Federated Uncertainty-Aware Aggregation for Fundus Diabetic Retinopathy Staging

Meng Wang[1], Lianyu Wang[2], Xinxing Xu[1], Ke Zou[3], Yiming Qian[1], Rick Siow Mong Goh[1], Yong Liu[1], and Huazhu Fu[1](✉)

[1] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore
hzfu@ieee.org

[2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, Jiangsu, China

[3] National Key Laboratory of Fundamental Science on Synthetic Vision and the College of Computer Science, Sichuan University, Chengdu 610065, Sichuan, China

Abstract. Deep learning models have shown promising performance in the field of diabetic retinopathy (DR) staging. However, collaboratively training a DR staging model across multiple institutions remains a challenge due to non-iid data, client reliability, and confidence evaluation of the prediction. To address these issues, we propose a novel federated uncertainty-aware aggregation paradigm (FedUAA), which considers the reliability of each client and produces a confidence estimation for the DR staging. In our FedUAA, an aggregated encoder is shared by all clients for learning a global representation of fundus images, while a novel temperature-warmed uncertainty head (TWEU) is utilized for each client for local personalized staging criteria. Our TWEU employs an evidential deep layer to produce the uncertainty score with the DR staging results for client reliability evaluation. Furthermore, we developed a novel uncertainty-aware weighting module (UAW) to dynamically adjust the weights of model aggregation based on the uncertainty score distribution of each client. In our experiments, we collect five publicly available datasets from different institutions to conduct a dataset for federated DR staging to satisfy the real non-iid condition. The experimental results demonstrate that our FedUAA achieves better DR staging performance with higher reliability compared to other federated learning methods. Our proposed FedUAA paradigm effectively addresses the challenges of collaboratively training DR staging models across multiple institutions, and provides a robust and reliable solution for the deployment of DR diagnosis models in real-world clinical scenarios.

Keywords: Federated learning · Uncertainty estimation · DR staging

M. Wang and L. Wang contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_21.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 222–232, 2023.
https://doi.org/10.1007/978-3-031-43895-0_21

1 Introduction

In the past decade, numerous deep learning-based methods for DR staging have been explored and achieved promising results [10,11,19,27]. However, most current studies focus on centralized learning, which necessitates data collection from multiple institutions to a central server for model training. This approach poses significant data privacy security risks. Additionally, in clinical practice, different institutions may have their own DR staging criteria [3]. Consequently, it is difficult for the previous centralized DR staging method to utilize data of varying DR staging criteria to train a unified model.

Federated learning (FL) is a collaborative learning framework that enables training a model without sharing data between institutions, thereby ensuring data privacy [15,21]. In the FL paradigm, FedAvg [24] and its variants [1,4,9,16,18,22,23] are widely used and have achieved excellent performance in various medical tasks. However, these FL methods assign each client a static weight for model aggregation, which may lead to the global model not learning sufficient knowledge from clients with large heterogeneous features and ignoring the reliability of each client. In clinical practice, the data distributions of DR datasets between institutions often vary significantly due to medical resource constraints, population distributions, collection devices, and morbidity [25,29]. This variation poses great challenges for the exploration of federated DR staging methods. Moreover, most existing DR staging methods and FL paradigms mainly focus on performance improvement and ignore the exploration of the confidence of the prediction. Therefore, it is essential to develop a new FL paradigm that can provide reliable DR staging results while maintaining higher performance. Such a paradigm would reduce data privacy risks and increase user confidence in AI-based DR staging systems deployed in real-world clinical settings.

To address the issues, we propose **a novel FL paradigm, named FedUAA**, that employs a personalized structure to handle collaborative DR staging among multiple institutions with varying DR staging criteria. We utilize uncertainty to evaluate the reliability of each client's contribution. While uncertainty is a proposed measure to evaluate the reliability of model predictions [12,14,28,30], it remains an open topic in FL research. In our work, we introduce **a temperature-warmed evidential uncertainty (TWEU)** head to enable the model to generate a final result with uncertainty evaluation without sacrificing performance. Additionally, based on client uncertainty, we developed **an uncertainty-aware weighting module (UAW)** to dynamically aggregate models according to each client's uncertainty score distribution. This can improve collaborative DR staging across multiple institutions, particularly for clients with large data heterogeneity. Finally, we construct a **dataset for federated DR staging** based on five public datasets with different staging criteria from various institutions to satisfy the real non-iid condition. The comprehensive experiments demonstrate that FedUAA provides outstanding DR staging performance with a high degree of reliability, outperforming other state-of-the-art FL approaches.

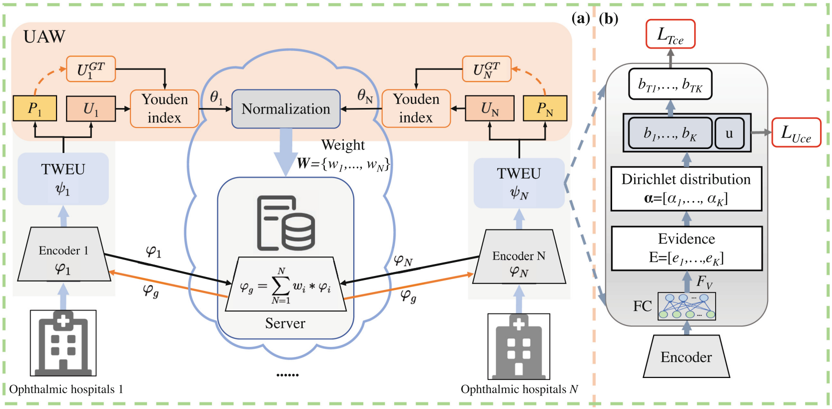

Fig. 1. The overview of FedUAA (a) with TWEU module (b). An aggregated encoder is shared by all clients for learning a global representation of fundus images, while a novel TWEU head is kept on the local client for local personalized staging criteria. Furthermore, a novel UAW module is developed to dynamically adjust the weights for model aggregation based on the reliability of each client.

2 Methodology

Figure 1 (a) shows the overview of our proposed FedUAA. During training, local clients share the encoder (φ) to the cloud server for model aggregation, while the TWEU (ψ) head is retained locally to generate DR staging results with uncertainty evaluation based on features from the encoder to satisfy local-specific DR staging criteria. The algorithm of our proposed FedUAA is detailed in **Supplementary A**. Therefore, the target of our FedUAA is:

$$\min_{\varphi\in\Phi,\psi\in\Psi} \sum_{i=1}^{N} \pounds\left(f_i\left(\varphi_i,\psi_i|X_i\right),Y_i\right), \tag{1}$$

where $\pounds$ is the total loss for optimizing the model, f_i is the model of i-th client, while X_i and Y_i are the input and label of i-th client. Different from previous personalized FL paradigms [2,4], our FedUAA dynamically adjusts the weights for model aggregation according to the reliability of each client, i.e., the client with larger distributional heterogeneity tends to have larger uncertainty distribution and should be assigned a larger weight for model aggregation to strengthen attention on the client with data heterogeneity. Besides, by introducing TWEU, our FedUAA can generate a reliable prediction with an estimated uncertainty, which makes the model more reliable without losing DR staging performance.

2.1 Temperature-Warmed Evidential Uncertainty Head

To make the model more reliable without sacrificing DR staging performance, we propose a novel temperature-warmed evidence uncertainty head (TWEU), which can directly generate DR staging results with uncertainty score based on the features from the encoder. The framework of TWEU is illustrated in Fig. 1 (b). Specifically, we take one of the client models as an example and we assume that the staging criteria of this client is K categories. Correspondingly, given a color fundus image input, we can obtain its $K+1$ non-negative mass values, whose sum is 1. This can be defined as $\sum_{i=1}^{K} b_i + u = 1$, where $b_i \geq 0$ is the probability of i-th category, while u represent the overall uncertainty score. Specifically, as shown in Fig. 1 (b), a local fully connected layer (FC) is used to learn the local DR category-related features F_V, and the *Softplus* activation function is adopted to obtain the evidence $E = [e_1, ..., e_K]$ of K staging categories based on F_V, so as to ensure that its feature value is greater than 0. Then, E is re-parameterized by Dirichlet concentration [5], as: $\boldsymbol{\alpha} = E + 1$, $i.e$, $\alpha_k = e_k + 1$ where α_k and e_k are the k-th category Dirichlet distribution parameters and evidence, respectively. Further calculating the belief masses ($\boldsymbol{b}$) and corresponding uncertainty score (u) by $b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S}$, $u = \frac{K}{S}$, where $S = \sum_{k=1}^{K} \alpha_{i,j}^{k}$ is the Dirichlet intensities. Therefore, the probability assigned to category k is proportional to the observed evidence for category k. Conversely, if less total evidence is obtained, the greater the uncertainty score will be. As shown in Fig. 1 (b), L_{Uce} is used to guide the model optimization based on the belief masses ($\boldsymbol{b}$) and their corresponding uncertainty score (u). Finally, temperature coefficients τ is introduced to further enhance the classifier's confidence in belief masses, i.e., $b_{Ti} = \frac{e^{(b_i/\tau)}}{\sum_{i=1}^{K} e^{(b_i/\tau)}}$, where $\boldsymbol{b_T} = [b_{T1}, ..., b_{Tk}]$ is the belief masses that were temperature-warmed. As shown in Fig. 1 (b), L_{Tce} is adopted to guide the model optimization based on the temperature-warmed belief features of $\boldsymbol{b_T}$.

2.2 Uncertainty-Aware Weighting Module

Most existing FL paradigms aggregate model parameters by assigning a fixed weight to each client, resulting in limited performance on those clients with large heterogeneity in their data distributions. To address this issue, as shown in Fig. 1 (a), we propose a novel uncertainty-aware weighting (UAW) module that can dynamically adjust the weights for model aggregation based on the reliability of each client, which enables the model to better leverage the knowledge from different clients and further improve the DR staging performance. Specifically, at the end of a training epoch, each client-side model produces an uncertainty value distribution (U), and the ground truth for incorrect prediction of U^{GT} also can be calculated based on the final prediction P by,

$$u_i^{GT} = 1 - \mathbf{1}\{P_i, Y_i\}, \text{ where } \mathbf{1}\{P_i, Y_i\} = \begin{cases} 1 & \text{if } P_i = Y_i \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where P_i and Y_i are the final prediction result and ground truth of i-th sample in local dataset. Based on U and U^{GT}, we can find the optimal uncertainty score θ, which can well reflect the reliability of the local client. To this end, we calculate the ROC curve between U and U^{GT}, and obtain all possible sensitivity ($Sens$)

and specificity ($Spes$) values corresponding to each uncertainty score (u) used as a threshold. Then, Youden index (J) [7] is adopted to obtain the optimal uncertainty score θ by:

$$\theta = \arg\max_{u} J\left(u\right), \text{ with } J\left(u\right) = Sens\left(u\right) + Spes\left(u\right) - 1. \tag{3}$$

More details about Youden index are given in **Supplementary B**. Finally, the optimal uncertainty scores $\Theta = [\theta_1, ..., \theta_N]$ of all clients are sent to the server, and a Softmax function is introduced to normalize Θ to obtain the weights for model aggregation as $w_i = e^{\theta_i} / \sum_{i=1}^{N} e^{\theta_i}$. Therefore, the weights for model aggregation are proportional to the optimal threshold of the client. Generally, local dataset with larger uncertainty distributions will have a higher optimal uncertainty score θ, indicating that it is necessary to improve the feature learning capacity of the client model to further enhance its confidence in the feature representation, and thus higher weights should be assigned during model aggregation.

3 Loss Function

As shown in Fig. 1 (b), the loss function of client model is:

$$L = L_{Uce} + L_{Tce}, \tag{4}$$

where L_{Uce} is adopted to guide the model optimization based on the features ($\boldsymbol{b}$ and u) which were parameterized by Dirichlet concentration. Given the evidence of $E = [e_1, ..., e_k]$, we can obtain Dirichlet distribution parameter $\boldsymbol{\alpha} = E + 1$, category related belief mass $\boldsymbol{b} = [b_1, ..., b_k]$ and uncertainty score of u. Therefore, the original cross-entropy loss is improved as,

$$L_{Ice} = \int \left[\sum_{k=1}^{K} -y_k \log\left(b_k\right)\right] \frac{1}{\beta\left(\alpha\right)} \prod_{k=1}^{K} b_k^{\alpha_k - 1} db = \sum_{k=1}^{K} y_k \left(\Phi\left(S\right) - \Phi\left(\alpha_k\right)\right), \tag{5}$$

where $\Phi(\cdot)$ is the digamma function, while $\beta\left(\alpha\right)$ is the multinomial beta function for the Dirichlet concentration parameter α. Meanwhile, the KL divergence function is introduced to ensure that incorrect predictions will yield less evidence:

$$L_{KL} = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K}\left(\tilde{\alpha}_k\right)\right)}{\Gamma\left(K\right)\sum_{k=1}^{K}\Gamma\left(\tilde{\alpha}_i\right)}\right) + \sum_{k=1}^{K}\left(\tilde{\alpha}_k - 1\right)\left[\Phi\left(\tilde{\alpha}_k\right) - \Phi\left(\sum_{i=1}^{K}\tilde{\alpha}_k\right)\right], \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function, while $\tilde{\alpha} = y + (1 - y) \odot \alpha$ represents the adjusted parameters of the Dirichlet distribution which aims to avoid penalizing the evidence of the ground-truth class to 0. In summary, the loss function L_{Uce} for the model optimization based on the features that were parameterized by Dirichlet concentration is as follows:

$$L_{Uce} = L_{Ice} + \lambda * L_{KL}, \tag{7}$$

Table 1. AUC results for different FL methods applied to DR staging.

Methods	APTOS	DDR	DRR	Messidor	IDRiD	Average
SingleSet	0.9059	0.8776	0.8072	0.7242	0.7168	0.8063
FedRep [4]	0.9372	0.8964	0.8095	0.7843	0.8047	0.8464
FedBN [23]	0.9335	0.9003	0.8274	0.7792	0.8193	0.8519
FedProx [22]	0.9418	0.8950	0.8127	0.7877	0.8049	0.8484
FedDyn [1]	0.9352	0.8778	0.8022	0.7264	0.5996	0.7882
SCAFFOLD [16]	0.9326	0.8590	0.7251	0.7288	0.6619	0.7815
FedDC [9]	0.9358	0.8858	0.7969	0.7390	0.7581	0.8236
Moon [18]	0.9436	0.8995	0.8117	0.7907	0.8115	0.8514
MDT [28]	0.9326	0.8908	0.7987	0.7919	0.7965	0.8421
Proposed	**0.9445**	**0.9044**	**0.8379**	**0.8012**	**0.8299**	**0.8636**

where λ is the balance factor for L_{KL}. To prevent the model from focusing too much on KL divergence in the initial stage of training, causing a lack of exploration for the parameter space, we initialize λ as 0 and increase it gradually to 1 with the number of training iterations. And, seen from Sect. 2.1, Dirichlet concentration alters the original feature distribution of F_v, which may reduce the model's confidence in the category-related evidence features, thus potentially leading to a decrease in performance. Aiming at this problem, as shown in Fig. 1 (b), we introduce temperature coefficients to enhance confidence in the belief masses, and the loss function L_{Tce} to guide the model optimization based on the temperature-warmed belief features $\boldsymbol{b_T}$ is formalized as:

$$L_{Tce} = -\sum_{i=1}^{K} y_i log\left(b_{Ti}\right). \tag{8}$$

4 Experimental Results

Dataset and Implementation: We construct a database for federated DR staging based on 5 public datasets, including APTOS (3,662 samples)[1], Messidor (1,200 samples) [6], DDR (13,673 samples) [20], KaggleDR (35,126 samples) (DRR)[2], and IDRiD (516 samples) [26], where each dataset is regarded as a client, More details of datasets are given in **Supplementary C**.

We conduct experiments on the Pytorch with 3090 GPU. The SGD with a learning rate of 0.01 is utilized. The batch size is set to 32, the number of epochs is 100, and the temperature coefficient τ is empirically set to 0.05. To facilitate training, the images are resized to 256×256 before feeding to the model.

[1] https://www.kaggle.com/datasets/mariaherrerot/aptos2019.
[2] https://www.kaggle.com/competitions/diabetic-retinopathy-detection.

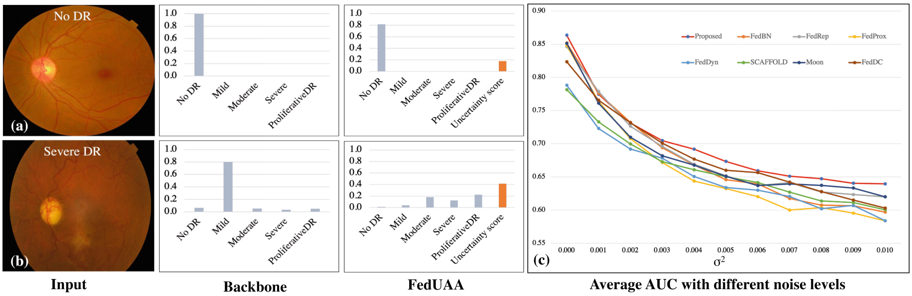

Fig. 2. (a) Instance of being correctly predicted (b) Sample with incorrect prediction result (c) Average AUC of different methods with increasing noise levels (σ^2).

Performance for DR Staging: Table 1 shows the DR staging AUC for different FL paradigms on different clients. Our FedUAA achieves the highest AUC scores on all clients, with a 1.48% improvement in average AUC compared to FedBN [23], which achieved the highest average AUC score among the compared methods. Meanwhile, most FL based approaches achieve higher DR staging performance than SingleSet, suggesting that collaborative training across multiple institutions can improve the performance of DR staging with high data privacy security. Moreover, as shown in Table 1, FL paradigms such as FedDyn [1] and SCAFFOLD [16] exhibit limited performance in our collaborative DR staging task due to the varying staging criteria across different clients, as well as significant differences in label distribution and domain features. These results indicate that our FedUAA is more effective than other FL methods for collaborative DR staging tasks. Furthermore, although all FL methods achieve comparable performance on APTOS and DDR clients with distinct features, our FedUAA approach significantly improves performance on clients with small data volumes or large heterogeneity distribution, such as DRR, Messidor, and IDRiD, by 1.27%, 1.33%, and 1.29% over suboptimal results, respectively, which further demonstrates the effectiveness of our core idea of adaptively adjusting aggregation weights based on the reliability of each client. In addition, we also conduct experiments demonstrate the statistical significance of performance improvement. As shown in Supplementary D, most average p-values are smaller than 0.05. These experimental results further prove the effectiveness of our proposed FedUAA.

Reliability Analysis: Providing reliable evaluation for final predictions is crucial for AI models to be deployed in clinical practice. As illustrated in Fig. 2 (b), the model without introducing uncertainty (Backbone) assigns high probability values for incorrect staging results without any alert messages, which is also a significant cause of low user confidence in the deployment of AI models to medical practices. Interestingly, our FedUAA can evaluate the reliability of the final decision through the uncertainty score. For example, for the data with obvious features (Fig. 2 (a)), our FedUAA produces a correct prediction result with a low uncertainty score, indicating that the decision is reliable. Conversely, even

Table 2. AUC results for different FL paradigms applied to DR staging.

Strategy	BC	EU	TWEU	UAW	APTOS	DDR	DRR	Messidor	IDRiD	Average
SingleSet	✓	✗	✗	✗	0.9059	0.8776	0.8072	0.7242	0.7168	0.8063
	✓	✓	✗	✗	0.9286	0.8589	0.8001	0.7404	0.6928	0.8042
	✓	✗	✓	✗	0.9414	0.8912	0.8279	0.7309	0.7616	0.8306
FL	✓	✗	✗	✗	0.9335	0.9003	0.8274	0.7792	0.8193	0.8519
	✓	✓	✗	✗	0.9330	0.8572	0.7938	0.7860	0.7783	0.8297
	✓	✗	✓	✗	0.9445	0.8998	0.8229	0.8002	0.8231	0.8581
	✓	✗	✓	✓	**0.9445**	**0.9044**	**0.8379**	**0.8012**	**0.8299**	**0.8636**

if our FedUAA gives an incorrect decision for the data with ambiguous features (Fig. 2 (b)), it can indicate that the diagnosis result may be unreliable by assigning a higher uncertainty score, thus suggesting that the subject should seek a double-check from an ophthalmologist to avoid mis-diagnosis. Furthermore, as shown in Fig. 2 (c), we degraded the quality of the input image by adding different levels of Gaussian noise σ^2 to further verify the robustness of FedUAA. Seen from Fig. 2 (c), the performance of all methods decreases as the level of added noise increases, however, our FedUAA still maintains a higher performance than other comparison methods, demonstrating the robustness of our FedUAA.

Ablation Study: We also conduct ablation experiments to verify the effectiveness of the components in our FedUAA. In this paper, the pre-trained ResNet50 [13] is adopted as our backbone (BC) for SingleSet DR staging, while employing FedBN [23] as the FL BC. Furthermore, most ensemble-based [17] and MC-dropout-based [8] uncertainty methods are challenging to extend to our federated DR staging task across multiple institutions with different staging criteria. Therefore, we compare our proposed method with the commonly used evidential based uncertainty approach (EU (L_{Uce})) [12].

For training model with SingleSet, as shown in Table 2, since Dirichlet concentration alters the original feature distribution of the backbone [12], resulting in a decrease in the model's confidence in category-related evidence, consequently, a decrease in performance when directly introducing EU (BC+EU (L_{Uce})) for DR staging. In contrast, our proposed BC+TWEU (L_{Uce}+L_{Tce}) achieves superior performance compared to BC and BC+EU (L_{Uce}), demonstrating that TWEU (L_{Uce}+L_{Tce}) enables the model to generate a reliable final decision without sacrificing performance. For training model with FL, as shown in Table 2, BC+FL outperforms SingleSet, indicating that introducing FL can effectively improve the performance for DR staging while maintaining high data privacy security. Besides, FL+EU (L_{Uce}) and FL+TWEU (L_{Uce}+L_{Tce}) also obtain a similar conclusion as in SingleSet, further proving the effectiveness of TWEU. Meanwhile, the performance of our FedUAA (FL+TWEU (L_{Uce}+L_{Tce})+UAW) achieves higher performance than FL+TWEU (L_{Uce}+L_{Tce}) and FL backbone, especially for clients with large data distribution heterogeneity such as DRR, Messidor, and IDRiD. These results show that our proposed UAW can further improve the performance of FL in collaborative DR staging tasks.

5 Conclusion

In this paper, focusing on the challenges in the collaborative DR staging between institutions with different DR staging criteria, we propose a novel FedUAA by combining the FL with evidential uncertainty theory. Compared to other FL methods, our FedUAA can produce reliable and robust DR staging results with uncertainty evaluation, and further enhance the collaborative DR staging performance by dynamically aggregating knowledge from different clients based on their reliability. Comprehensive experimental results show that our FedUAA addresses the challenges in collaborative DR staging across multiple institutions, and achieves a robust and reliable DR staging performance.

Acknowledgements. This work was supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141, A*STAR Central Research Fund "A Secure and Privacy Preserving AI Platform for Digital Health", and A*STAR Career Development Fund (C222812010).

References

1. Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263 (2021)
2. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)
3. Asiri, N., Hussain, M., Al Adel, F., Alzaidi, N.: Deep learning based computer-aided diagnosis systems for diabetic retinopathy: a survey. Artif. Intell. Med. **99**, 101701 (2019)
4. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International Conference on Machine Learning, pp. 2089–2099. PMLR (2021)
5. Connor, R.J., Mosimann, J.E.: Concepts of independence for proportions with a generalization of the dirichlet distribution. J. Am. Stat. Assoc. **64**(325), 194–206 (1969)
6. Decencière, E., Zhang, X., Cazuguel, G., et al.: Feedback on a publicly distributed image database: the Messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)
7. Fluss, R., Faraggi, D., Reiser, B.: Estimation of the Youden Index and its associated cutoff point. Biometrical J.: J. Math. Methods Biosci. **47**(4), 458–472 (2005)
8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
9. Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., Xu, C.Z.: FEDDC: federated learning with non-IID data via local drift decoupling and correction. In: CVPR, pp. 10112–10121 (2022)

10. Gulshan, V., Peng, L., Coram, M., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402 (2016)
11. Gunasekeran, D.V., Ting, D.S., Tan, G.S., Wong, T.Y.: Artificial intelligence for diabetic retinopathy screening, prediction and management. Curr. Opin. Ophthalmol. **31**(5), 357–365 (2020)
12. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification. arXiv preprint arXiv:2102.02051 (2021)
13. 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)
14. Huang, L., Denoeux, T., Vera, P., Ruan, S.: Evidence fusion with contextual discounting for multi-modality medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 401–411. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_39
15. Kairouz, P., McMahan, H.B., Avent, B., et al.: Advances and open problems in federated learning. Found. Trends® Mach. Learn. **14**(1–2), 1–210 (2021). https://doi.org/10.1561/2200000083
16. 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)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10713–10722 (2021)
19. Li, T., et al.: Applications of deep learning in fundus images: a review. Med. Image Anal. **69**, 101971 (2021)
20. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf. Sci. **501**, 511–522 (2019)
21. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. **37**(3), 50–60 (2020)
22. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. **2**, 429–450 (2020)
23. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FEDBN: federated learning on non-IID features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
24. 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)
25. Nguyen, T.X., et al.: Federated learning in ocular imaging: current progress and future direction. Diagnostics **12**(11), 2835 (2022)
26. Porwal, P., et al.: Indian diabetic retinopathy image dataset (IDRID): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
27. Ting, D.S.W., Cheung, C.Y.L., Lim, G., et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA **318**(22), 2211 (2017)

28. Yu, Y., Bates, S., Ma, Y., Jordan, M.: Robust calibration with multi-domain temperature scaling. Adv. Neural. Inf. Process. Syst. **35**, 27510–27523 (2022)
29. Zhou, Y., Bai, S., Zhou, T., Zhang, Y., Fu, H.: Delving into local features for open-set domain adaptation in fundus image analysis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 682–692. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_65
30. Zou, K., Yuan, X., Shen, X., Wang, M., Fu, H.: TBraTS: trusted brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 503–513. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_48