MICCAI

CAR-MFL: Cross-Modal Augmentation by Retrieval for Multimodal Federated Learning with Missing Modalities

Pranav Poudel[2,5], Prashant Shrestha[2*], Sanskar Amgain[2*], Yash Raj Shrestha[3], Prashnna Gyawali[4], and Binod Bhattarai[1]

[1] University of Aberdeen, Aberdeen, UK
[2] NepAl Applied Mathematics and Informatics Institute for research, Nepal
[3] University of Lausanne, Switzerland
[4] West Virginia University, USA
[5] Fogsphere (Redev.AI), UK

Abstract. Multimodal AI has demonstrated superior performance over unimodal approaches by leveraging diverse data sources for more comprehensive analysis. However, applying this effectiveness in healthcare is challenging due to the limited availability of public datasets. Federated learning presents an exciting solution, allowing the use of extensive databases from hospitals and health centers without centralizing sensitive data, thus maintaining privacy and security. Yet, research in multimodal federated learning, particularly in scenarios with missing modalities—a common issue in healthcare datasets—remains scarce, highlighting a critical area for future exploration. Toward this, we propose a novel method for multimodal federated learning with missing modalities. Our contribution lies in a novel cross-modal data augmentation by retrieval, leveraging the small publicly available dataset to fill the missing modalities in the clients. Our method learns the parameters in a federated manner, ensuring privacy protection and improving performance in multiple challenging multimodal benchmarks in the medical domain, surpassing several competitive baselines. Code Available: https://github.com/bhattarailab/CAR-MFL

Keywords: Multimodal Federated Learning · Missing Modality

1 Introduction

Multimodal AI is poised to revolutionize healthcare by integrating data from various modalities such as medical scans, omic data, and pathology reports [1]. By synthesizing information across these different streams, such models offer a more holistic view of diseases, significantly enhancing diagnostic precision [25,22]. Yet, to learn such models, extensive training data is required, which poses a critical challenge, especially within healthcare, where privacy and security constraints severely limit data availability.

---

* These authors made equal contributions.

Federated learning offers a privacy-preserving approach to utilizing private datasets across medical centers, where learnings from private data can be aggregated to improve shared healthcare models without sharing raw data [17,31,26]. However, challenges arise in multimodal settings, as the modality distribution can be different across medical centers, due to clinical equipment unavailability, different acquisition procedures, or storage limitations [23]. Relevant approaches for handling missing modality scenarios in centralized settings often involve prompt-based [21,28], generative-based [3,16] or dropout-based approaches [14,24]. A notable work by Chen et al. [4] in the centralized domain employs cross-modal retrieval to impute the missing modality, where the complementary modality sample is retrieved by comparing features across modalities. However, this approach assumes shared representations between modalities, requiring additional contrastive objectives to enforce alignment.

Previous works that assume multimodal data to be available to all clients in a federated setting [20,19,2] disregard the challenges of missing modalities, where clients trained on a single modality may fail to properly capture the relationship between modalities, resulting in a divergence from multimodal clients and suboptimal federated training [12,29]. A naive approach to dealing with missing modality is to simply replace the missing data with zeros [30,15]. However, such a strategy can induce bias in the global model, causing it to potentially neglect representations learned from rarely present modalities. Existing works considering missing modality in federated setting either assume access to a large-scale public multimodal dataset [29], impractical in the context of medical data, or utilize class prototypes [15] which may not be suitable for multi-label classification. Some approaches like sharing amplitude spectrum [27] are ill-defined for modalities like text and tabular data.

Towards this end, we propose a novel approach utilizing Cross-modal data Augmentation by Retrieval at the client level to address the missing modality in Multimodal Federated Learning. Our method, CAR-MFL leverages a small, publicly available multimodal dataset during training to retrieve the most likely absent modality samples for unimodal inputs. In doing so, we capitalize on the significant efforts made by the medical community to make certain datasets publicly available, while also scaling to include both multimodal and unimodal private datasets across medical centers. To the best of our knowledge, this is the first work to augment a small-scale publicly available data set for Multimodal Federated Learning. While a recent work by Hao et al. [8] augments data at the client level for federated learning to mitigate the effects of data heterogeneity, it is limited to an unimodal setup. Our focus, however, extends to maximizing the use of publicly available real multimodal data, often overlooked in federated learning scenarios. We summarize our key contributions below:

- We introduce CAR-MFL, a novel multimodal federated learning framework employing retrieval-based cross-modal augmentation for missing modalities.
- We perform extensive evaluations, both quantitative and qualitative, on challenging benchmarks and compare with the competitive baselines, surpassing their performances.

2   Method

A typical multimodal federated setting consists of K clients, each with its private dataset D_k of n_k samples. The i_{th} data sample in D_k along with its corresponding label, $Y^{(i)}$ is denoted by the tuple $(\{X_m^{(i)}\}_{m=1}^{M_k}, Y^{(i)})$, where M_k represents the number of modalities in the kth client. Without loss of generality, we consider the presence of two modalities, image I and text T. All clients share the same model architecture, comprising of modality-specific encoders f_e, a concatenation-based fusion module $\oplus$, and a classifier head f_c. We thus represent the entire model $\boldsymbol{w}$ as a set $\{f_e^I, f_e^T, f_c\}$. The general global loss function to be optimized in multimodal federated setting is expressed as:

$$\underset{w}{\operatorname{argmin}}\ L(\boldsymbol{w}) = \sum_{k=1}^{K} \frac{|D_k|}{|D|}\ l_k(\boldsymbol{w}, D_k) \tag{1}$$

$$\text{where,}\ l_k(\boldsymbol{w}, D_k) = \frac{1}{|D_k|} \sum_{(X_I^{(i)}, X_T^{(i)}, Y^{(i)}) \in D_k} \mathcal{L}(f_c(f_e^I(X_I^{(i)}) \oplus f_e^T(X_T^{(i)})), Y^{(i)}) \tag{2}$$

Here, D denotes the union of all client datasets, and $\mathcal{L}$ refers to the client-level local loss for the model $\{f_e^I, f_e^T, f_c\}$ on a data sample $(X_I^{(i)}, X_T^{(i)}, Y^{(i)})$. A common approach to handling missing modality is zero-filling. Considering the case of text sample to be absent,. the local loss using zero-filling becomes $\mathcal{L}(f_c(f_e^I(X_I) \oplus 0), Y)$. Naively optimizing Eqn 1 in this scenario leads to modality biases where the resulting model tends to overlook the features of modality absent in majority of the clients, yielding suboptimal solutions.

Cross-Modal Augmentation: To address this challenge, we propose CAR-MFL, which utilizes cross-modal data augmentation for unimodal clients by utilizing a small publicly available multimodal dataset D_p annotated and curated by the community. This method creates new multimodal instances in unimodal clients by pairing the unimodal data with complementary modalities retrieved from the public dataset. This results in the following cases of augmented input:

$$X = \begin{cases} X_I, T & \text{if image-only client} \\ I\ , X_T & \text{if text-only client} \\ X_I, X_T & \text{if multimodal client} \end{cases} \tag{3}$$

Here, I, T are complementary modalities retrieved from the public dataset using cross-modal augmentation. Our overall method is as illustrated in Fig. 1.

Commonly, complementary modalities are retrieved through direct cross-modal retrieval. This assumes aligned representation between modalities, necessitating additional cross-modal objectives for enforcement. Furthermore, such representations may not be optimal for classification tasks that could benefit from complementary information. Our method first performs intra-modal retrieval and uses the associated modality of the retrieved public sample to impute the missing modality. Consider an unimodal client with only image data (a
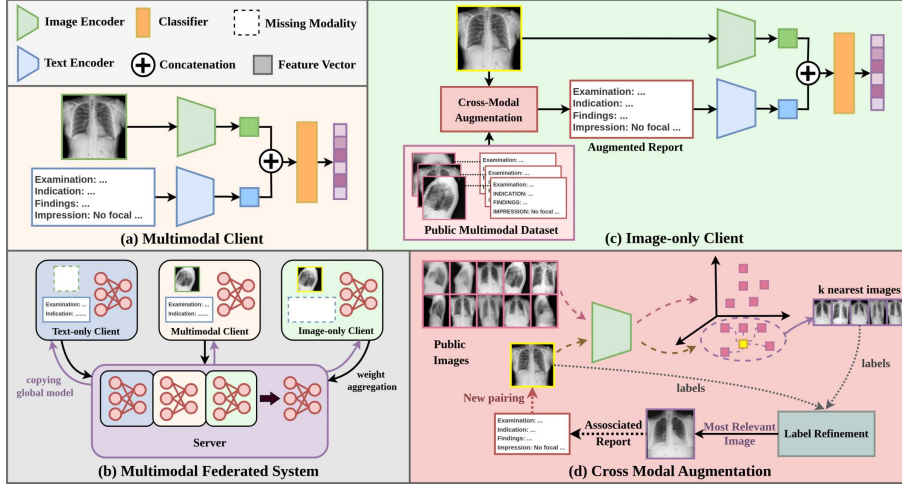
Fig. 1: Illustration of **CAR-MFL**. **(a)** Multimodal client with access to multi-modal data. **(b)** Multimodal federated system with missing modality. **(c)** Image client with only image samples; missing text modality is retrieved via our Cross-Modal Augmentation module. **(d)** Cross-Modal augmentation procedure for a query image (yellow): Most relevant image from public data is retrieved based on distance in feature space and label similarity. Then, the associated text of the retrieved image is paired with the query image forming a paired input.

similar process is applied to text-only clients). We begin with a pool of images $I_p = \{I_1, I_2, .., I_{N_p}\}$ from the public multimodal dataset D_p. At the start of each communication round, we select top-k closest images, I_{pk} from the public data pool for each image X_I^c in the client dataset D_c. The selection is made based on the distance metric, δ between feature vectors of public images and local client images generated by the encoder of the global model, which is calculated as:

$$\delta(X_I^{c(i)}, I_i) = ||f_e^I(X_I^{c(i)}) - f_e^I(I_i)||^2 \tag{4}$$

Even intra-modal retrieval may yield semantically dissimilar samples at the beginning of the training procedure, owing to the lack of discriminative information in representation as the model is either pretrained on semantically different datasets or randomly initialized. We thus propose to augment *label refinement*, which further ranks the top-k set I_{pk} based on labels. This ranking prioritizes images with labels closely matching the query image, measured by Jaccard similarity between the label sets:

$$J(L_q, L_{pk}^j) = \frac{|L_q \cap L_{pk}^j|}{|L_q \cup L_{pk}^j|} \tag{5}$$

where, L_q denote the label set of $X_I^{c(i)}$ and L_{pk}^j denote label sets of jth image in I_{pk}; I_j. We select the image, $I_r^{(i)}$, from I_{pk} with the highest Jaccard similarity

score using Eqn 5. This selected image has its associated text, $T_r^{(i)}$, which we pair with $X_I^{c(i)}$ to form complete multimodal data instances (Eqn. 6). This process is repeated for all the images in the local dataset, and a new augmented dataset, D_c' is created, which is used for training the local model in unimodal client c:

$$D_c' = \bigcup_{(X_I^{c(i)}, Y^{(i)}) \in D_c} (X_I^{c(i)}, T_r^{(i)}, Y^{(i)}) \tag{6}$$

Using the augmented samples, the local loss function in Eqn. 2 becomes $\mathcal{L}(f_c(f_e^I(X_I^{(i)}) \oplus f_e^T(T_r^{(i)})), Y^{(i)})$. Retrieving these samples provides vital task-related information compared to the naive zero-filling approach. Consequently, the trained models learn to incorporate information from both modalities, mitigating the issue of modality bias. In this way, we utilize the associated modality of the sample obtained by intra-model retrieval with label refinement to achieve cross-modal augmentation for missing modalities.

Weight Re-adjustment: It is possible that the retrieved sample may not completely agree with the original label set Y in every case. Using Eqn 1 directly could introduce noise and corrupt the global model. Hence, we perform aggregation weight re-adjustment by scaling down the existing weights ($\frac{|D_k|}{|D|}$) of the complementary modality encoders of unimodal clients with the hyper-parameter α, followed by *softmax* normalization to derive new re-adjusted weights. We use these adjusted weights only when updating the respective encoders, keeping the weights of the classification layer unchanged. We determine $\alpha = 0.3$ through cross-validation (Refer to Supplementary Material).

How do we address the Privacy Concerns? It may raise questions for the reader that augmenting cross-modal data from publicly known identities may breach privacy protection. Our augmentation is a dynamic process that varies with the iterations (See Fig. 4). It is evident that averaging the features from multiple identities obfuscates the identity and protects privacy through k-anonymity [7]. Moreover, the augmentations are done at the client level and pairing information is never shared with the other clients.

3   Experiments and Results

Datasets and Setups: We leverage three publicly available datasets: MIMIC-CXR [11], NIH Open-I [5], and CheXpert [10], to create two experimental setups: Homogeneous and Heterogeneous. Both setups exclusively contain frontal views of chest X-rays and share a common public dataset comprising image-text pairs of 1000 patients sampled from MIMIC-CXR. The homogenous setup comprises 10 clients, each with training data for 810 patients sampled from the same dataset MIMIC-CXR. In contrast, the heterogeneous setup includes 8 image-only clients, each with data for 900 patients from CheXpert, and 2 multimodal clients, each with data for 1116 patients from NIH Open-I datasets. This setup mirrors a real-world scenario where the data characteristics and label distribution across different categories of clients vary significantly. In both setups, the global model

Table 1: AUC↑ Performance in Homogeneous and Heterogeneous partitions.

| Partitions	Homogeneous									Hetero
(I:T:M)	4:0:6	0:4:6	2:2:6	6:0:4	0:6:4	3:3:4	8:0:2	0:8:2	4:4:2	8:0:2
mFedAvg	86.64	87.95	87.4	82.99	87.01	84.48	79.69	86.32	80.47	72.76
mFedAvg$P$	87.77	88.81	88.69	85.49	87.64	85.78	81.95	87.13	83.46	82.52
CreamFL	78.42	75.39	75.79	78.47	75.47	75.65	78.36	75.28	75.49	78.22
CAR-MFL	**89.78**	**90.06**	**89.94**	**88.22**	**90.12**	**89.62**	**87.31**	**89.22**	**88.93**	**86.14**

is validated and tested using common validation and test sets derived from the official splits of MIMIC-CXR. Additionally, each client can have multiple data instances. For brevity, we express client configurations in the format **I:T:M** where I and T denote the number of image-only and text-only clients respectively, and M indicates the number of multimodal clients in the setup.

Implementation Details: We utilize pretrained Resnet50 [9] and BERT-base [6] architecture as image and text encoders respectively. Their outputs are reduced to 256-dimensional features and L2-normalized before fusion. The intra-modal retrieval is performed on these normalized features. We adopt a simple concatenation approach for multimodal fusion, followed by a linear layer for classification. The models are optimized using Adam [13] with a learning rate of $1e^{-4}$ and are locally trained for 3 epochs per communication round (30 total rounds). We retrieve the top 10 similar data points before subjecting them to label refinement to generate an augmented pair. We evaluate models using macro AUC (Average area under the Receiver Operating Characteristic Curve) on the multimodal test set. We use the labels generated by MedViLL [18] for Open-I. All reported values represent means from experiments conducted with 3 random seeds. All figures and studies, besides Table 1 were conducted on the setting with 8 unimodal clients (8:0:2 and/or 0:8:2) within the homogeneous setup.

Baselines: We compared our method with three competitive baselines. First, we adapted FedAvg to handle multimodal data which we refer to as mFedAvg. Originally, FedAvg does not utilize publicly available datasets while training the model. To ensure a fair comparison, we train mFedAvg w and w/o the public data set, denoted as mFedAvg and mFedAvg$P$ respectively. In both cases, we perform zero-fillings for the missing modalities. Both our method and mFedAvg$P$ train a separate model on the public data as a virtual client which is also used in aggregation. Finally, we compare our method to CreamFL [29], a state-of-the-art method for multimodal federated learning with missing modalities (ICLR'23).

3.1   Quantitative Results

Table 1 shows the AUC of all baselines and our method across various client configurations in both setups. Our method, CAR-MFL significantly outperforms all baselines at all settings of modality distribution in the homogeneous as well as challenging heterogeneous partition demonstrating its applicability in real-world
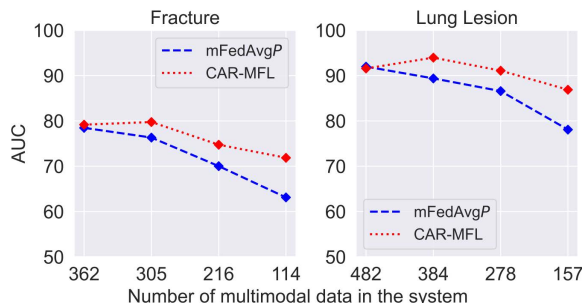
Fig. 2: Comparison between mFedAvg$P$ and CAR-MFL on rare pathologies.

settings. In the homogeneous setup, CAR-MFL, with just 2 multimodal clients, surpasses the mFedAvg$P$ model employing 6 multimodal clients. Furthermore, CAR-MFL outperforms mFedAvg$P$ in the heterogeneous setup by 4%. The poor performance of CreamFL is attributed to their assumption of aligned representations between modalities and their imposition through data-intensive intermodal loss. Furthermore, their utilization of 50,000 public data samples is in stark contrast to our setups which utilize just 2701 paired samples as the public data. For additional results and findings, please refer to Supplementary section.

Clinical Relevancy: To investigate the clinical applicability of our proposed framework, we analyzed its performance with a focus on rare pathological classes. This evaluation specifically targeted how the absence of multimodal data influences the diagnostic accuracy of such rare conditions. We focused on two rare categories: "Fracture" and "Lung Lesion", with less than 1.5% of instances out of 43,802 total instances, progressively reducing the share of multimodal clients to simulate a transition towards a predominantly unimodal dataset within the system. As shown in Fig. 2, despite an anticipated decrease in the effectiveness of the CAR-MFL approach under these conditions, it exhibited robustness compared to the mFedAvg$P$ framework, which demonstrated a significant drop in performance when moving from a multimodal to a unimodal client setup. This stark difference in performance, particularly noted in the handling of these rare classes, underscores the superior clinical relevance of our model.

Study on Public Data Size: In Fig. 3 a, We evaluate the effectiveness of our approach at various levels of public training data keeping the test set unchanged. CAR-MFL demonstrates substantial performance improvement compared to mFedAvg$P$ and robustness to decreasing public data. Remarkably, even with data from just 250 patients, CAR-MFL outperforms mFedAvg$P$ by 6%.

3.2   Qualitative Results

Modality Bias: We examine the distribution of classification layer weights to assess the effectiveness of our approach in mitigating modality bias. We calculate the distribution by summing the absolute values of the weights for image and

(a)   (b)
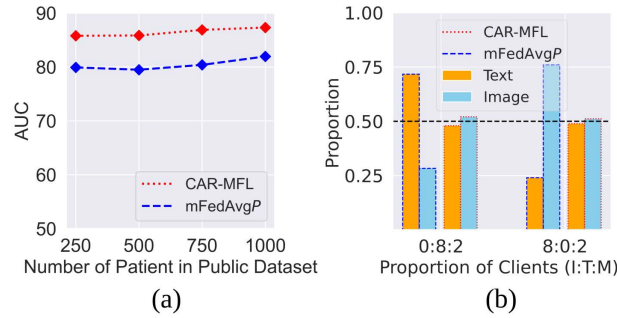
Fig. 3: (a) Study of model AUC on varying public data size. **(b)** Distribution of weight values of classification layer across image and text features.

Original Pair	Retrieved Complementary Modality		
	Round 1	**Round 15**	**Round 30**
	the heart size is normal . the hilar and mediastinal contours are normal . the lungs are without chf or evidence of focal consolidations concerning for pneumonia . there is no pleural effusion or pneumothorax . the visualized osseous structures are unremarkable .	pa and lateral radiographs of the chest demonstrate clear lungs and normal hilar and cardiomediastinal contours . there is no pneumothorax or pleural effusion . pulmonary vascularity is normal . there is extensive bridging osteophytosis of the thoracic spine consistent with dish . a cholecystectomy clip can be seen in the right upper quadrant of the abdomen .	lung volumes are low but the lungs are grossly clear without lobar consolidation pleural effusion pneumothorax or pulmonary edema . the cardiomediastinal silhouette is within normal limits .
lungs are well expanded . a vague opacity is again seen in the left lateral lung base similar to multiple priors back to and likely representing an area of possible scarring or atelectasis . there is no pleural effusion or pneumothorax . the cardiomed-iastinal silhouette is unremarkable			

Fig. 4: Qualitative Analysis of retrieved samples across different training rounds. Column 1 contains a paired image text sample. The first row displays retrieved text reports from the public dataset when the text modality is missing. The second row displays retrieved images when the image modality is missing.

text features separately, then normalize them by the total weights. As depicted in Fig. 3 b, mFedAvg$P$ with zero-filling, exhibits a distinct bias, with most weight values allocated to the modality present in the majority of clients. In contrast, CAR-MFL shows a near-equal weight distribution for both modalities.

Dynamic Augmentation: We additionally evaluate retrieval samples obtained by our cross-modal augmentation module at various stages of training in Fig. 4. We observed that, on average, an image sample in the unimodal image client is paired with 16 unique text samples, while a text sample in a text client is paired with 9 unique image samples. This evaluation was done on the 8 unimodal client settings (0:8:2 and 8:0:2) over 30 training rounds.

4   Conclusion

We presented a retrieval-based cross-modal augmentation approach for multi-modal federated learning with missing modalities. This approach imputes the

absent modality samples in unimodal clients utilizing a small public multimodal dataset. Our simple yet effective method surpasses previous federated learning frameworks, including the recently proposed state-of-the-art method, CreamFL, while requiring significantly fewer public data samples. Moreover, our approach does not assume a common representation space for different modalities and adds no extra computational overhead to the training procedure, ensuring its broad applicability across various multimodal tasks.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature Medicine **28**(9), 1773–1784 (2022)
2. Chen, J., Pan, R.: Medical report generation based on multimodal federated learning. Computerized Medical Imaging and Graphics p. 102342 (2024)
3. Chen, Y., Liu, C., Huang, W., Cheng, S., Arcucci, R., Xiong, Z.: Generative text-guided 3d vision-language pretraining for unified medical image segmentation. arXiv preprint arXiv:2306.04811 (2023)
4. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. arXiv preprint arXiv:2302.08958 (2023)
5. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Gross, R., Airoldi, E., Malin, B., Sweeney, L.: Integrating utility into face de-identification. In: Danezis, G., Martin, D. (eds.) Privacy Enhancing Technologies. pp. 227–242. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
8. Hao, W., El-Khamy, M., Lee, J., Zhang, J., Liang, K.J., Chen, C., Duke, L.C.: Towards fair federated learning with zero-shot data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3310–3319 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

12. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International conference on machine learning. pp. 5132–5143. PMLR (2020)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lau, K., Adler, J., Sjölund, J.: A unified representation network for segmentation with missing modalities. arXiv preprint arXiv:1908.06683 (2019)
15. Le, H.Q., Thwal, C.M., Qiao, Y., Tun, Y.L., Nguyen, M.N., Hong, C.S.: Cross-modal prototype based multimodal federated learning under severely missing modality. arXiv preprint arXiv:2401.13898 (2024)
16. Lee, H., Kim, W., Kim, J.H., Kim, T., Kim, J., Sunwoo, L., Choi, E.: Unified chest x-ray and radiology report generation model with multi-view chest x-rays. arXiv preprint arXiv:2302.12172 (2023)
17. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
18. Moon, J.H., Lee, H., Shin, W., Kim, Y.H., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Journal of Biomedical and Health Informatics 26(12), 6070–6080 (2022)
19. Qayyum, A., Ahmad, K., Ahsan, M.A., Al-Fuqaha, A., Qadir, J.: Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. IEEE Open Journal of the Computer Society 3, 172–184 (2022)
20. Sachin, D., Annappa, B., Ambasange, S., Tony, A.E.: A multimodal contrastive federated learning for digital healthcare. SN Computer Science 4(5), 674 (2023)
21. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–700. Springer (2022)
22. Shrestha, P., Amgain, S., Khanal, B., Linte, C.A., Bhattarai, B.: Medical vision language pretraining: A survey. arXiv preprint arXiv:2312.06224 (2023)
23. Thrasher, J., Devkota, A., Siwakotai, P., Chivukula, R., Poudel, P., Hu, C., Bhattarai, B., Gyawali, P.: Multimodal federated learning in healthcare: a review. arXiv preprint arXiv:2310.09650 (2023)
24. van Tulder, G., de Bruijne, M.: Learning cross-modality representations from multimodal images. IEEE transactions on medical imaging 38(2), 638–648 (2018)
25. Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D.: Multimodal deep learning models for early detection of alzheimer's disease stage. Scientific reports 11(1), 3254 (2021)
26. Wang, M., Wang, L., Xu, X., Zou, K., Qian, Y., Goh, R.S.M., Liu, Y., Fu, H.: Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging. arXiv preprint arXiv:2303.13033 (2023)
27. Yan, Y., Feng, C.M., Li, Y., Goh, R.S.M., Zhu, L.: Federated pseudo modality generation for incomplete multi-modal mri reconstruction. arXiv preprint arXiv:2308.10910 (2023)
28. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
29. Yu, Q., Liu, Y., Wang, Y., Xu, K., Liu, J.: Multimodal federated learning via contrastive representation ensemble. arXiv preprint arXiv:2302.08888 (2023)

30. Zheng, T., Li, A., Chen, Z., Wang, H., Luo, J.: Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. arXiv preprint arXiv:2302.08646 (2023)
31. Zhou, Q., Zheng, G.: Fedcontrast-gpa: Heterogeneous federated optimization via local contrastive learning and global process-aware aggregation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 660–670. Springer (2023)