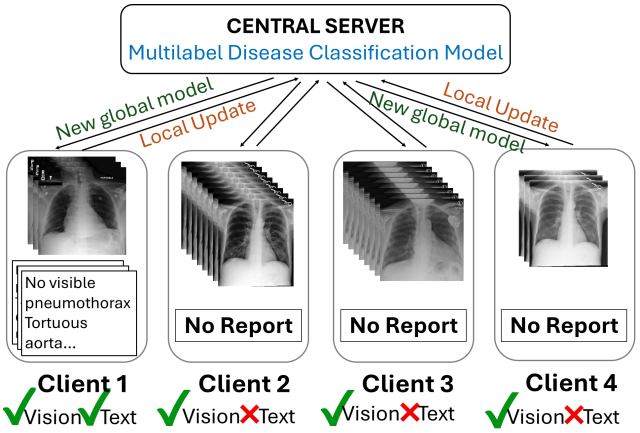


Figure 1: Problem overview. Only 1 out of 4 clients have both modalities, *i.e.*, CXR image and radiology report.

Typically, MML works focus on centralized training, requiring collection and storage of multimodal data on a server for training the models, leading to privacy concerns. The drawbacks of centralized learning has inspired researchers to develop and apply Federated Learning (FL) that enables various clients to collaboratively train models without sharing local data (Mammen 2021; Aledhari et al. 2020; Li et al. 2021; Zhu et al. 2021; Huang, Ye, and Du 2022; Wagner et al. 2023; Hernandez-Cruz et al. 2024; Saha et al. 2024; Wagner et al. 2024). Tackling modality incongruity is crucial in realistic Multimodal Federated Learning (MMFL) as presence of particular modalities across clients might vary, leading to poor performance.

Most existing MMFL works (Xiong et al. 2022; Agbley et al. 2021; Salehi et al. 2022; Qayyum et al. 2022;

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Nandi and Xhafa 2022; Chen and Li 2022; Wei 2021) assume the presence of all modalities in each client. Despite being a critical question, investigation on the impact of missing modality (Le et al. 2023; Zhao, Barnaghi, and Haddadi 2022; Chen and Zhang 2022; Yu et al. 2023) during training has been limited. Intuitively, multimodal models are expected to be more powerful than unimodal models. Therefore, it follows that multimodal clients involved in FL should have better performance than their unimodal versions owing to the availability of complementary information from additional modalities. However, it is not evident how the presence of both unimodal and multimodal clients impact the performance of MMFL in practice. MML models are assumed to be more robust to missing modalities owing to possible redundancy between modalities. As a result, even if some clients are missing modalities, the other modalities should be able to compensate the loss. On the other hand, multimodal integration has been observed to be vulnerable to incomplete or missing modalities in centralized setting as MML models possess larger input dimension than unimodal models and the missing input dimensions may hamper the model training. To summarise, the impact of clients missing some modalities in MMFL is not well-known.

This is particularly crucial **in low- and middle-income countries** (LMICs), where healthcare infrastructure often faces significant challenges, including a **shortage of medical experts like radiologists**. This shortage can lead to **gaps in the availability of comprehensive medical reports**, which are critical for accurate disease diagnosis. Many healthcare facilities in these regions may have access to imaging equipment but lack the expertise to interpret these images and provide detailed reports. This creates a crucial need for solutions that can enhance diagnostic capabilities despite these limitations. MMFL with missing modality tackling scheme presents a promising research avenue to address this challenge. Approaches for handling missing modalities in FL can enable healthcare facilities in LMICs to benefit from collaborative model training with institutions that have access to more comprehensive datasets, including both images and reports.

Through this work, we attempt to address the incongruent MMFL issues and answer: **Does an incongruent MMFL system benefit over unified FL by leveraging the extra modality present in multimodal clients?** Another related question is: Does the modality incongruity vary based on client heterogeneity? These questions are particularly crucial as addressing these questions can potentially help set up a practically beneficial MMFL system among clients in real-world healthcare scenarios. We strongly believe that this paper will facilitate decision making and provide easy, feasible solutions to alleviate the impact of modality incongruity.

In this paper, we attempt to address these critical questions related to the absence of text modality in incongruent MMFL settings. However, we are aware that the investigation is task-specific and model architecture-sensitive. In other words, varying MMFL settings, target tasks, modalities, model architectures *etc.*, can bias the results due to the presence of multiple variables influencing the learning outcome. Notably, there are many multimodal tasks and innumerable existing model architectures that could be explored in this context. However, the objective of this work is to primarily reveal different insightful aspects of MMFL by varying the presence of the primary modality in clients (text) instead of varying model architectures or proving its generalizability across a large number of MML tasks. In this work, we particularly choose a real-world multimodal problem using Medical Vision and Text (Report) modalities. We address a long-tailed multi-label disease classification problem (with 14 categories) from Chest X-Ray images and radiology reports. In this setting, some clients possess both images and radiology reports whereas the others possess only images as shown in Fig. 1. Our primary contributions are:

1. To the best of our knowledge, this is the first work that investigates modality incongruity effects in heterogeneous MMFL. We empirically determine the conditions under which an incongruent MMFL system performs worse than the corresponding unimodal FL system in the context of non-IID data distribution. This reveals important considerations of designing a practical MMFL system with mixed unimodal and multimodal clients and suggests plausible modifications to improve performance.
2. We first demonstrate how the **variation of self-attention masks** in the multimodal client(s) vary the effectiveness of information fusion in incongruent MMFL system.
3. We transform the incongruent MMFL problem to pseudo-congruent MMFL by introducing a **Modality Imputation Network (MIN)** in unimodal clients and demonstrate its performance across varied MMFL settings as a direct way of mitigating modality incongruity.
4. We introduce regularization schemes in unimodal and multimodal clients to achieve a client-invariant representation despite modality incongruity that includes the incorporation of proximal loss (**FedMultiProx**), contrastive loss (**MultiMOON**), and modality-aware knowledge distillation loss (**MAD**).
5. We also demonstrate the potential of leveraging unlabeled data (both unimodal and multimodal) on the server to mitigate the modality incongruity issues. For this, we first leverage ensemble distillation (**FedDF** (Lin et al. 2020)) and then propose a novel client model fine-tuning strategy called Leave-one-out teacher (**LOOT**).

## Preliminaries and Problem Setup

**Problem Formulation:** We consider a multilabel classification task within MMFL setting with $q$ multimodal and $n$ unimodal clients denoted as $\{K^1, K^2, ..., K^q\}$ and $\{K^{q+1}, K^{q+2}, ..., K^{q+n}\}$ respectively. A sample datapoint in the dataset $D_m$ for a multimodal client $K_m$ along with its label(s) is denoted as $\{(X_i^m, Y^m)\}_{i=1}^{N_m}, m = 1, 2, ..., q$, where $N_m$ denotes the number of modalities in $D_m$. The dataset $D_u$ for a unimodal client $K_u$ is denoted as $\{(X^u, Y^u)\}, u = q+1, ..., q+n$. In this work, we only consider two modalities $(m = 2)$ for multimodal clients (*i.e.*, image and text) whereas only image for unimodal clients. Our goal is to minimize the following loss: $\mathcal{L}(w) =$

$\sum_{m=1}^{q} \mathbb{E}_{\{(X_i^m, Y^m)\}_{i=1}^{N_m} \sim \mathcal{D}^m} \left[ \mathcal{L}_m(w; \{(X_i^m, Y^m)\}_{i=1}^{N_m}) \right] + \sum_{u=1}^{n} \mathbb{E}_{\{(X^u, Y^u)\} \sim \mathcal{D}^u} \left[ \mathcal{L}_u(w; (X^u, Y^u)) \right]$

**Datasets:** We utilized the **MIMIC-CXR** (Johnson et al. 2019) and **NIH Open-I** (Demner-Fushman et al. 2016) datasets. Although both MIMIC-CXR and Open-I consist of chest X-ray images and report pairs, the two datasets have different characteristics since they were collected from separate institutions and the diagnostic information represented by the two X-ray image sets are differently distributed (See Fig. 6 and 7 of **Appendix** $A$). The MIMIC-CXR dataset comprises 377,110 Chest X-ray images along with their corresponding free-text reports. Our experiments were conducted exclusively on 91,685 unique frontal view image-report pairs. These were divided according to the official MIMIC-CXR split (89,395 for training, 759 for validation, and 1,531 for testing). The other dataset, Open-I, includes 3,851 reports and 7,466 Chest X-ray images out of which 3,547 frontal view image-report pairs have been used. There are 14 disease classes in the datasets, *viz.*, Support Device, Pleural Effusion, Consolidation, Pneumothorax, Lung Opacity, Enlarged Cardiomediastinum, Atelectasis, Others, Cardiomegaly, Lung lesion, Edema, Fracture, Pneumonia, Pleural other and No finding. A mild imbalance was observed in MIMIC-CXR where the class ratios ranged from $13.39\%$ (support devices) to $1.2\%$ (pneumonia, and pleural other). A severe imbalance was observed in Open-I with the maximum class ratios of $28.8\%$ (Others, and cardiomegaly) and the minimum of $1.07\%$ (support devices).

**FL settings:** We investigate the modality incongruity effects in both IID and non-IID settings. Following previous works (Chen and Chao 2020; Xiong et al. 2023; Saha, Mishra, and Noble 2023; Acar et al. 2021; Li, He, and Song 2021; Xiong et al. 2023), we use Dirichlet distributions with $\gamma = 100$ for simulating IID client data partition and $\gamma = 0.1, 0.5$ for non-IID partition. We evaluate the model performances primarily with 4 clients under fully multimodal and unimodal settings. We confine our study to only 4 clients as in most cases this depicts a realistic number of collaborating institutions in healthcare. Besides, it is a deliberate methodological choice. Limiting the study to 4 clients allows for a more controlled analysis. With a higher number of clients, the complexity increases, potentially diluting the clarity and specificity of insights into the individual contributions and interactions of multimodal and unimodal clients. This focused approach ensures a more precise and meaningful understanding of the dynamics at play in such FL environments. For analyzing modality incongruent MMFL, we vary the ratio of multimodal and unimodal clients as 1:3 and 3:1.

**Multimodal Learning settings and notations:** For a given Chest X-Ray $v$, we denote the flattened visual feature from the last CNN layer as $v = \{v_1, v_2, ..., v_K\}$ and location feature as $l = \{l_1, l_2, ..., l_K\}$ where $K$ denotes the number of visual features. The final visual embedding is $\tilde{v}_i = v_i + l_i + s_V$ where $s_V$ is a semantic embedding vector for visual features. These features are projected into the final embedding space with same dimension as language embedding space via a fully connected layer. For the corresponding report, the text embedding is denoted as $w = \{w_1, w_2, ..., w_N\}$ and the corresponding positional embedding as $p = \{p_1, p_2, ..., p_N\}$. The final language features are expressed as $\tilde{w} = w_i + p_i + s_L$, where $s_L$ is semantic embedding vector for language features. The visual and language embeddings are concatenated to form joint embedding for feeding into the multimodal transformer in multimodal clients as $\tilde{J} = \{S, v_1, v_2, ..., v_K, S\bar{E}P, w_1, w_2, ..., w_N, E\}$ where the embedding length $L_{emb} = N + K + 3$. Here, we obtain the start, separation and end tokens $S, S\bar{E}P, E$ by adding the special tokens with corresponding position and semantic embedding. For unimodal clients, we apply padding for the missing text embeddings. We learn unified, contextualized representation of CXR and reports using single BERT-based transformer encoder model (Kenton and Toutanova 2019; Moon et al. 2022) and attach 14 linear heads to the transformer to address the 14-class multilabel classification model. More details are in **Appendix** $B$.

| Data | Open-I | | MIMIC CXR | | Open-I | | MIMIC CXR | |
|---|---|---|---|---|---|---|---|---|
| **Partition** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** |
| | **M:U = 3:1** | | | | **M:U = 1:3** | | | |
| IID | 77.64 | 29.35 | 96.43 | 81.90 | 67.61 | 20.43 | 94.51 | 80.85 |
| $\gamma$= 0.5 | 74.27 | 26.60 | 87.92 | 77.28 | 67.33 | 20.36 | 80.01 | 72.75 |
| $\gamma$= 0.1 | 58.56 | 25.38 | 76.84 | 69.67 | 53.69 | 22.38 | 70.14 | 66.86 |
| | **Fully multimodal (M:U = 4:0)** | | | | **Fully unimodal (M:U = 0:4)** | | | |
| IID | 93.01 | 46.54 | 98.00 | 84.63 | 66.85 | 19.76 | 94.05 | 80.70 |
| $\gamma$= 0.5 | 88.79 | 39.84 | 96.89 | 83.37 | 76.12 | 29.77 | 92.26 | 79.85 |
| $\gamma$= 0.1 | 84.37 | 42.12 | 96.10 | 82.55 | 75.18 | 36.75 | 91.12 | 78.80 |

Table 1: MMFL Performance with varying degree of heterogeneity. M and U denotes multimodal and unimodal clients.

## Modality Incongruity in MMFL

In this section, we first ask the following question:

**Question:** Is incongruent MMFL more beneficial than unimodal FL (primarily in healthcare scenario)?

For this, we start by defining how modality incongruity can be quantified in MMFL. We particularly explore three different settings: (a) a fully multimodal setting where all clients have multimodal data - both Chest X-Ray (CXR) images and radiology reports, (b) a fully unimodal setting where all clients have unimodal data, *i.e.*, only CXR, and (c) a mixed unimodal-multimodal setting where some clients have both CXR and reports while others only possess CXR. We further vary the last setting by varying the proportion of unimodal and multimodal clients. We evaluate modality incongruity by comparing model performance in each setting. The higher the performance difference between (a) and (c), the more the modality incongruity effects. If (c) performs poorer than (b), we consider it severely modality incongruent.

**Observation:** We empirically conclude that incongruent MMFL outperforms its unimodal version only under homogeneous setting, *i.e.*, IID data partition. The unimodal FL performance surpasses that of incongruent MMFL for both multimodal client proportions under heterogeneous or non-IID data distribution across clients.

First, we empirically validate that the model performance of (b) improves over (a) under homogeneous setting with Dirichlet coefficient $\gamma = 100$ in Table 1. As observed, when 3 (or 1) out of 4 clients are multimodal clients in IID set-

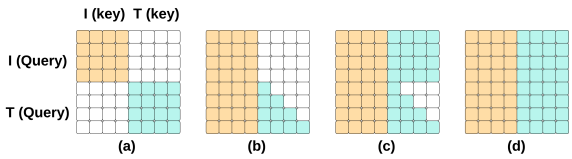


Figure 2: Four self-attention schemes used: (a) Isolated (b) Causal (c) Partially Bidirectional (d) Bidirectional.

tings, the model AUC drops respectively by 15.37 and 1.57 (or 25.40 and 3.49) below fully multimodal FL settings in Open-I and MIMIC respectively. However, the model performance is better than unimodal FL performance in all the above cases. However, we observe that under non-IID partitioning with Dirichlet coefficient $\gamma = 0.1,\ 0.5$, the MMFL performance severely deteriorates and is even worse than the unimodal settings for both the datasets. For $\gamma = 0.1$, when only 3 (or 1) out of 4 clients possess both CXR and report, AUC drops respectively by 16.62 and 14.28 (or 21.49 and 20.98) below fully unimodal settings in Open-I and MIMIC respectively. This indicates that the presence of unimodal clients adversely impacts the multimodal ones in mixed, heterogeneous MMFL which results in sub-optimal utilization of the reports even where they are available. It is observed that the impact of modality incongruity increases with increase in heterogeneity. Besides, even though the model performance decreases with decreasing proportion of multimodal clients, the degradation is relatively low. The replacement of first multimodal client by unimodal client decreases the performance by 25.81 and 19.26, whereas the replacement of two more clients only drops the performance by another 4.87 and 6.70 in Open-I and MIMIC respectively.

To address this issue, we propose and evaluate 3 different solution pathways in the next 3 sections.

## Method 1: Variation of Self-attention Masks

In this section, we investigate 4 self-attention mechanisms (Moon et al. 2022) to facilitate the model's learning of multimodal representation that is more robust to the adverse influence of unimodal clients in incongruent MMFL. Each of these masks offers a unique way of handling the interactions between image and text modalities, which is crucial in our study of modality incongruity in MMFL. By experimenting with these different masks, we gain insights into how different levels and types of modality integration impact the learning process, especially in the presence of unimodal clients.

The self-attention mask $M \in R^{L_{emb} \times L_{emb}}$ is denoted as:

$$M_{jk} = \begin{cases} 0, & (\text{ attention allowed }) \\ -\infty, & (\text{ attention not allowed }) \end{cases} \qquad j,k = 1,\ldots,L_{emb}. \tag{1}$$

In the self attention module, each attention head can be represented as: $\textit{Attention} = \text{softmax}(SA + M)V$, $SA = \frac{QK^T}{\sqrt{d_k}}$ where $Q, K, V, d_k$ respectively indicates queries, keys, values and dimension of keys. Based on modality type, self attention matrix (SA) can be expressed in terms of four subparts: $SA_{q,k} = SA_{S_q:SEP_q,S_k:SEP_k} + SA_{S_q:SEP_q,W_{1k}:E_{Lk}} + SA_{W_{1q}:SEP_q,S_k:SEP_k} + SA_{W_{1q}:E_q,W_{1k}:E_k}$. Below we discuss four types of self-attention and justification behind their usage in the work. For more details, see **Appendix C**.

| | $\gamma = 0.5$ | | | | $\gamma = 0.1$ | | | |
|---|---|---|---|---|---|---|---|---|
| **Self** | **Open-I** | | **MIMIC CXR** | | **Open-I** | | **MIMIC CXR** | |
| **Attention** | **AUC** | **F1** | **AUC** | **F1 score** | **AUC** | **F1 score** | **AUC** | **F1** |
| | **M:U = 1:3** | | | | | | | |
| Isolated | 67.33 | 20.36 | 80.01 | 72.75 | 53.69 | 22.38 | 70.14 | 66.86 |
| Causal | 68.11 | 25.55 | 82.89 | 73.85 | 54.05 | 22.09 | 72.50 | 67.91 |
| parBi | 70.71 | **25.77** | 84.75 | 76.50 | 54.46 | **22.66** | 75.36 | 68.60 |
| Bi | **70.79** | **25.77** | **85.66** | **77.07** | **57.55** | 19.75 | **76.09** | **68.63** |
| | **M:U = 3:1** | | | | | | | |
| Isolated | 74.27 | 26.60 | 87.92 | 77.28 | 58.56 | 25.38 | 76.84 | 69.67 |
| Causal | 74.48 | 26.97 | 88.89 | 78.32 | 58.86 | **28.06** | 78.73 | 71.08 |
| parBi | 75.38 | 27.77 | 90.05 | 79.88 | 59.43 | 26.30 | 80.10 | **72.75** |
| Bi | **76.76** | **30.99** | **90.20** | **80.76** | **61.84** | 26.11 | **81.13** | 72.42 |

Table 2: Performance with varying self-attention schemes

**(i) Isolated Self-Attention:** It restricts the interaction between image and text modalities in multimodal clients. This is particularly important in our context, where some clients only have one modality (CXR), and we need to understand how much each modality can contribute on its own.

**(ii) Causal Self-Attention:** It introduces a controlled interaction between the modalities by allowing language features to attend to both preceding words and visual features, but prevents visual features from attending to language features. This is especially relevant in our case as this restricts the image embeddings to attend to the text which is missing in some clients while allowing the text to be guided by image.

**(iii) Bidirectional Self-Attention:** By allowing unrestricted interaction between the image and text modalities, the bidirectional mask facilitates comprehensive context learning. This is essential for exploring the full potential of multimodal learning, especially in cases where the integration of modalities can lead to a more holistic understanding than either modality alone.

**(iii) Partially Bidirectional Self-Attention**: This aims to combine the benefits of both bidirectional and causal masks. It allows for the integration of image features with language features (like the bidirectional mask) while preserving the causal nature of language (like the causal mask).

**Performance analysis:** The performance of various self-attention schemes in MMFL is summarized in Table 2. As shown in the table, all the other masks improve the performance over isolated masks. Overall, Bidirectional self-attention mask shows the best performance and outperforms the isolated mask by around 3.27% and 4.54% for Open-I and MIMIC respectively in terms of AUC score. The improvement is relatively higher with higher heterogeneity in data partition and for lesser proportion of multimodal clients. However, while variation of self-attention masks show slight improvement in performance, it fails to enable MMFL to surpass the corresponding unimodal performance. This demonstrates that varying self-attention masks can only act as an assisting agent to boost the MMFL performance but not as a stand-alone factor towards achieving better performance than unimodal FL in heterogeneous settings.

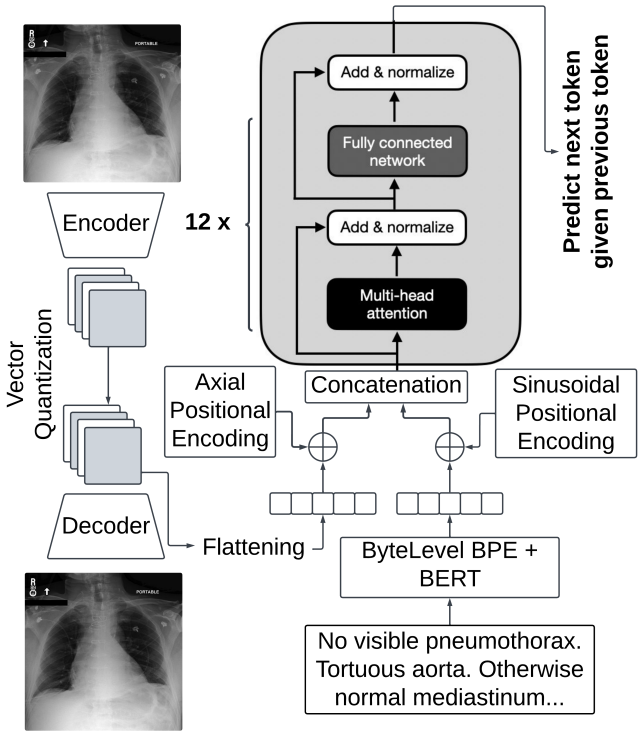


Figure 3: Modality Imputation Network (MIN)

## Method 2: Incongruent to Pseudo-congruent

**Modality Imputation Network:** We convert the incongruent MMFL setting to pseudo-congruent MMFL by introducing a Modality Imputation Network (MIN) to generate radiology reports based on CXR in the unimodal clients as shown in Fig. 3. This imputation is performed prior to the start of FL procedure and does not add any computational overhead. For this, we first utilize VQ-GAN as the image tokenizer, which is composed of an encoder, a decoder, and a learnable codebook of fixed size. The encoder first transforms the CXR image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ into a continuous feature space $\mathbf{z} \in \mathbb{R}^{h \times w \times d_z}$. Consequently, it is quantized into a series of discrete tokens $\{v_1, v_2, ..., v_{h \times w}\}$ by identifying the nearest code embedding in the codebook through nearest neighbor search. The decoder reconstructs the original input from these discrete codes. This approach enables the model to develop a concise and discrete representation of the images. Next, we split each report in the multimodal client into individual word tokens using a byte-level BPE tokenizer, and encase these tokens with specific markers. The ultimate embeddings for the report are derived by combining the word embeddings with a sinusoidal positional embeddings as shown in Fig. 3.

We introduce a BERT-based cross-modal Transformer architecture (Kenton and Toutanova 2019; Lee et al. 2023) and train our model with a causal attention mask in the multimodal client which allows the model to learn about the radiological report in a sequence, conditioned on the CXR images. In order to efficiently manage long-range sequences under limited computational resources, we employ an efficient attention mechanism called Performer (Choromanski et al. 2020). During training in a multimodal client, we concatenate CXR and report embeddings from the same subject as depicted in Fig. 3 and feed it into the model. The problem is considered to be a sequence generation task and model is trained to minimize the negative log-likelihood of predicting the next token based on the preceding tokens. The loss function is: $L = \sum_{i=1}^{n} -\log P\left(w_i \mid w_{0:i-1}\right) + \sum_{i=1}^{m} -\log P\left(v_i \mid w, v_{0:i-1}\right)$ where $n$ = text sequence length + 2 and $m = h \times w + 2$ as $w_0, w_n, v_0, v_m$ are special tokens. See **Appendix** $B$ for more details. After the training procedure is completed, we freeze the pre-trained model and use it for generating reports in unimodal clients.

| Data Partition | Open-I BLEU-4 | | | | MIMIC CXR BLEU-4 | | | |
|---|---|---|---|---|---|---|---|---|
| | C1 (T) | C2 | C3 | C4 | C1 (T) | C2 | C3 | C4 |
| $\gamma$= 0.5 | 0.051 | 0.048 | 0.046 | 0.046 | 0.067 | 0.064 | 0.061 | 0.061 |
| $\gamma$= 0.1 | 0.048 | 0.043 | 0.040 | 0.041 | 0.061 | 0.052 | 0.054 | 0.054 |
| | **AUC** | **Recall** | **Prec** | **F1** | **AUC** | **Recall** | **Prec** | **F1** |
| | **M:U = 1:3** | | | | | | | |
| $\gamma$= 0.5 | 78.42 | 28.74 | 54.19 | 37.56 | 92.86 | 78.16 | 83.33 | 81.23 |
| $\gamma$= 0.1 | 76.78 | 46.26 | 30.94 | 37.08 | 92.08 | 82.96 | 78.64 | 80.74 |
| | **M:U = 3:1** | | | | | | | |
| $\gamma$= 0.5 | 81.24 | 86.68 | 29.49 | 44.01 | 93.45 | 82.31 | 84.74 | 83.68 |
| $\gamma$= 0.1 | 79.61 | 32.48 | 50.00 | 38.67 | 93.36 | 74.25 | 90.09 | 81.30 |

Table 3: MMFL Performance with MIN

**Performance analysis:** Table 3 shows the report generation performance of MIN across all clients in terms of BLEU-4 score. As the model is trained on Client 1 (C1), we first validate its performance on the test set of the same client (C1(T)). For the other clients (C2-C4), we test the report generation performance on all local data samples. The mean BLEU-4 scores for Open-I (MIMIC) with $\gamma = 0.1$ and $\gamma = 0.5$ are 0.043 (0.055) and 0.048 (0.063) respectively. It is also observed from Table 3 that MIN enables the incongruent MMFL system to be more beneficial than unimodal FL in almost all cases. *Eg:* For downstream classification task with $\gamma = 0.1$, incongruent MMFL with 1 and 3 multimodal clients surpass the respective unimodal FL AUC by $1.6$ and $4.43$ for Open-I and by $0.96$ and $2.24$ for MIMIC.

## Method 3: Towards Modality-invariance

The heterogeneous data distribution and modality incongruity lead to distributional modality gaps between the unimodal and multimodal clients thereby posing significant challenges. In this section, we aim to learn modality- and client-invariant representations to aid the information fusion process by bridging the gap between clients. In FL context, this can be achieved either from the client side by constraining the clients or from the server side by leveraging some unlabeled publicly available dataset to learn generalizable representation despite modality shift as discussed below:

### Client-level solutions

Overall, we consider three primary ways of constraining or regularizing the clients to learn modality-invariant representations. In each of the following techniques, we improve upon the naive unimodal FL strategy by incorporating prior knowledge regarding the presence of particular modalities in different clients as shown in Fig.4.

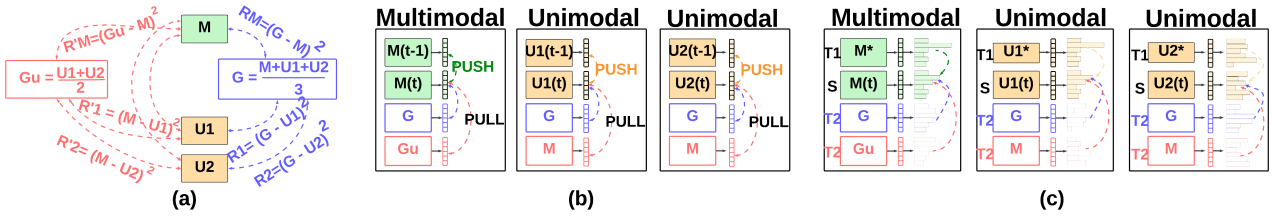


Figure 4: Client-level solutions in 3-client FL - one multimodal client ($M$) and two unimodal clients ($U_1$ and $U_2$). (a) shows the model-based regularization technique of FedProx (in blue) and FedMultiProx (in red). The global model $G$ is replaced by $G_u$ in multimodal clients and $M$ in unimodal clients. (b) shows the representation-based regularization technique of MOON (in blue) and MultiMOON (in red). (c) shows the Modality-aware Knowledge Distillation technique (MAD) and MAD+. $M^*, U_1^*, U_2^*$ represent pre-trained models, *i.e.*, the first teacher model $T_1$. $G$ denotes the second teacher $T_2$ in all clients for MAD. For MAD+, $G_u$ denotes the second teacher model in the multimodal client and $M$ denotes the second teacher model in the unimodal clients.

| | $\gamma = 0.5$ | | | | | | | | $\gamma = 0.1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | **Open-I** | | **MIMIC CXR** | | **Open-I** | | **MIMIC CXR** | | **Open-I** | | **MIMIC CXR** | | **Open-I** | | **MIMIC CXR** | |
| **Methods** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** | **F1** |
| | **M:U = 1:3** | | | | **M:U = 3:1** | | | | **M:U = 1:3** | | | | **M:U = 3:1** | | | |
| FedAvg | 67.33 | 20.36 | 80.01 | 72.75 | 74.27 | 26.60 | 87.92 | 77.28 | 53.69 | 22.38 | 70.14 | 66.86 | 58.56 | 25.38 | 76.84 | 69.67 |
| | **Client-level solutions** | | | | | | | | | | | | | | | |
| FedProx | 69.26 | 24.40 | 83.44 | 73.28 | 74.90 | 27.88 | 88.67 | 80.16 | 56.37 | 23.85 | 71.08 | 69.15 | 60.20 | 24.73 | 76.96 | 71.32 |
| FedMultiProx | 70.92 | 24.91 | 85.67 | 75.06 | 76.24 | 28.05 | 90.03 | 79.48 | 58.12 | 24.51 | 72.34 | 68.26 | 62.27 | 28.91 | 78.20 | 71.68 |
| MOON | 68.29 | 22.45 | 82.00 | 73.98 | 74.75 | 28.82 | 85.73 | 75.24 | 55.64 | 24.04 | 74.05 | 70.88 | 60.48 | 25.05 | 77.54 | 73.27 |
| MultiMOON | 70.38 | 24.73 | 83.98 | 72.41 | 76.02 | 29.64 | 88.98 | 77.05 | 57.92 | 25.59 | 76.69 | 70.61 | 61.96 | **29.38** | 80.18 | 74.47 |
| MAD | 73.82 | **26.95** | 84.69 | 75.20 | 78.28 | 29.36 | 89.90 | 76.87 | 60.77 | 26.80 | 77.23 | 72.20 | 64.70 | 28.84 | 81.87 | 74.43 |
| MAD+ | **74.39** | 26.74 | **86.92** | **78.65** | **79.05** | **29.96** | **91.34** | **80.23** | **61.80** | **27.49** | **80.83** | **73.18** | **66.62** | 29.18 | **84.43** | **75.02** |
| | **Server-level solutions** (utilizing additional data) | | | | | | | | | | | | | | | |
| FedDF (I) | 74.03 | 29.36 | 84.88 | 74.82 | 78.29 | 32.03 | 88.20 | 78.90 | 61.28 | 27.84 | 77.18 | 70.58 | 65.46 | 28.74 | 82.06 | 73.85 |
| FedDF (T) | 71.49 | 26.00 | 82.09 | 72.49 | 77.80 | 30.65 | 85.81 | 77.01 | 59.49 | 26.77 | 73.65 | 71.24 | 63.04 | 27.99 | 80.43 | 74.89 |
| FedDF (I+T) | 76.72 | 33.94 | 85.13 | 75.45 | 80.22 | 35.18 | 90.82 | 79.90 | 63.93 | **29.80** | 80.01 | 74.16 | 68.10 | 30.30 | 86.50 | 76.78 |
| LOOT (I) | 75.06 | 34.98 | 86.85 | **78.19** | 79.94 | 33.74 | 90.33 | 81.28 | 62.49 | 28.02 | 82.22 | 74.54 | 66.02 | 28.85 | 86.08 | 75.47 |
| LOOT (T) | 73.84 | 27.43 | 83.10 | 71.36 | 78.55 | 31.67 | 89.39 | 79.88 | 60.36 | 27.88 | 78.10 | 71.93 | 65.38 | 28.33 | 83.14 | 74.30 |
| LOOT (I+T) | **79.36** | **38.73** | **89.97** | 77.85 | **83.75** | **40.30** | **92.47** | **83.34** | **65.25** | 29.32 | **84.29** | **75.17** | **70.94** | **30.65** | **89.60** | **77.09** |

Table 4: MMFL Performance with client- and server-level solutions. T and I indicate the presence of Text and Image respectively

**Model parameter-based regularization:** This approach incorporates a regularization term to effectively mitigate the influence of varying local updates, as in FedProx (Li et al. 2020). Motivated by this, we introduce **FedMultiProx** in this work to reduce the model diversity among unimodal and multimodal clients originating from the variation of information content from client to client. Rather than constraining each client model to be more aligned with the global model, we specifically regularize the models in unimodal client groups to match the averaged model from multimodal client group and vice versa. This forces the model to focus particularly on modality incongruity effects by penalizing large deviations between the unimodal and multimodal client(s), thereby effectively keeping the local updates in these client groups closer to each other. Accordingly, the optimization objective in $m^{th}$ multimodal client is denoted as $\min_{\theta_t^m} \mathbb{E}_{\{(X_i^m, Y^m)\}_{i=1}^{N_m} \sim \mathcal{D}^m}[\mathcal{L}_{CE}^m(\theta_t^m; \{(X_i^m, Y^m)\}_{i=1}^{N_m}) + \lambda \left\| \theta_t^m - \frac{1}{n}\sum_{u=1}^{n} \theta_{t-1}^u \right\|^2]$, where $\theta$ denotes network parameter. The objective in $u^{th}$ unimodal client can be denoted as $\min_{\theta_t^u} \mathbb{E}_{\{(X_1^u, Y^u)\}) \sim \mathcal{D}^u}[\mathcal{L}_{CE}^u(\theta_t^u; \{(X_1^u, Y^u)\}) + \lambda \left\| \theta_t^u - \frac{1}{q}\sum_{m=1}^{q} \theta_{t-1}^m \right\|^2]$. $\lambda$ is a tuning hyperparameter.

**Representation-based regularization:** Another way of ensuring that the local updates are closely aligned with the representations learned by the global model is applying contrastive learning at the representation level or embedding space, thereby comparing and contrasting the feature representations derived from different models as in MOON (Li, He, and Song 2021). Since the global model is expected to yield modality heterogeneity-agnostic representations, the objective is to minimize the disparity between the client representation ($z_t^k$) and global representation ($z_{t-1}^G$), while simultaneously maximizing the disparity between the client representation at current step ($z_t^k$) and previous step ($z_{t-1}^k$). In this work, we propose **MultiMOON** by replacing the global model of MOON by averaged multimodal client model in unimodal client group and averaged unimodal client model in multimodal client group. For this, we individually replace $z_{t-1}^G$ by $z_{t-1}^M$ for unimodal clients and by $z_{t-1}^U$ for multimodal clients that enforces a stronger constraint that essentially bridges the modality heterogeneity gap between unimodal and multimodal clients. *Eg:* The loss in a unimodal client can be denoted as: $\mathcal{L}_{\text{con}}^k(\theta_t^k; \theta_{t-1}^k; \theta_{t-1}^G; \{X_i^k\}_{i=1}^{N_k}) = -\log \frac{\exp(\text{sim}(z_t^k, z_{t-1}^M)/\tau)}{\exp(\text{sim}(z_t^k, z_{t-1}^M)/\tau) + \exp(\text{sim}(z^k, z_{t-1}^k)/\tau)}$ where $sim(\cdot, \cdot)$ and $\tau$ denote cosine similarity function and temperature parameter respectively.

**Consistency regularization:** The inter-client modality gap can also be addressed by applying consistency regularization at the logit level via knowledge distillation. To this end, we propose **Modality-Aware knowledge Distillation (MAD)** exploiting the global and local knowledge. We introduce a dual teacher model with the global model as one teacher and a frozen local model pre-trained solely on the local client data (unimodal or multimodal) as the other. The student network is trained via guidance from the logit outputs of both the teacher models, thereby indirectly reducing the gap between unimodal and multimodal feature representations in a given client. For this, we minimize the KL divergence of the student logits with respect to the logits of both the teacher models denoted as $\mathcal{L}^k_{MAD} = \sigma\left(z^k_{pre}\right) \log \frac{\sigma\left(z^k_{pre}\right)}{\sigma\left(z^k_t\right)} + \sigma\left(z^G_{t-1}\right) \log \frac{\sigma\left(z^G_{t-1}\right)}{\sigma\left(z^k_t\right)}$ where $z_{pre}$ denotes the locally pre-trained model embedding and $\sigma$ denotes softmax function. Next, following our previous modifications, we propose a variant of MAD, which we term **MAD+**, by replacing the global model with averaged multimodal (or unimodal) client model for unimodal (or multimodal) client groups. Employing knowledge distillation under this setting forces the training to focus on effective balancing of distilled knowledge between unimodal and multimodal clients thereby achieving better modality invariance. The loss function is: $\mathcal{L}^k_{MAD+} = \sigma\left(z^k_{pre}\right) \log \frac{\sigma\left(z^k_{pre}\right)}{\sigma\left(z^k_t\right)} + I\{k=u\}\left[\sigma\left(z^M_{t-1}\right) \log \frac{\sigma\left(z^M_{t-1}\right)}{\sigma\left(z^k_t\right)}\right] + I\{k=m\}\left[\sigma\left(z^U_{t-1}\right) \log \frac{\sigma\left(z^U_{t-1}\right)}{\sigma\left(z^k_t\right)}\right]$. I is indicator function.

## Server-level solutions

We investigate whether the presence of some unlabeled data on server can help us to reduce the modality gap between unimodal and multimodal client models. For this, we particularly consider three different modality settings (only CXR, only report, and both modalities) each for two datasets - with and without domain gap with respect to the client data in the server. For the latter, we utilize a subset of the same source dataset (Open-I or MIMIC) as the clients that is not a part of client data. For the server dataset with domain gap, we utilize a different CXR dataset (See **Appendix** $D$).

**Ensemble Distillation:** We first leverage **FedDF** (Lin et al. 2020) that uses ensemble distillation to train a single student model via guidance from multiple teacher models where each teacher represents the updated local model from each client. The distillation is done using KL divergence by constraining the student model to yield the same output logits as the average logits from the teacher models.

**Leave-one-out Teacher:** We propose a **Leave-one-out teacher (LOOT)** model to finetune each client model in the server by enforcing constraints in the feature representation space targeted towards matching the embeddings of other client models. To this end, we define a mean cosine similarity matrix across all models in a mini-batch based on the embeddings and finetune a given client model (student model) by maximizing the similarity of its mean embeddings with respect to that of the other client models (teacher models). For given $K$ models coming from local updates in K clients, we leave one model (which is used as student model) and use $K-1$ other models as teacher model to bridge the gap between the models. This is performed for all the $K$ client models.

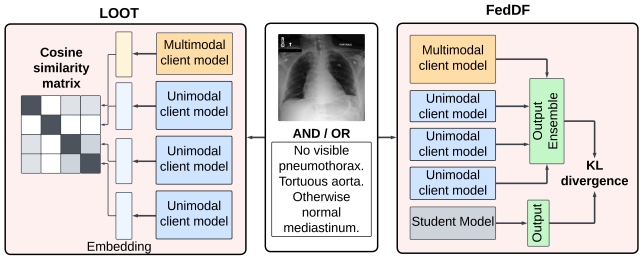


Figure 5: Server-level solutions - LOOT vs FedDF

## Performance analysis

As shown in Table 4, FL algorithms like FedProx, MOON, and FedDF perform slightly better than FedAvg in dealing with modality heterogeneity. The proposed multimodal extensions, enforcing a stronger constraint, further improve the accuracy in each case. MAD+ consistently outperforms the other client-based methods as it leverages a locally-expert teacher model pre-trained on its own client to provide additional supervision on the unimodal task. The server-based models utilizing both image and text of the additional unlabeled data in the server performs better than others. LOOT (T+I) consistently performs better than all other methods as it fine-tunes each client model by trying to match the embeddings of other client models. It is particularly effective as it enforces the unimodal clients to produce multimodal-like embeddings that reduces the modality incongruity effects. Another interesting observation is that LOOT is the only model capable of surpassing the unimodal FL performance for both M:U=3:1 and 1:3 for $\gamma = 0.5$. However, for $\gamma = 0.1$, while LOOT still achieves the best performance, it cannot outperform the corresponding unimodal FL.

## Conclusion

This paper investigates the issue of modality incongruity in MMFL in the context of multilabel disease classification from CXR and radiology reports. Our investigation is based on better utilization of existing techniques from FL literature, adaptation of known methods from other areas as well as introduction of novel yet intuitive MMFL methods. Our comprehensive evaluation demonstrates that modality imputation is the most effective method for tackling modality heterogeneity, closely followed by server-level finetuning of the client models leveraging unlabeled data on server (See **Appendix D and E**). Our work achieves a remarkable advancement in the area of **multimodal learning with missing modality**. It addresses the pressing challenges of the scarcity of skilled health professionals in rural regions, highlighting the potential to improve healthcare outcomes globally.

## Acknowledgments

This work was supported in part by the UK EPSRC (Engineering and Physical Research Council) Programme Grant EP/T028572/1 (VisualAI), a UK EPSRC Doctoral Training Partnership award, the UKRI grant EP/X040186/1 (Turing AI Fellowship), and the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography). FW is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1), by the Anglo-Austrian Society, and by an Oxford-Reuben scholarship. PS acknowledges Yash Bhalgat for the insightful discussions and valuable inputs.

## References

Acar, D. A. E.; Zhao, Y.; Navarro, R. M.; Mattina, M.; Whatmough, P. N.; and Saligrama, V. 2021. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*.

Agbley, B. L. Y.; Li, J.; Haq, A. U.; Bankas, E. K.; Ahmad, S.; Agyemang, I. O.; Kulevome, D.; Ndiaye, W. D.; Cobbinah, B.; and Latipova, S. 2021. Multimodal melanoma detection with federated learning. In *ICCWAMTIP*. IEEE.

Aguilar, G.; Rozgić, V.; Wang, W.; and Wang, C. 2019. Multimodal and multi-view models for emotion recognition. *arXiv preprint arXiv:1906.10198*.

Aledhari, M.; Razzak, R.; Parizi, R. M.; and Saeed, F. 2020. Federated learning: A survey on enabling technologies, protocols, and applications. *IEEE Access*, 8: 140699–140725.

Bayoudh, K.; Knani, R.; Hamdaoui, F.; and Mtibaa, A. 2021. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 1–32.

Chen, H.-Y.; and Chao, W.-L. 2020. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*.

Chen, J.; and Zhang, A. 2022. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Chen, S.; and Li, B. 2022. Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending. In *IEEE INFOCOM*. IEEE.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; et al. 2020. Rethinking Attention with Performers. In *ICLR 2020*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*.

Hernandez-Cruz, N.; Saha, P.; Sarker, M. M. K.; and Noble, J. A. 2024. Review of Federated Learning and Machine Learning-Based Methods for Medical Image Analysis. *Big Data and Cognitive Computing*, 8(9).

Huang, W.; Ye, M.; and Du, B. 2022. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*.

Jaques, N.; Taylor, S.; Sano, A.; and Picard, R. 2017. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *ACII 2017*. IEEE.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Le, H. Q.; Nguyen, M. N.; Thwal, C. M.; Qiao, Y.; Zhang, C.; and Hong, C. S. 2023. FedMEKT: Distillation-based Embedding Knowledge Transfer for Multimodal Federated Learning. *arXiv preprint arXiv:2307.13214*.

Lee, H.; Kim, W.; Kim, J.-H.; Kim, T.; Kim, J.; Sunwoo, L.; and Choi, E. 2023. Unified Chest X-ray and Radiology Report Generation Model with Multi-view Chest X-rays. *arXiv preprint arXiv:2302.12172*.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; and He, B. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

Mammen, P. M. 2021. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.

Moon, J. H.; Lee, H.; Shin, W.; Kim, Y.-H.; and Choi, E. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*.

Nandi, A.; and Xhafa, F. 2022. A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods*, 204: 340–347.

Parthasarathy, S.; and Sundaram, S. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint

representations by cyclic translations between modalities. In *AAAI Conference on Artificial Intelligence*.

Qayyum, A.; Ahmad, K.; Ahsan, M. A.; Al-Fuqaha, A.; and Qadir, J. 2022. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, 3: 172–184.

Saha, P.; Mishra, D.; and Noble, J. A. 2023. Rethinking Semi-Supervised Federated Learning: How to co-train fully-labeled and fully-unlabeled client imaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 414–424. Springer.

Saha, P.; Wagner, F.; Mishra, D.; Peng, C.; Thakur, A.; Clifton, D.; Kamnitsas, K.; and Noble, J. A. 2024. $F^3$OCUS – Federated Finetuning of Vision-Language Foundation Models with Optimal Client Layer Updating Strategy via Multi-objective Meta-Heuristics. arXiv:2411.11912.

Salehi, B.; Gu, J.; Roy, D.; and Chowdhury, K. 2022. Flash: Federated learning for automated selection of high-band mmwave sectors. In *IEEE INFOCOM 2022*. IEEE.

Wagner, F.; Li, Z.; Saha, P.; and Kamnitsas, K. 2023. Post-Deployment Adaptation with Access to Source Data via Federated Learning and Source-Target Remote Gradient Alignment. In *International Workshop on Machine Learning in Medical Imaging*, 253–263. Springer.

Wagner, F.; Xu, W.; Saha, P.; Liang, Z.; Whitehouse, D.; Menon, D.; Newcombe, V.; Voets, N.; Noble, J. A.; and Kamnitsas, K. 2024. Feasibility of Federated Learning from Client Databases with Different Brain Diseases and MRI Modalities. arXiv:2406.11636.

Wei, X. 2021. A multi-modal heterogeneous data mining algorithm using federated learning. *The Journal of Engineering*, 2021(8): 458–466.

Xiong, B.; Yang, X.; Qi, F.; and Xu, C. 2022. A unified framework for multi-modal federated learning. *Neurocomputing*, 480: 110–118.

Xiong, Y.; Wang, R.; Cheng, M.; Yu, F.; and Hsieh, C.-J. 2023. Feddm: Iterative distribution matching for communication-efficient federated learning. In *IEEE CVPR*.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, Q.; Liu, Y.; Wang, Y.; Xu, K.; and Liu, J. 2023. Multimodal Federated Learning via Contrastive Representation Ensemble. *arXiv preprint arXiv:2302.08888*.

Zhao, Y.; Barnaghi, P.; and Haddadi, H. 2022. Multimodal federated learning on iot data. In *2022 IEEE/ACM IoTDI*.

Zhu, H.; Xu, J.; Liu, S.; and Jin, Y. 2021. Federated learning on non-IID data: A survey. *Neurocomputing*, 465: 371–390.