MULTI-SOURCE-FREE DOMAIN ADAPTATION VIA UNCERTAINTY-AWARE ADAPTIVE DISTILLATION

Yaxuan Song, Jianan Fan, Dongnan Liu, Weidong Cai

School of Computer Science, University of Sydney, Australia

ABSTRACT

Source-free domain adaptation (SFDA) alleviates the domain discrepancy among data obtained from domains without accessing the data for the awareness of data privacy. However, existing conventional SFDA methods face inherent limitations in medical contexts, where medical data are typically collected from multiple institutions using various equipment. To address this problem, we propose a simple yet effective method, named ***U**ncertainty-aware **A**daptive **D**istillation* (UAD) for the multi-source-free unsupervised domain adaptation (MSFDA) setting. UAD aims to perform well-calibrated knowledge distillation from (i) model level to deliver coordinated and reliable base model initialisation and (ii) instance level via model adaptation guided by high-quality pseudo-labels, thereby obtaining a high-performance target domain model. To verify its general applicability, we evaluate UAD on two image-based diagnosis benchmarks among two multi-centre datasets, where our method shows a significant performance gain compared with existing works. The code is available at `https://github.com/YXSong000/UAD`.

***Index Terms*—** Unsupervised Domain Adaptation, Multi-source-free, Uncertainty-ware

1. INTRODUCTION

Unsupervised domain adaptation (UDA) is a promising streamline of works to compensate for the distributional discrepancy [1]. It seeks to utilise existing transferable knowledge from labelled data drawn from one or more source domains to recognise unlabelled data in the target domain [2]. UDA has shown great success in a broad spectrum of downstream applications, including classification [3] [4] [5], segmentation [6] [7] [8] and object detection [9] [10] by mitigating this domain shift.

Despite its great promises in general visual perception tasks, existing UDA approaches inherently fall short in medical scenarios where additional regulations on data sharing restrictions. To address the problems on medical images, source-free DA methods [4] have been developed, providing the pre-trained source model only instead of directly accessing the source data to preserve the privacy issue.

In this work, we investigate multi-source-free unsupervised domain adaptation (MSFDA) [11] [12] and improve the typical SFDA settings [4] [13] by introducing multiple source domains. It therefore holds the potential to serve as an appealing solution for real-world large-scale medical image analysis studies involving multiple centres. Several recent efforts have been made [12] [14] with preliminary attempts to the self-supervised clustering pseudo-labelling method [15], which is commonly adopted for MSFDA. However, they tend to be suboptimal particularly for medical image processing. Since the distinctions of the data from multiple centres are large, the models trained on datasets derived from single or multiple healthcare institutions have not demonstrated a consistent ability to generalise their applicability to external sites [16].

To transcend the aforementioned bottlenecks, in this paper, we propose a framework for MSFDA for medical image analysis. Our contributions include:

1) We propose a novel algorithm termed as ***U**ncertainty-aware **A**daptive **D**istillation* (UAD). Our algorithm first recognises the source model with the most comparable underlying data distribution to the target domain to deliver coordinated model initialisation, and then further leverages the complementary knowledge among source models for precise distillation to the target domain; 2) To avoid over- and under-confidence issues, we apply the Temperature Scaling (TS) method for comprehensive confidence calibration over source models towards a well-regulated knowledge distillation procedure; 3) We substantiate the effectiveness of the proposed method by comparison experiments and ablation studies across diverse scenarios, demonstrating its practical benefits towards various endpoints with clinical significance.

2. METHODS

2.1. Problem Setting

Without involving any source domain data in training the final model, we aim to transfer a series of models, pre-trained on multiple source domains, to a new target domain without any human annotation. In this work, we will consider the K-way classification-model adaptation. We are given a source model zoo $\{\theta_S^j\}_{j=1}^N$, which contains N source classification models from N source domains. For the j-th source model

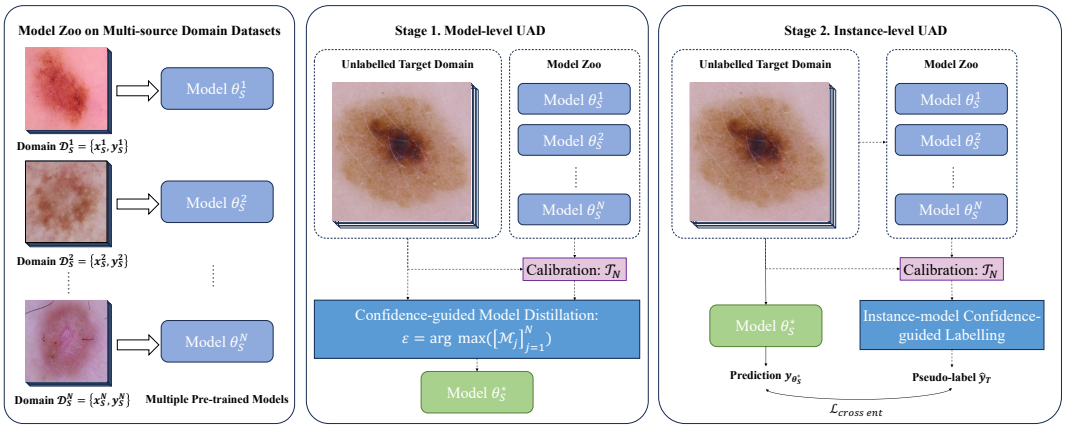

Fig. 1. **Overview of the proposed framework.** Our framework follows a multi-source domain model pre-training process with a two-stage uncertainty-aware adaptive distillation (UAD) process of model initialisation and pseudo-labelling.

θ_S^j in the source model zoo, with the input space being $\mathcal{X}$ and the output space being $\mathcal{Y}$, it is learned by the source dataset $\mathcal{D}_S^j = \{x_{S_j}^i, y_{S_j}^i\}_{i=1}^{n_j}$ with n_j instances, where $x_{S_j}^i \in \mathcal{X}_{S_j}$, $y_{S_j}^i \in \mathcal{Y}_{S_j}$. A target classification model $\theta_T : \mathcal{X} \to \mathbb{R}^K$ is learned by only $\{\theta_S^j\}_{j=1}^N$ and the unlabelled target domain dataset $\mathcal{D}_T = \{x_T^i\}_{i=1}^{n_T}$ with n_T instances.

2.2. Uncertainty-aware Adaptive Distillation

In the proposed framework, we transfer the knowledge from multiple source models to adapt the target domain with pseudo-labels generated by distilling the proper source model. Technically, we learn a set of uncertainty (or its opposite, confidence) measures for both overall domain-wise and individual instance-wise distillation corresponding to each source model in the source model zoo. It evaluates the distributional distance of certain source models working on the target domain dataset and the quality of pseudo-labelling. Specifically, we introduce *margin*, defined as the difference between the predicted probabilities of the first and second most probable classes [17], as the metric to estimate the confidence measure:

$$\mathcal{M} = \underset{k=1}{\text{Topk}}(\delta(\theta(x))) - \underset{k=2}{\text{Topk}}(\delta(\theta(x))), \tag{1}$$

where $\delta(\cdot)$ denotes the Softmax Layer operation with $\delta_j(v) = \frac{\exp(v_j)}{\sum_{i=1}^K \exp(v_i)}$ for $j = 1, ..., K$ and $v \in \mathbb{R}^K \mapsto (0, 1)^K$. Intuitively, if a model θ has a larger value of the margin $\mathcal{M}$ while predicting an instance, it is regarded as more optimal to extract the instance's feature and finally does the classification task.

In order to prevent the trained target domain model from being interrupted by confounding factors incurred by attributed irrelevant to the target task (e.g., image appearance discrepancy due to inconsistent imaging protocols) or avoid local minima problems, we propose to perform Uncertainty-aware Adaptive Distillation (UAD) from two complementary perspectives, (i) model-level and (ii) instance-level, towards directed and well-regularised multi-source model adaptation. The overview of our proposed framework is illustrated in Fig. 1.

Model-level UAD: In previous work related to multi-source domain adaptation [11], it was a common practice to involve all source models with varying weights in the subsequent fine-tuning stage. However, we found that if there is a significant domain gap between a particular source model and the target domain, negative transfer [18] could be incurred which results in biased adaptation. Thus, to initialise a base target model with minimal disturbance, we collect all pre-trained source models from each domain and estimate the overall confidence measure of each source model for predicting the target domain data. Specifically, for assessing the confidence of a source model θ_S^j's inference results on the target domain data, we average all confidence measures estimated for each instance of the target domain data as follows: $\mathcal{M}_\text{j} = \frac{\sum_{i=1}^{n_T} \mathcal{M}_\text{i}}{n_T}$. The source model with the largest confidence measure which is defined as ε for the target domain, θ_S^*, is regarded as the model conforming to the underlying data distribution closest to the target domain and can be considered as the optimal teacher:

$$\varepsilon = \arg\max([\mathcal{M}_\text{j}]_{j=1}^N). \tag{2}$$

We assign the source model θ_S^* as the initial model for SFDA learning on the target data to minimise the gap between the multiple source domains and the target domain.

Instance-level UAD: As the target domain data are not annotated, we propose to use the instance-level UAD method for self-supervised learning on the target data with pseudo labels. Specifically, we sequentially estimate the confidence measure (margin) of each model in the source model zoo for predicting each instance x_T^i, for $i = 1, ..., n_T$, in the target domain and select the most confident source model to generate the pseudo-label:

$$\varepsilon_i = \arg\max([\mathcal{M}_\text{i}]_{i=1}^{n_T}), \tag{3}$$

where $\mathcal{M}_\text{i}$ denotes the margin values of source models pre-

dicting the target domain instance with:

$$\mathcal{M}_{\mathrm{i}} = \Big[\operatorname*{Topk}_{k=1}(\delta(\theta_S^j(x_T^i))) - \operatorname*{Topk}_{k=2}(\delta(\theta_S^j(x_T^i))) \Big]_{i=1,j=1}^{n_T,N}. \quad (4)$$

For the instance x_T^i, the corresponding pseudo-label is obtained by prediction of the source model with $\mathcal{M}_{\mathrm{i}} = \varepsilon_i$, which we define as θ_T^i: $\hat{y}_T^i = \Big[\theta_T^i(x_T^i)\Big]_{i=1}^{n_T}$. $\{x_T^i, \hat{y}_T^i\}_{i=1}^{n_T}$ is leveraged to fine-tune the target initial model $\theta_T = \theta_S^*$ by minimising the standard cross-entropy loss:

$$\mathcal{L}_{tar} = -\mathbb{E}_{(x_T,\hat{y}_T)\in\mathcal{X}_T\times\hat{\mathcal{Y}}_T} \sum\nolimits_{k=1}^{K} \mathbb{1}_{[k=\hat{y}_T]} \log \delta_k(\theta_T(x_T)), \quad (5)$$

where $\mathbb{1}(\cdot)$ gives value 1 when the argument is true.

2.3. Temperature Scaling

In certain models, domain shift and limited data in source domains may result in over- and under-confidence in predicting target domain data which potentially triggers a mismatch between model prediction accuracy and confidence [19]. In other words, when this phenomenon occurs, the confidence measure ε will no longer be an optimal measure for improving model prediction accuracy.

To address this problem, we embedded Temperature Scaling (TS) which acts on prediction probabilities to calibrate the logits prior to confidence measurement. In our approach, TS is capable of effectively regularising the representation of uncertainty in model predictions, and a more precise and unbiased representation of uncertainty is preferable for the process of knowledge distillation. The parameter $\mathcal{T}$ is the so-called temperature, which yields softer probability estimates with larger a temperature to alleviate over-confidence in the model. For every source model $[\theta_S^j]_{j=1}^N$, we learn $\mathcal{T}_{\mathrm{j}}$ by setting an initialisation value $\mathcal{T}_{\text{initial}}$ and applying temperature scaling on the target domain data $\mathcal{D}_T$: $\mathcal{T}_{\mathrm{j}} = \text{TS-Alg}([\theta_S^j]_{j=1}^N, \mathcal{D}_T)$. Specifically, the temperature scaling models are tuned by minimising expected calibration error (ECE), a.k.a., calibration gap, which is defined as the difference between accuracy and confidence for a given bin [20]:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n_T} \Big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\Big|, \quad (6)$$

where M denotes the number of interval bins that we group predictions, and B_m represents the batch of indices of instances allocated in the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$.

Given the logit vector $\theta_S^j(x_T^i)$ obtained from each source model, the calibrated probabilities are estimated by the formula: $z_j = \theta_S^j(x_T^i)/\mathcal{T}_{\mathrm{j}}$, where z_j is the calibrated pre-softmax output (logits) that will be utilised in Sec. 2.2.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Implementation Details

Datasets: We evaluate the proposed multi-source-free domain adaptation framework for classification tasks on two series of datasets:

- Multi-centre Diabetic Retinopathy (*DR*) dataset: The multi-centre DR dataset, which measures DR grades (no DR, mild DR, moderate DR, severe DR and proliferative DR), consists of three public datasets (domains) *APTOS 2019* [21], *DDR* [22], and *IDRiD* [23] with counts 3660, 13673, and 516 correspondingly.
- Skin Cancer MNIST *HAM10000* [24]: To investigate the classification of lesions as benign or malignant in different parts of the human body, we split it into four domains by skin lesion locations which are *back*, *face*, *lower extremity*, and *upper extremity* with counts 2192, 745, 2077 and 1118 respectively.

In our experimental process, we reprocess the data by first resizing into 256×256 and cropping into size 224; then, we assign one domain as the target in turn while considering the others as source domains.

Implementation Details: Following the top-rank solution for medical image classification [25], we employ DenseNet-121 as the backbone. In the source model training process, we use smooth labels instead of the usual one-hot labels to reduce overfitting and label noise. The maximum number of epochs $\mathcal{N}_{\text{epoch}}$ for both DR and HAM10000 datasets is set to 100; while during the UAD process, the $\mathcal{N}_{\text{epoch}}$ is set to 15 with a series of updated pseudo-labels at the start of each. The batch size is set to 32. For each epoch, there are $\mathcal{N}_{\text{training data}}/32$ iterations in domains. We use $\mathcal{T}_{\text{initial}} = \log(1/1.5)$ and 1.5 for the DR dataset and the HAM10000 dataset, respectively. For both source models pre-training and adaptive distillation, we leverage stochastic gradient descent with momentum value 0.9 and weight decay 10^{-3}, with the learning rate scheduling method [3] during the model learning progress.

3.2. Comparison Experiments

For experimental comparison, we included one existing SFDA framework *AaD* [13] with multi-source extension and two MSFDA frameworks *DECISION* [11] and *CAiDA* [12] as baseline methods. We re-implement them following their default settings. The experimental results are reported in Table 1. The multi-source extension of AaD is implemented via an ensemble that passes the target data through each of the adapted source model and takes an average of the soft prediction to obtain the test label. By exploring the experimental results of iterations during the SFDA process for DECISION, we noted that, except for the target domain I in DR and F in HAM10000, the performance of the DECISION model deteriorates as the iterations increase for training the target model. This phenomenon is also observed in the CAiDA framework, although the degradation in model performance in the domain adaptation process is not as severe as in the DECISION framework. Intuitively, in a domain-biased and unsupervised setting, the model overfits to noisy labels when training on the target data. It is due to the effect of the involvement of inappropriate source models and low-quality

Table 1. **Comparison experiments with baselines and ablation study.** For method, M-UAD, I-UAD and TS are abbreviations of *model-level UAD*, *instance-level UAD* and *temperature scaling*. For datasets, A, D and I are abbreviations of *APTOS 2019*, *DDR* and *IDRiD*; B, F, L and U are abbreviations of *back*, *face*, *lower extremity* and *upper extremity*. The first three rows are baselines, and the last four rows are ablation study. All values are adaptation accuracy (%). The last row is our default method setting and corresponding experimental result.

Method	DR				HAM10000				
	D, I → A	A, I → D	A, D → I	AVG.	F, L, U → B	B, L, U → F	B, F, U → L	B, F, L → U	AVG.
AaD (22') [13]	36.13	33.07	46.32	38.51	64.55	64.30	65.14	72.36	66.59
DECISION (21') [11]	57.32	45.43	58.33	53.69	74.27	**76.24**	71.06	78.98	75.14
CAiDA (21') [12]	71.74	44.98	50.97	55.90	73.68	73.83	79.59	78.80	76.48
M-UAD	71.49	62.03	50.39	61.30	81.84	68.19	87.48	83.27	80.20
I-UAD	72.91	63.71	53.10	63.24	84.58	69.66	88.78	83.09	81.53
M-UAD + I-UAD	74.47	64.39	53.88	64.25	85.40	71.41	89.41	84.08	82.58
M-UAD + I-UAD + TS	**74.52**	**65.27**	**58.72**	**66.17**	**85.40**	73.29	**89.70**	**84.44**	**83.21**

pseudo-labels generated.

In comparison with existing frameworks, our proposed method effectively mitigates both factors that could potentially diminish the performance of the target domain model: we identify the most confident source model, excluding inappropriate ones from participating in the training of the target model, and generate the most reliable pseudo-labels through the optimal source model. The last row in Table 1 shows that the average accuracy of domain adaptation via UAD (our method) in both datasets significantly outperforms all the baselines.

3.3. Ablation Study

Furthermore, we also performed an ablation study on the domain adaptation process: the model-level UAD only without training implementation, the instance-level UAD only without training implementation, and the model-level and instance-level UAD with training but without temperature scaling.

Effectiveness on Model-level and Instance-level UAD: To avoid inappropriate source model(s), which are learned by the source domain data that deviates significantly from the target domain data distribution, from disrupting the final performance of the target domain model, we first propose the exclusion of such disruptive source model(s) during the training process. Instead, using the model-level UAD (M-UAD) method, we pick the most confident source model, which is also the optimal choice among existing models, to serve as the initialisation of training the target model process. This establishes a solid foundation in the early stages of model training. The first row of the ablation study (M-UAD) in Table 1 demonstrates the result that implementing only M-UAD leads to an improvement of approximately 5% on average compared to the baseline results.

In an unsupervised learning setting, the generation of pseudo-labels is a crucial step in driving the eventual high-performance model. Instead, the generation of low-quality pseudo-labels leads the target model to gradually fit into these noisy labels, thereby reducing the final performance of the target model. To prevent this from occurring, we propose using the instance-level UAD (I-UAD) method to identify the most confident label corresponding to an individual instance as its pseudo-label. The second row of the ablation study (I-UAD) in Table 1 gives the experimental result that applying the I-UAD method leads to a higher accuracy for the target model compared to the M-UAD approach.

The third row of the ablation study (M-UAD + I-UAD) in Table 1 gives the experimental result that the performance can be further improved by jointly applying the two-level UAD.

Effectiveness on Temperature Scaling: According to Sec. 2.3, to mitigate the problem of over- and under-confidence in certain model(s) predicting the target domain data, TS is an effective method to calibrate the model. The last row of Table 1 gives the experimental result of applying the TS approach to our combined UAD framework, showing an improvement in the average accuracy compared to without applying the TS model calibration method. This effect is particularly pronounced on some target domains with relatively low accuracy, such as domains I and F in the DR and HAM10000 datasets respectively.

4. CONCLUSION

In this study, we proposed a two-level uncertainty-aware adaptive distillation method termed UAD, a novel deep learning framework for multi-source-free unsupervised domain adaptation on medical imaging data, with successful application on datasets across diseases and human anatomical regions. Both initialising the target domain training process by identifying the optimal source model and generating reliable pseudo-labels by leveraging a post-calibrated source model zoo, our method significantly outperforms the existing frameworks performing on the medical imaging data. In conclusion, our proposed method can fill the gap in the MSFDA setting in the field of medical image processing and analysis.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [21–24]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, and Weidong Cai, "Taxonomy adaptive cross-domain adaptation in medical imaging via optimization trajectory distillation," in *ICCV*, 2023, pp. 21174–21184.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," *Machine Learning*, pp. 151–175, 2010.

[3] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[4] Jian Liang, Dapeng Hu, and Jiashi Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *ICML*, 2020, pp. 6028–6039.

[5] Canran Li, Dongnan Liu, Haoran Li, Zheng Zhang, Guangming Lu, Xiaojun Chang, and Weidong Cai, "Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling," in *MICCAI*, 2022, pp. 715–724.

[6] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai, "Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting," in *CVPR*, 2020, pp. 4243–4252.

[7] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu, "Cris: Clip-driven referring image segmentation," in *CVPR*, 2022, pp. 11686–11695.

[8] Jianan Fan, Dongnan Liu, Hang Chang, and Weidong Cai, "Learning to generalize over subpartitions for heterogeneity-aware domain adaptive nuclei segmentation," *International Journal of Computer Vision*, 2024.

[9] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang, "Progressive domain adaptation for object detection," in *WACV*, 2020, pp. 749–757.

[10] Dongnan Liu, Chaoyi Zhang, Yang Song, Heng Huang, Chenyu Wang, Michael H Barnett, and Weidong Cai, "Decompose to adapt: Cross-domain object detection via feature disentanglement," *IEEE Transactions on Multimedia*, pp. 1333–1344, 2022.

[11] Sk Miraj Ahmed, Dripta S. Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K. Roy-Chowdhury, "Unsupervised multi-source domain adaptation without access to source data," in *CVPR*, 2021, pp. 10103–10112.

[12] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu, "Confident anchor-induced multi-source free domain adaptation," in *NeurIPS*, 2021, pp. 2848–2860.

[13] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer, "Attracting and dispersing: A simple approach for source-free domain adaptation," in *NeurIPS*, 2022.

[14] Zhongyi Han, Zhiyan Zhang, Fan Wang, Rundong He, Wan Su, Xiaoming Xi, and Yilong Yin, "Discriminability and transferability estimation: a bayesian source importance estimation approach for multi-source-free domain adaptation," in *AAAI*, 2023, pp. 7811–7820.

[15] Zhaoqing Wang, Ziyu Chen, Yaqian Li, Yandong Guo, Jun Yu, Mingming Gong, and Tongliang Liu, "Mosaic representation learning for self-supervised visual pre-training," in *ICLR*, 2023.

[16] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, p. e1002683, 2018.

[17] Burr Settles, "Active learning literature survey," *University of Wisconsin-Madison Department of Computer Sciences*, 2009.

[18] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell, "Characterizing and avoiding negative transfer," in *CVPR*, 2019, pp. 11285–11294.

[19] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs, "Soft calibration objectives for neural networks," in *NeurIPS*, 2021, pp. 29768–29779.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017, pp. 1321–1330.

[21] Sohier Dane Karthik, Maggie, "Aptos 2019 blindness detection," *Kaggle*, 2019.

[22] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Information Sciences*, pp. 511–522, 2019.

[23] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," *Data*, 2018.

[24] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, pp. 1–9, 2018.

[25] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng, "Federated semi-supervised medical image classification via inter-client relation matching," *MICCAI*, 2021.