applied sciences

Article

Class Imbalanced Medical Image Classification Based on Semi-Supervised Federated Learning

Wei Liu [1], Jiaqing Mo [1,*] and Furu Zhong [2]

[1] Xinjiang Key Laboratory of Signal Detection and Processing, College of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
[2] School of Physics and Electronic Science, Zunyi Normal College, Zunyi 563006, China
* Correspondence: xjumojiaqing@163.com; Tel.: +86-139-9999-7550

Abstract: In recent years, the application of federated learning to medical image classification has received much attention and achieved some results in the study of semi-supervised problems, but there are problems such as the lack of thorough study of labeled data, and serious model degradation in the case of small batches in the face of the data category imbalance problem. In this paper, we propose a federated learning method using a combination of regularization constraints and pseudo-label construction, where the federated learning framework consists of a central server and local clients containing only unlabeled data, and labeled data are passed from the central server to each local client to take part in semi-supervised training. We first extracted the class imbalance factors from the labeled data to participate in the training to achieve label constraints, and secondly fused the labeled data with the unlabeled data at the local client to construct augmented samples, looped through to generate pseudo-labels. The purpose of combining these two methods is to select fewer classes with higher probability, thus providing an effective solution to the class imbalance problem and improving the sensitivity of the network to unlabeled data. We experimentally validated our method on a publicly available medical image classification data set consisting of 10,015 images with small batches of data. Our method improved the AUC by 7.35% and the average class sensitivity by 1.34% compared to the state-of-the-art methods, which indicates that our method maintains a strong learning capability even with an unbalanced data set with fewer batches of trained models.

Keywords: federated learning; semi-supervised algorithm; pseudo-labels; classification

check for updates

Citation: Liu, W.; Mo, J.; Zhong, F. Class Imbalanced Medical Image Classification Based on Semi-Supervised Federated Learning. *Appl. Sci.* **2023**, *13*, 2109. https://doi.org/10.3390/app13042109

Academic Editor: Jan Egger

Received: 8 December 2022
Revised: 24 January 2023
Accepted: 3 February 2023
Published: 6 February 2023

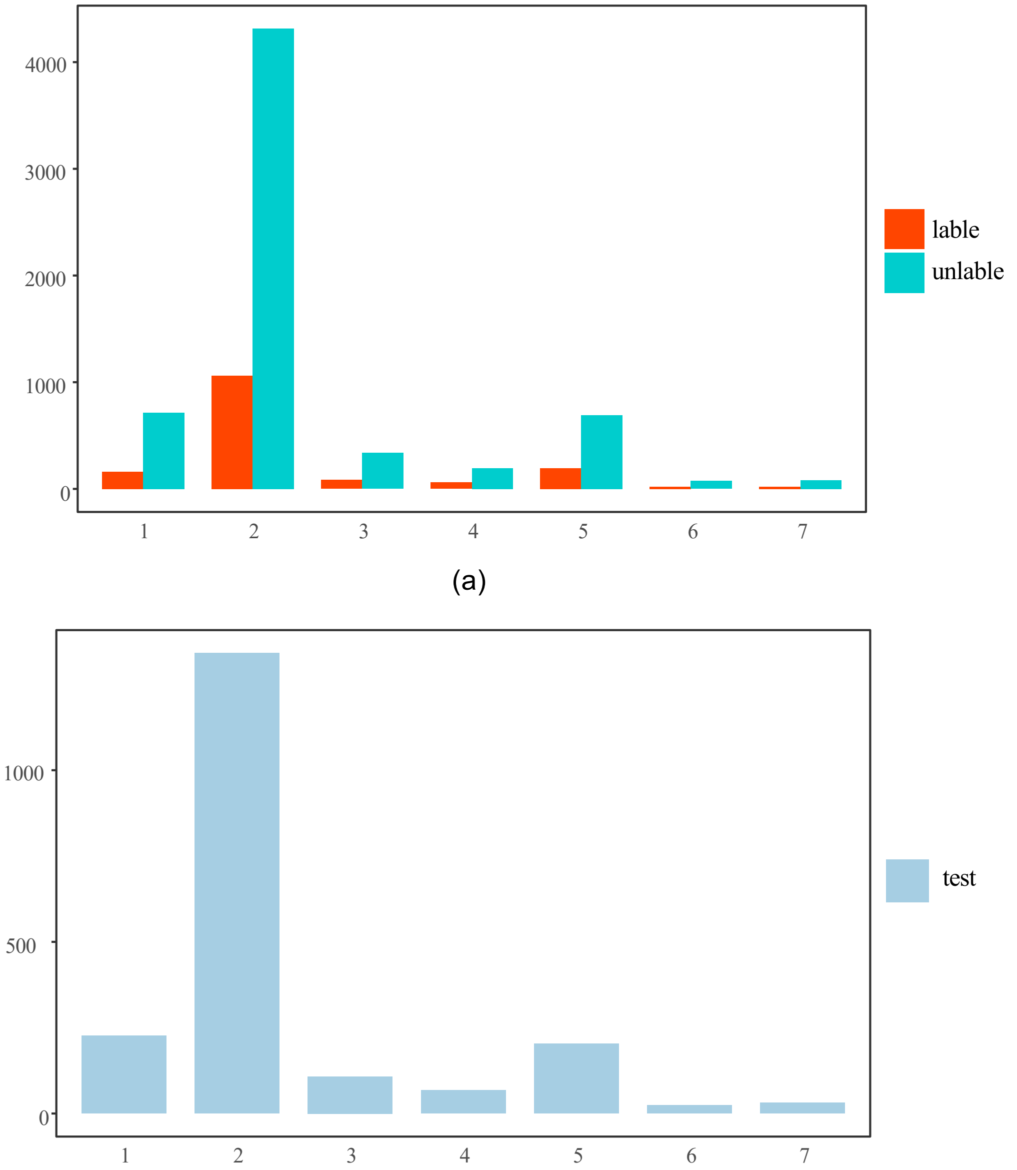
1. Introduction

Compared with traditional natural image classification, the following two problems are common for all kinds of medical image data in medical image classification tasks: One is data scarcity and category imbalance caused by privacy protection [1,2], and the other is that it is difficult to get tags [3] due to high tagging cost and lack of professional talents in some regions. In recent years, federated learning has received a lot of attention from researchers due to its ability to total various idiosyncratic data by fusing distributed training network parameters with no interaction between data in various places. This guarantees data privacy, based on these advantages, which makes it compatible with medical tasks [4–9]. Through our research survey we noticed that concerning federated learning methods most of the current research revolves around supervised training [10,11], some of the research applies valid features from labeled data to unlabeled data [12], and there is less research on class imbalance within the data. On the difficulty of acquiring balanced data: Due to the large differences in the incidence of disease types among localities [13] and the inability of some hospitals to effectively label the huge amount of data, how to apply a large amount of unlabeled data to federated learning and overcome the internal class imbalance problem to improve the accuracy of the network becomes the main problem to be solved in this paper.

To this end, we investigated an improved semi-supervised federated learning problem that introduces a treatment of the unlabeled-data class-imbalance problem based on the fact that only a small amount of labeled data is involved. To obtain useful information from unlabeled data, two methods are commonly used: One uses regularization to constrain unlabeled data using labeled data features [12,14–17], and the other uses a specific way to construct pseudo-labels [18–20] and set confidence thresholds to filter the predicted features. Liu et al. [12] set a small amount of labeled data in a local client and achieved regularization constraints by extracting the intrinsic relationships of the categories in the labeled images in the form of a correlation matrix to laterally constrain the unlabeled data from other clients. Although further exploitation of labeled data is carried out, the training on unlabeled data only incorporates common consistent regularization methods to improve the robustness of the model and does not effectively address the problem of category imbalance at the data level. The difference is that Bdair et al. [21] used labeled data only as a model initialization training and focused their research on the training of local clients, where they applied the peer anonymous (PA) learning method to unlabeled clients, integrating similar clients to the community and obtaining pseudo-labeling of unlabeled data with the aid of peer learning. This method improved the training efficiency but still suffers from category imbalance at the data level. To solve this problem, Jiang et al. [18] used labeled data for pre-training and performed dynamic library building in a local client containing only unlabeled data. They transformed the original classification task into classification of the library by randomly setting different label ratios to avoid the problem of blurred decision boundaries in low categories caused by unbalanced data categories. This approach largely alleviates the problem of data-level imbalance, but ignores the value of labeled data and requires training in large batches to optimize the class imbalance problem, otherwise the client experiences severe model degradation in some places, which leads to serious degradation in the performance of the aggregated model, as also confirmed in the experimental Section 3.2. Meanwhile, in FedAvg [22], FedIRM [12], FedPerl [21], and imFed-Semi [18] studies, which have federated learning as a framework, it is common to divide the ratio of labeled to unlabeled data in the experimental data into 2:8, where the labeled data are divided into a small number (1–4) of labeled local clients. The unlabeled data are divided into a majority of unlabeled local clients, and the studies do not investigate the effect of a different number of local clients on model classification. In addition, the performance of the model is only measured using various metrics in the model validation phase, and there is no research on how to represent the robustness of the model.

Based on the above problems, how to effectively extract valuable information from labeled data and better overcome the class imbalance problem has become a key issue. The thought is to apply and process the labeled data several times so that it can play an auxiliary role in the process of client-side training. Based on the similarity index of disease presentation [23], it was indicated that the distribution of diseases between different categories of the same type still presents a high degree of similarity across geographic areas. With this discovery, we consider that the validity of training in each client can be guaranteed with a reasonable distribution of data, and this makes it possible to supervise the pseudo-labeling of its predictions. In addition, we think it is also significant to choose what kind of unlabeled data should participate in the training, because the data selection can mitigate the task loss of unlabeled clients and effectively utilize the unlabeled samples.

In our article, we present a semi-supervised federated learning approach based on a combination of regularization and pseudo-label construction. We labeled data for model pre-training, local client fusion training, and constraining the generation of pseudo-labels inside each client. Additionally, to better shape the decision boundaries for fewer categories in unlabeled data, we considered using a data selector to filter partially stable unlabeled data in combination with the existing SSL algorithm [24]. To validate the performance of the model, we present a method to measure the robustness of the model. We experimentally validated it on the ISIC 2018: skin lesion analysis for melanoma detection data set [25] with the class sample distribution of the training and test sets as shown in Figure 1. Our method improves all metrics compared to other advanced federated learning methods [12,18,22] for the same data sample, and the model has optimal robustness.



(a)

(b)

Figure 1. (**a**) The distribution of samples with labels in the central server is shown in red, and the distribution of samples without labels in the local client is shown in green. (**b**) Sample distribution of test set categories for model detection.

The main contributions of our article are as follows:

- To the extent of our knowledge, we present the first approach that combines regularization constraints with pseudo-label construction in solving a federated learning for medical information classification tasks.
- We propose a stable selector to filter unlabeled data to improve the robustness of the model and the pseudo-labeling information.
- We construct controllable data samplers that can divide labeled and unlabeled data in arbitrary proportions to explore the impact of different numbers of clients on the model.

- We suggest a measure of model robustness that measures the sensitivity of the model to the class of data.

2. Materials and Methods

2.1. Data Set and Task Setup

Data set: We used the ISIC 2018: dermal lesion analysis for melanoma detection data set [25] for dermal lesion diagnosis, which collected dermoscopic images from different populations, consisting of 10,015 dermoscopic images of cases. These include a representative collection of all important diagnostic categories in the field of pigmented lesions: actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibromas (df), melanomas (mel), melanocytic nevi (nv), and vascular lesions (vasc).

In our task, we let x_l be labeled data, x_u be unlabeled data, y be true labels, $\overline{y}$ be pseudo-labels, $D_L = \{x_l^1, x_l^2 \cdots x_l^N\}$ be the labeled data set, $D_L = \{x_u^{N+1}, x_u^{N+2} \cdots x_u^{N+M}\}$ be the unlabeled data set, and $C = \sum_{i=1}^{n} c_i$ be the local client set, where $c_i = \sum_{j=1}^{M/n} O_j(D_U)$, n is the number of clients, O is the random sampler, $K = \{m_1, m_2 \cdots m_k\}$ is the distribution of each category in the labeled data set, k is the number of categories, and $y = \min\limits_{K \in D_L}(K)/K$ is the imbalance factor.

2.2. Federated Learning Method

As shown in Figure 2, our approach is to pre-warm the labeled data using the central server network $f(\theta)$, pass the network parameters to the local network by broadcasting, pass the labeled data to each local client to train the local network $f_i(\theta)$ jointly with the unlabeled data, use the imbalance factor to construct the equilibrium auxiliary to do regularization constraints at the representation layer of the local network, and use the aggregation algorithm FedAvg [22] to aggregate the network of each local client to get the updated central server network $f(\theta) = \sum_{i=1}^{n} c_i \times f_i(\theta)/C$. Then, the global network is passed to the local client network again by broadcasting for the next iteration until convergence.

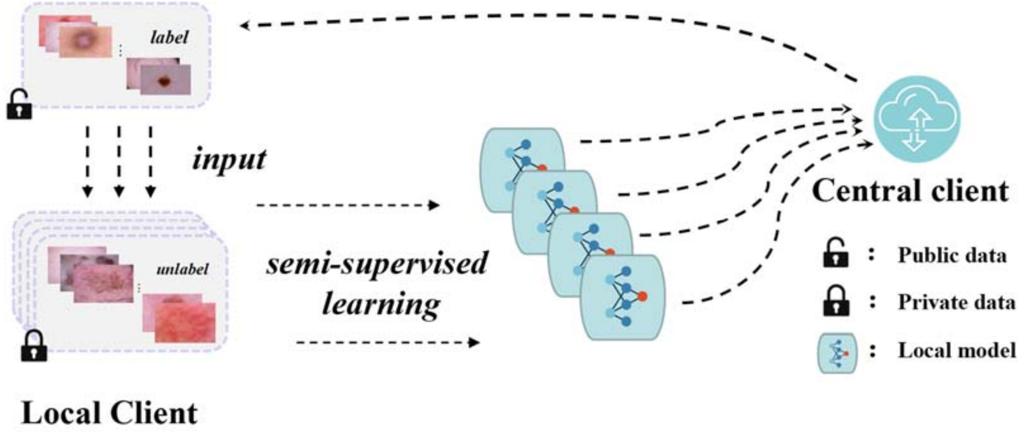


Figure 2. Framework of our proposed federated learning approach.

In the local client training process we used ReMixMatch [24], which has the most advanced performance in the field of semi-supervised learning, as the main body. We first used the pre-trained network for pseudo-label prediction, taking the average of the output of each batch as one prediction, and using a distributed alignment method to match the pseudo-label with the real label, with the following expression

$$Y = K/f_i(p(\eta|x_u), \theta) \tag{1}$$

where $p(\eta|x_u)$ is the primary prediction of the unlabeled data. We finally obtained the pseudo-label $\overline{y} = \tau(Y)$, where $\tau(\cdot)$ is the temperature sharpening function. The second method used was anchored augmentation, i.e., the same input data were randomly inverted and cropped to form weakly augmented data, and the augmented labeled data were mixed with unlabeled data for regularization. Finally, using the cross-entropy loss as a metric, we introduced a self-supervised rotation loss l_r for consistent regularization of unlabeled images in order to improve the sensitivity and stability of the network to unlabeled images under different perturbations [26,27].

2.3. Data Distribution

We followed the previous method of data assignment and divided the ratio of data with and without labels into 2:8. However, in our method, labeled data were very significant, and since the labeled information needs to be fully utilized to extract the class imbalance factor and fuse it with the unlabeled data to join the training network, as shown in Figure 3, we set the number of labeled local clients to 0. The unlabeled data can be assigned to 0 to 8 local clients by parameter control, and the advantage of this setting is that we can change the training pattern at any time. When the unlabeled local client is 0, the model can be pre-trained to achieve the initialization of the model; when the unlabeled local client is 1, the training mode degenerates from federated learning to ordinary semi-supervised learning, so that the differences between the two modes can be compared under the same data. Furthermore, the optimal number of unlabeled clients can be derived through experiments to achieve optimal model performance.

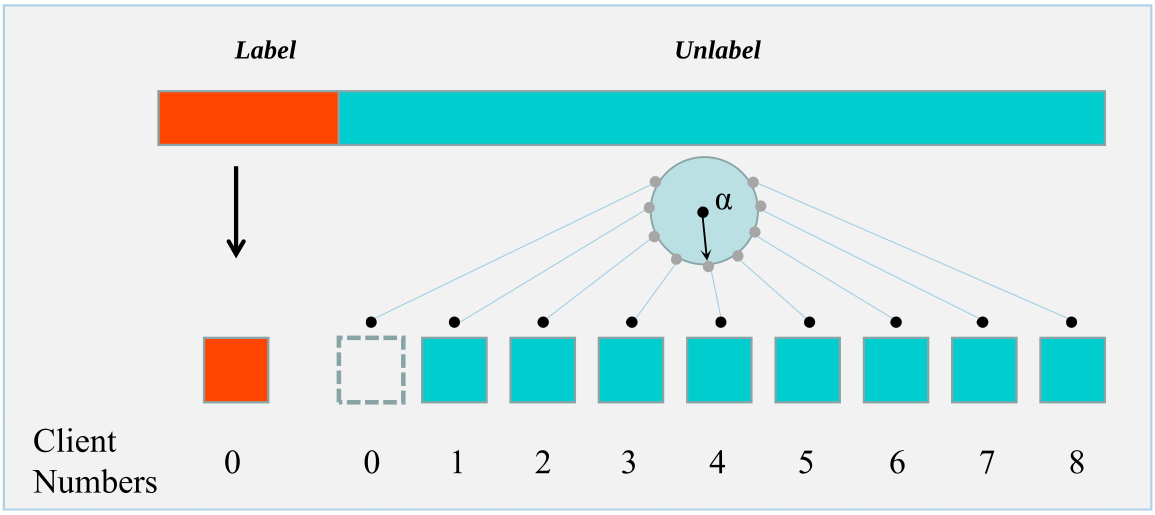


Figure 3. Controlled label-free data distribution method.

2.4. Class Balance Auxiliary

To solve the experimental class imbalance problem, we introduced the class balancing aid proposed by Lee et al. [20]. The schematic diagram is shown in Figure 4. Experimentally, the central client has labeled data for category statistics to obtain the imbalance factor ς . Using the Bernoulli distribution $B(\cdot)$ to generate the mask $I = B(y \times \varsigma)$, so that a few class labels have a higher probability of generating 1, while most category labels generate 0 with a higher probability, with label balancing loss can be calculated as follows:

$$l_{abc} = I \times H(f_i(p(\eta|x_l), \theta), y) \tag{2}$$

where $H(\cdot)$ is the standard cross-entropy loss and $(p(\eta|x_l), \theta)$ is the prediction of the local network on the labeled data.

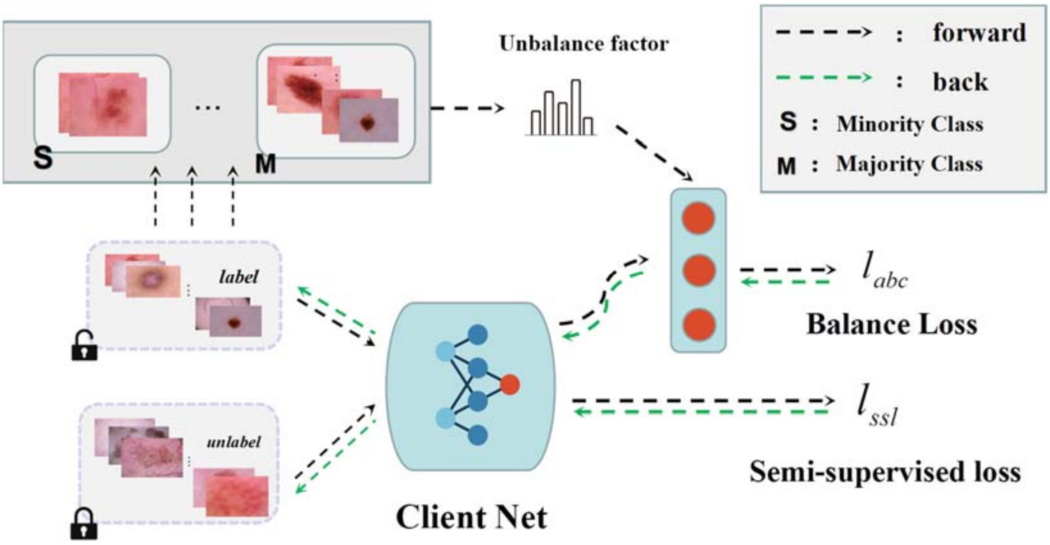


Figure 4. The training method in the local client network and the composition of the class-balanced auxiliary.

To improve the accuracy of label-free prediction, we introduced a confidence level $\gamma(\cdot)$. The data are used as a loss calculation only when the probability of predicting a label is greater than the confidence threshold. The unlabeled equalizer is also adjusted, because the unlabeled data do not have accurate label information; if the category is directly balanced, it is easy to cause the model to forget most of the category information. In order to ensure that the model can both fully learn the multi-category information and adapt to the class imbalance challenge, we introduced dynamic weights. The weight parameter is affected by the iteration cycle, in the initial training the imbalance factor is constant to 1, so it loses its balancing role. As the iteration cycle increases the imbalance factor gradually recovers to continue working; the unlabeled balance loss is calculated as follows:

$$l_{abcu} = \gamma(f_i(p(\eta|x_u),\theta),\omega) \times I \times H(f_i(p(\eta|x_u),\theta),y) \quad (3)$$

where $\gamma(\cdot)$ is the confidence function and ω is the confidence threshold.

2.5. Pseudo-Label Construction

In the construction of semi-supervised loss, we borrowed the pseudo-label construction framework of ReMixMatch [24]. However, medical images require higher stability of the model in the classification task compared to natural images, because the classification of disease recognition often depends on local pixel points. If the model is not strong against interference, the model obtains more ambiguous label information when the unlabeled images are less informative or when the image quality is poor. To address this problem, we present the stable selector pseudo-label construction method, which aims to further improve the stability of the network for unlabeled images by screening more stable samples of unlabeled images to generate pseudo-labels with sufficient label information.

The stability selector estimates the predictive stability of the model for unlabeled samples by calculating the difference in probability distribution between the features of the unlabeled augmented samples and the perturbed adversarial samples after going through the model. The process is as follows: we pass a batch of unlabeled weakly enhanced data x_u through the pre-training network $f_i(\theta)$ to obtain the general label prediction y_i^u. At the same time, we pass the added perturbed unlabeled adversarial sample $d(x_u)$ into the network to obtain the perturbed label prediction $d\left(y_i^u\right)$. Finally, we input both y_i^u and $d\left(y_i^u\right)$ into the generator to obtain the adversarial perturbation vector r_i^u. r_i^u is denoted as

$$r_i^u = arg\min_K Div(y_i^u, d(y_i^u)) \tag{4}$$

where K is the maximum selection hyperparameter and $Div(\cdot)$ is a function of the scatter between the predicted labels of the augmented samples and the perturbed adversarial samples. In our experiments we used the KL scatter.

However, the above selector has a problem, because the selector uses randomly generated perturbation terms in the interval of ±0.5, because when the perturbation is 0, the gradient of $Div(\cdot)$ with respect to the perturbation is constant 0. At this time, the selector is not able to improve the network stability. For this problem we use a sample once for the network training scatter, with respect to the average gradient of the perturbation to approximate the perturbation, which is represented in the experiment as follows.

$$\overline{g} = \nabla_{d(x_u)} Div(y_i^u, d(y_i^u)) \tag{5}$$

$$r_i^u = arg\min_K Div(f_i(p(\eta|x_u), \theta); f_i(p(\eta|\overline{g}), \theta)) \tag{6}$$

where ∇ is the average gradient operation. Finally, we selected the predicted labels corresponding to the first K-minimum scatter to participate in the pseudo-label construction.

Then we introduced the distribution alignment proposed by ReMixMatch [24], which makes the predicted labels fit the class ratio to further overcome the class imbalance problem; The semi-supervised pseudo-label construction process is shown in Figure 5. First, we designate the K stable predicted pseudo-labels screened as the prediction q of one batch of unlabeled data and define the edge class distribution as $Y_{dist} = \varsigma \times \left(N_{D_L}/N_{D_U}\right)$. Then, the K unlabeled data after each round of screening are input into the network to get the label prediction and are stored. When the next round of prediction is finished, it will be averaged with the previous prediction, and when the 129th round of prediction is finished, it will replace the prediction of the 1st round, and the average of each round. The prediction after each round is averaged so that it is the distribution of unlabeled data predictions $\overline{Y}_{dist}$. The final pseudo-label is defined as

$$\overline{y} = Norm(q \times (Y_{dist}/\overline{Y}_{dist})) \tag{7}$$

where $Norm(\cdot)$ is the normalization process.

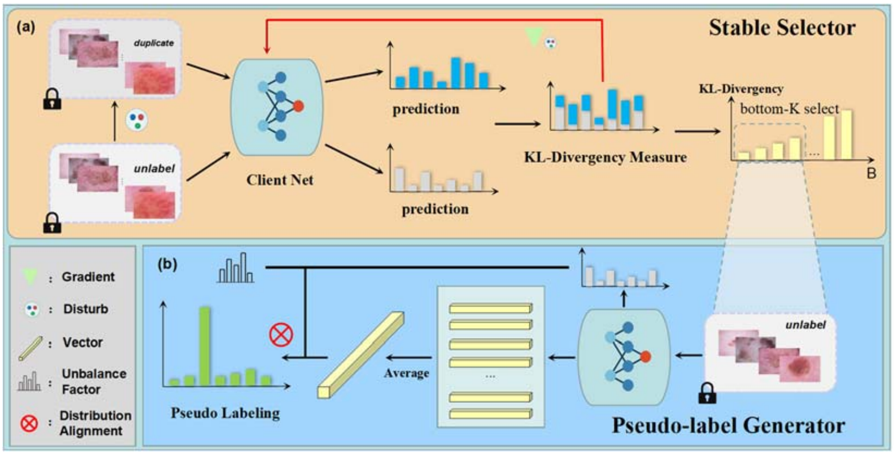


Figure 5. The pseudo-label construction method consists of two parts: (**a**) Data stability selector, which selects the optimal data by measuring the difference of data distribution in different states of the same picture. (**b**) Pseudo-label generator, which combines the probability distribution and distribution alignment generated by the screened pictures into the network to generate label predictions.

2.6. Model Stability Validation

To verify that our method combining label regularization with pseudo-label construction and a data selector is helpful for the stability of the model, we propose a simple way to measure the stability of the training completion model, i.e., two different validation methods were selected to experimentally validate the test set: One is the traditional single-label multi-classification validation method, used by Jiang et al. [18] in their experiments, which is characterized by selecting the larger probability ordinal number in a single image as the predicted label y_1. However, there are cases where the probability of multiple ordinal numbers is approximated, which indicates a bias in the model's classification of that image. The second is the multi-label multiclassification validation method, used by Lee et al. [20] in their experiments, which features a hyperparameter threshold (<0.5) and then compares the data by category and selects the ordinal number corresponding to a probability greater than the threshold as the label y_2 for that image. This allows for the presence of multiple labels and clearly shows the bias of the model on the classification of that image. When the model is sensitive to each category, the probability corresponding to the correct category is much higher than the incorrect category, and y_1 and y_2 are equal, but if the model is less sensitive to certain categories, y_1 and y_2 are not equal. According to this characteristic, we chose to use the sensitivity metric to calculate the error of the two validation methods to assess the stability of the model, and the error representation formula is as follows:

$$\varepsilon = \left| S\left(\sum_{i=1}^{N} y_1^i\right) - E\left[\sum_{j=1}^{K} S_j(y_2)\right] \right| \tag{8}$$

where N is the number of validation sets, K is the number of categories, $S(\cdot)$ is the sensitivity metric calculation method, E is the expectation calculation, and $|\cdot|$ indicates that the absolute value is taken.

3. Results

3.1. Experimental Setup and Details

Experimental setup: We trained 80% of the image data for the federated network and 20% for the final network test, and followed the COVID-Net [28] scheme for the preprocessing of the data to perform a random transformation of the image with translation, rotation, and flip. The image size was adjusted to 224 × 224 and normalized, and the pre-network initialization was performed using the network parameters input to the network pre-trained by ImageNet. Through experimental observations, we set the number of local clients as four, and used random sampling to equally divide the randomly disrupted image data into five subsets, of which one subset was assigned to the central server as labeled data and four subsets were assigned to the local clients as unlabeled data. The evaluation metrics consisted of AUC, sensitivity, specificity, accuracy, precision, and F1. The performance of the central server network was analyzed comprehensively, and the mean value of the best metrics from three independent trainings was used as the final result. Their computational relationships are as follows:

$$\text{Sensitivity} = \text{TP}/\text{TP} + \text{FN} \tag{9}$$

$$\text{Specificity} = \text{TN}/\text{TN} + \text{FP} \tag{10}$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{TN} + \text{FP}) \tag{11}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{12}$$

$$\text{F1} = 2(\text{Precision} \times \text{Sensitivity})/(\text{Precision} + \text{Sensitivity}) \tag{13}$$

Among them, sensitivity and specificity metrics are very important metrics in medical image classification tasks, but these two metrics are contradictory, so we chose to use Sensitivity metrics to measure the performance of the model in our experiments. Additionally, in our study, the AUC metric is very informative because it is not sensitive to whether the samples are balanced or not and is more suitable for dealing with unbalanced data distributions. In addition, the AUC is more sensitive to model changes and the larger the value of the AUC, the more likely the current classification algorithm will rank the positive samples, i.e., will be able to classify them better [29].

Implementation details: To ensure fair consistency of the experiments we chose the network DenseNet121 consistent with FedIRM [12] and 30 cycles of warm-up; performed 200 cycle iterations, data batch size was set to 10; did random augment [30] data enhancement on the input images; and added Dropout layer to the network to prevent overfitting. Confidence level of network parameters threshold was 0.95, batch training was performed using Adam optimizer with momenta of 0.9 and 0.99, and the learning rate was set to 2×10^{-4}.

3.2. Comparison with Advanced Methods

We set up controlled experiments including (1) a FedAvg [22] model using only 20% of the central client with labeled data and using it as a baseline criterion; (2) FedIRM [12] using a new inter-client matching relationship to obtain similar information about labeled client diseases to constrain unlabeled clients and reduce task loss; and (3) imFed-Semi [18] used to build a dynamic class library, filter out high confidence samples, re-estimate the client class distribution, subdivide it into sub-banks with different pseudo-labels, and construct a priori transformation functions through Bayesian criteria to transform the task into classification of the sub-banks, thus avoiding the impact caused by the majority class of the central client.

Due to the limitations of the experimental equipment, the upper limit of our method batches was 10, which is also the optimal batch that our method can achieve. We performed two sets of control experiments for the state-of-the-art method separately: One set of models effects under 10 batches of data training (fixed variables, let the state-of-the-art method be compared with our method under the same batches). The other set shows the effect of the model trained with the optimal batch (greater than 10) (fixed upper limit to allow the state-of-the-art method to be compared with our method at the optimal effect that can be achieved). A comparison of the results of our method and the state-of-the-art method is shown in Table 1.

Table 1. Comparison with state-of-the-art methods on the same task.

Method	Client Num		Metrics					
	Label	Unlabel	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
FedAvg [22]	2	0	88.27	92.34	61.43	92.28	66.50	62.53
FedIRM(10) [12]	2	8	88.17	89.87	40.76	91.66	34.38	37.02
FedIRM(best) [12]	2	8	90.38	90.30	67.86	92.87	61.20	62.02
imFed-Semi(10) [18]	2	8	92.40	93.30	58.29	92.10	76.87	63.10
imFed-Semi(best) [18]	2	8	88.40	94.75	67.75	94.04	79.12	71.77
our	2	8	95.75	95.58	72.92	95.47	73.88	72.90

Among them, the FedIRM [12] and imFed-Semi [18] methods produce a degradation in model performance due to the reliance on inter-data relationship-driven representation and the need for sub-base-alike redistribution of data in small batches of data. This leads to insufficient information on class relationships during a single training session, resulting in the loss of a few class decision boundaries in the aggregated and broadcast models. When increasing the batch size, there is a significant improvement in model performance beyond the baseline level as more information on class relationships can be obtained from the labeled data and there is a greater probability of reassigning a few classes, indicating that the additional unlabeled information is beneficial for federated learning through the integration algorithm.

Our combination of regularization constraints and pseudo-label construction solves these problems well. In contrast to FedIRM [12], which can only find limited relational matches from each batch of labeled image data, our imbalance factors extracted from the labeled data are global and unique, and thus are not affected by the size of the data training batches. Relying only on regularized relational matches between clients in FedIRM [12] greatly reduces the learning ability of the model for fewer classes in the face of small batches of imbalanced data, while our regularized class balancing aid, constructed using imbalance factors, significantly improves the learning ability of the model.

Compared with the imFed-Semi [18] method, the method attempts to solve the class balance problem by constructing sub-banks for the contradictory transfer of tasks, which transforms the classification problem for unbalanced classes into a classification problem for sub-banks after reconstructing the data. However, when constructing the transfer function, due to the randomness of the data allocation ratio, in order to ensure full data coverage, it often appears that the first few libraries have less data allocated and the last library has too much data allocated. This allocation is unlearnable, which also means that this phenomenon will continue throughout the training, and when the batch is small, it may even appear that some of the libraries are empty, which makes it easy to have model degradation during the training. The pseudo-label construction we use solves these problems by first using a data selector to filter out stable samples, second by averaging the 128 batches of unlabeled predictions to alleviate the problem of limited learning range due to small batches of data, generating dynamic edge distributions in a cyclic manner (replacing the 1st prediction when the 129th prediction occurs). Additionally, it incorporates a constant edge distribution imbalance factor to jointly constrain the class probability distribution generated by the network, and finally generates pseudo-labels. This ensures the smoothness of the model in learning unlabeled data and improves the category imbalance problem.

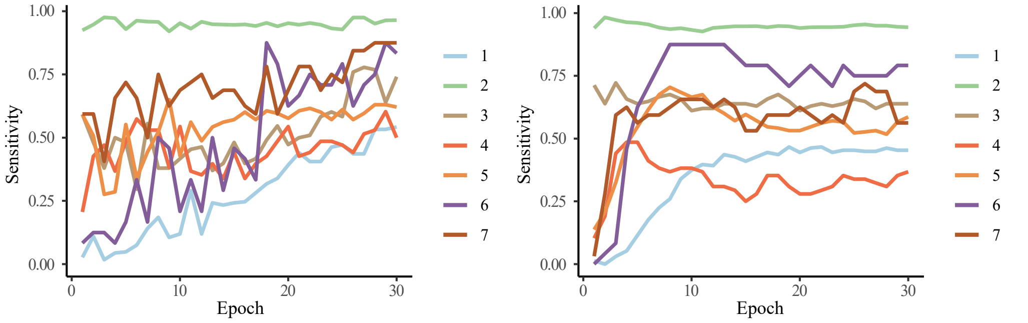
Our method shows significant improvement in most metrics (especially AUC, Sensitivity, and specificity) compared to FedIRM [12] and imFed-Semi [18] on the skin lesion classification task for the same batch of dermoscopic images, thanks to the combination of class distribution regularization with pseudo-label representation. In general, category distribution regularization solves the problem that the aggregated model is more likely to favor the majority class in having labels, as shown in the FedIRM(10) results in Table 2 of Section 3.3, which tends to cause the decision boundary of the minority class to vanish. In addition, the application of our pseudo-labeled representation solves the problem of perturbation of the training model by the adversarial samples, as shown in Figure 5 of Section 3.3. Our method makes full use of the unlabeled data, so that the robustness of the model is significantly improved, and the cyclic dynamically generated distribution also avoids model degradation.

Table 2. Comparison of the sensitivity of each category under different methods.

Method	Category							
	1	2	3	4	5	6	7	Average
FedAvg	44.40	92.99	69.63	42.05	57.83	60.83	70.00	62.53
FedIRM(10)	61.23	88.73	0	0	60.59	0	0	30.07
FedIRM(best)	58.15	80.76	87.96	45.59	66.01	33.33	11.75	54.79
imFed-Semi(10)	29.07	99.40	74.07	50.00	39.40	16.66	78.12	55.24
imFed-Semi(best)	52.86	99.11	77.78	57.35	49.75	87.50	78.13	71.78
our	53.30	96.34	63.88	60.29	63.05	87.50	87.50	73.12

3.3. Internal Comparative Analysis of Methods

Dynamic learning within categories: In Figure 6a, the sensitivity change curves of each category under our method, and in Figure 6b the sensitivity change curves of each category under FedAvg [22] model are shown and set as the baseline level for comparison with our method. The horizontal axis shows 30 class sampling points sampled in 10 cycles, and the vertical axis indicates the change in sensitivity of each class in the test set as the iteration cycle progresses. From the observation we can infer that with the convergence of training almost all sensitivities of all classes are higher than the baseline level. Our method can learn knowledge effectively from unlabeled data and improve the network performance, to not only have higher recognition for multi-category data, but also have a greater improvement in the recognition rate for fewer category data. Figure 7 reflects the changes of each loss for the training set and the test set. It can be observed that at the beginning of the 170th cycle, each loss curve shows an obvious decreasing trend, among which the change of auxiliary loss is the most obvious and runs through the other curves. This is, because with the increase in training cycles, the class balance aid gradually produces a constraining effect on the pseudo-label of unlabeled data, and each loss curve has the same change trend among them. This indicates that there is no overfitting phenomenon in the training of the model. This result also demonstrates the effectiveness of our approach of combining consistency constraints with pseudo-labeling in the federated learning for the class imbalance task.



(a) (b)

Figure 6. (**a**) A graph of the sensitivity changes of each category under our method. (**b**) A graph of the sensitivity changes of each category under the basic network.

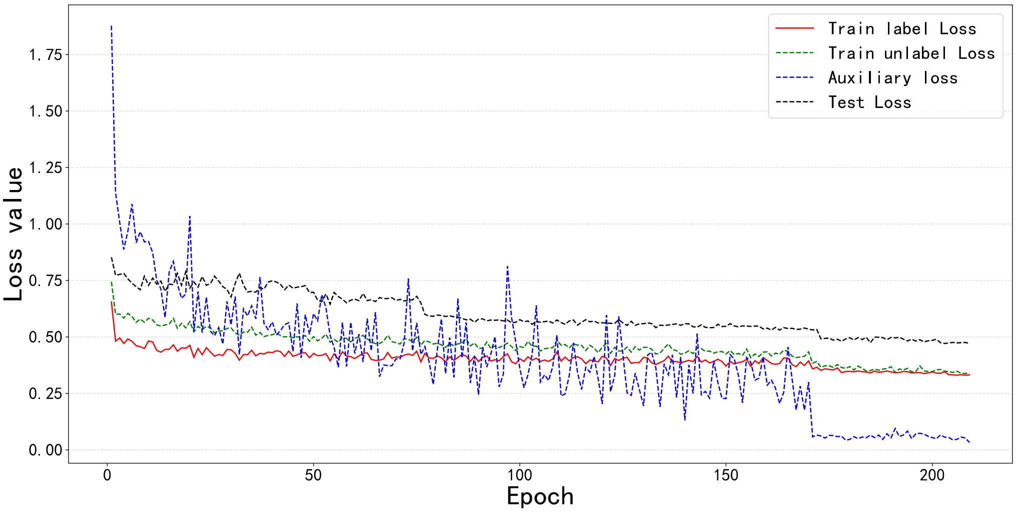


Figure 7. Variation of loss curve between training set and test set.

Performance Analysis of Fusion Methods: As shown in Table 2, for the comparison of the category analysis performed using the FedAvg [22], FedIRM [12], imFed-Semi [18] method, and our method under different batches, first we can observe that FedAvg [22] can guarantee a higher accuracy rate under the same small batch of training with fewer category samples, which indicates that the category is more distinguishable compared to other categories and reduces the impact caused by fewer data samples. FedIRM [12] shows severe model degradation, with smaller batches of training affecting the supervision of inter-client relationships, as can be observed in comparison to the control in Figure 1b, where, although most categories show a significant improvement in accuracy, a smaller number of four categories show the disappearance of decision boundaries. imFed-Semi [18] shows a significant improvement in the average category sensitivity compared to FedIRM [12] and a significant improvement in the impact caused by data class imbalance. However, due to the uneven distribution of sub-bases during data redistribution and the fact that this phenomenon is constant, which leads to a new imbalance problem in small batches of training, we can find a significant decrease in the accuracy of the first category compared to the other methods. In contrast, the regularization constraint in our method is global and unique, unaffected by small batches of data, and the pseudo-label construction is smooth. Combining the two methods ensures the stability of the model while solving the class imbalance problem, and it can be observed that our class sensitivity is generally higher than the baseline level, and the average class sensitivity shows a large improvement compared to the advanced methods tested. Secondly, under the optimal batch training, our method shows an average category sensitivity improvement of 1.34% compared to the imFed-Semi [18] method with class balancing effect. As shown in Figure 8, under our model evaluation criteria, it can be observed that our method produces a lower error, which indicates that the model trained by our method has higher robustness.

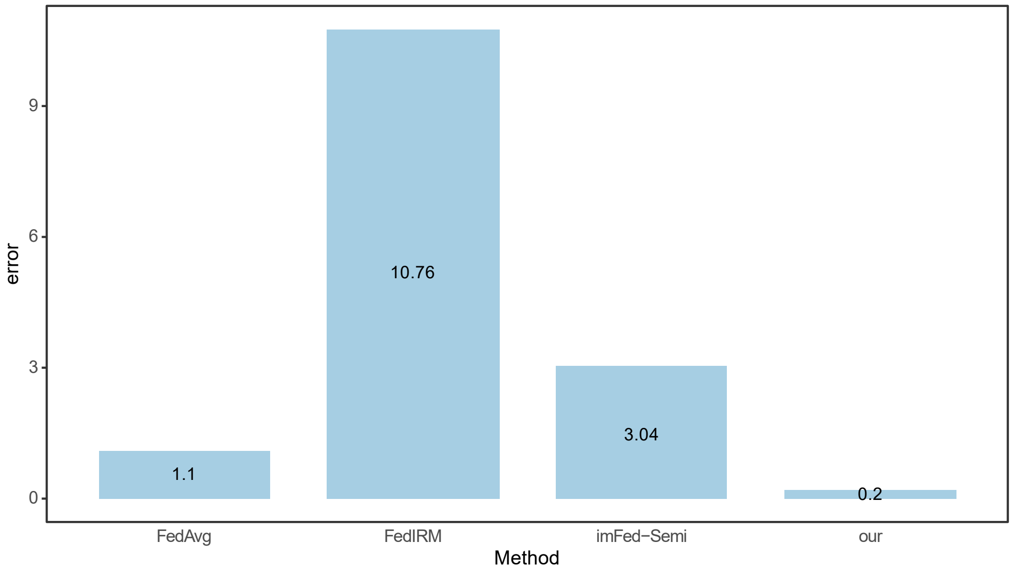


Figure 8. Sensitivity errors of different methods under different evaluation criteria.

Exploratory study of internal parameters: We fixed the amount of tagged data for the central client at 20% and kept increasing the number of local clients to change the ratio of tagged and untagged data. By observing Figure 9 we can conclude that the accuracy is highest when the local clients are four and the ratio of tagged to untagged data for a single local client is 1:1. It is worth mentioning that when the local client is 1, the training method degenerates to ordinary semi-supervised learning, which also proves the effectiveness of using federated learning. In addition, we continued to do a dynamic analysis of the confidence threshold of the key parameter in the method, because when the threshold is too low, the training is disturbed and produces errors, while a high threshold causes the regularization to lose its effect. As shown in Table 3, we detected that the model accuracy reaches the highest values when the threshold is set to 0.95.

Table 3. Sensitivity of our method at different confidence thresholds.

η	1	0.98	0.95	0.9	0.85	0.8	0.75	0.7
Sensitivity	71.09	71.63	72.92	71.53	72.16	71.5	69.89	69.94

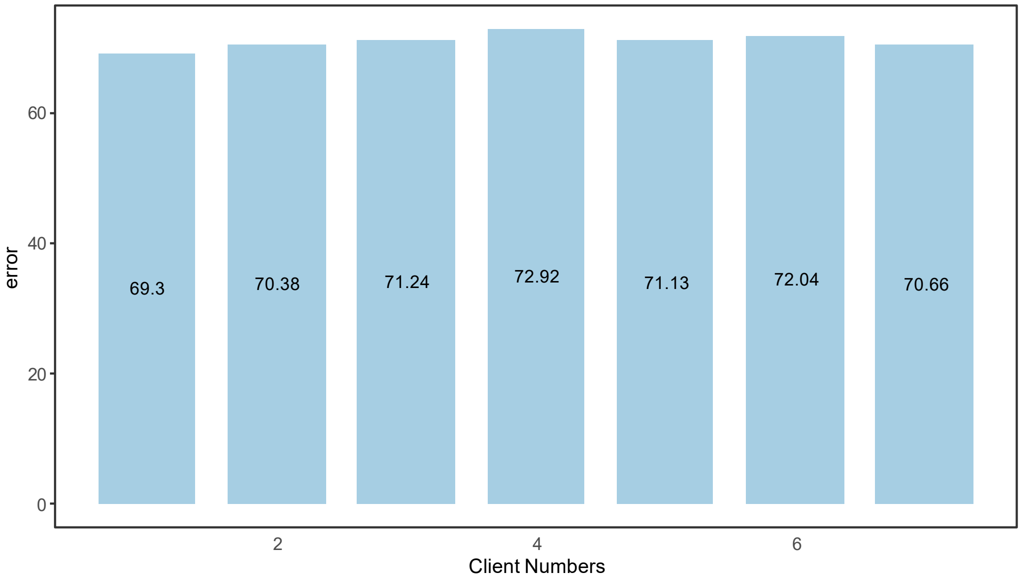


Figure 9. The sensitivity variation of our method with the same amount of data and the number of clients in different places.

4. Discussion

We present a new federated learning method that can effectively solve the class imbalance problem in the melanoma dermoscopy image classification task. This is the first time that regularization constraints are combined with pseudo-labeled representations in semi-supervised federated learning, and this classification model demonstrates the improvement in the accuracy of the model using our approach by two validation methods. The classification model trained under our method has an accuracy of 95.75%, sensitivity of 72.92%, specificity of 95.47%, precision of 73.88%, F1 score of 72.90%, and AUC of 95.75%.

In the process of proving the validity of our proposed method, we compared our method with the baseline level FedAvg [22] and two of the current state-of-the-art methods FedIRM [12] and imFed-Semi [18] in the field of federated learning. With the same data set, the same data distribution, the same batch, and the same model optimal performance, our method leads to the optimal level of the model. The sensitivity (recall), the most important metric in medical image classification, is improved by 5.17% compared to the state-of-the-art level, and the specificity, another important metric, is improved by 1.43%.

In addition, we present a simple model-robustness evaluation criterion, which demonstrates the effectiveness of our proposed method in improving the robustness of the model by showing that our model yields the smallest prediction distribution error compared to the models trained by FedAvg [22], FedIRM [12] and imFed-Semi [18]. When we change the sample ratio between the medium and local clients by a controlled data sampler, as shown in Figure 9, we find that the sensitivity of important metrics is improved by 3.62% under the federated learning framework compared to the degraded ordinary semi-supervised method. This demonstrates the effectiveness of combining federated learning with semi-supervised learning for the improvement of classification performance of medical image classification models.

Notably, by looking at Table 1, it can be observed that the accuracy index decreases by 5.24% compared to the state-of-the-art level, while Equation (14) reflects a significant decrease in accuracy when false positives (FP) rise. Through analysis, we believe that the reason is that the pseudo-labeling is constrained while relaxing the prediction of the few-category data labels, thus improving the sensitivity of the model to the few categories to some extent. In contrast to the decrease in sensitivity for most categories, described in Table 2, according to Equation (15) we observe a significant decrease in sensitivity when false negatives (FN) rise, due to the constraint of pseudo-labeling while suppressing the prediction of most category data labels. In short, this phenomenon occurs uniformly, i.e., all images from the majority category are predicted into the minority category, but we believe that this phenomenon can be tolerated, because from a theoretical point of view, the sensitivity of the minority category is substantially improved, which nicely solves the category imbalance problem. From a practical point of view, we believe that a larger number of image categories corresponds to diseases with higher probability of occurring. This means, there are relatively more experience and methods to deal with that category of diseases in clinical treatment, and doctors have higher recognition of that category of diseases. This is a deficiency that can be largely compensated for manually, while a few categories of diseases lack the above elements, so we believe that the classification model trained by our method has more advantages than disadvantages.

In addition to these shortcomings of our approach, in terms of model initialization only simple supervised training is used, and we believe that such a migration of the parameter distribution is not suitable for semi-supervised learning. In terms of task selection, although our method has been validated effectively in single-category multiclassification tasks, it is still inadequate in the face of multi-category multiclassification tasks and needs further exploration and research. Similarly, our proposed model robustness validation method is only applicable to single-category multiclassification task models. In terms of model generalizability, our method contends for better classification of dermoscopic images, but whether the method is generalizable in the field of medical images requires further research, e.g., Amyar et al. [31] proposed a multi-task learning framework to predict diseases, and they argued that encoders can use multiple tasks to extract meaningful and powerful features, allowing better generalization of the model. We believe that such an approach is highly informative in the field of medical image classification. In terms of practical application, at present, our research is still theoretical basic research. Theoretically, we believe that the method is clinically applicable, but at present it is not applied in clinical applied research. In a subsequent work we will attempt to address the shortcomings presented above, and in a recent work we preliminarily constructed a model initialization method applicable to semi-supervised learning and extended the approach of this paper in combination with it to other medical imaging tasks to demonstrate the generality of the method.

5. Conclusions

In this study we present an improved semi-supervised federated learning problem in solving federated learning for medical information classification tasks, and we propose the first method that combines regularization constraints with pseudo-label construction. To address the category imbalance at the data level, our specific approaches include (1) using valid information in tagged data to constrain pseudo-labels in local clients and (2) proposing a data selector to make it filter interference-resistant when using untagged images as new data sets to participate in pseudo-label generation. In addition, we propose a method to measure the robustness of the model in the experimental validation phase. The validity of our proposed method is verified in an experimental validation on a publicly available data set. In the future, we will strive to investigate the direction of model parameter initialization in semi-supervised learning in order to explore a model parameter initialization method applicable to universal medical image classification tasks.

Author Contributions: Conceptualization, methodology, software, visualization, validation, formal analysis, investigation, data curation, writing—original draft preparation, W.L.; writing—review and editing, supervision, J.M.; resources, J.M. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported the Natural Science Foundation of Xinjiang Uygur Autonomous Region, under grant number No.2019D01C072, Education Department of Guizhou Province, under grant number No. QJHKY[2019]115.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Since the data set comes from the Skin Lesion Analysis Cha-llenge for Melanoma Detection at the 2017 International Symposium on Biomedical Imaging (ISBI) hosted by the International Skin Imaging Collaboration (ISIC) (https://www.kaggle.com/datasets/wanderdust/skin-lesion-analysis-toward-melanoma-detection (15 September 2022)), informed consent was not applicable in our case.

Data Availability Statement: Challenges of Melanoma Detection in Skin Lesion Analysis at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC) can be obtained from 2018 IEEE 15th International Sy-mposium on Biomedical Imaging (ISBI2018) (https://www.kaggle.com/datasets/wanderdust/skin-lesion-analysis-toward-melanoma-detection (15 September 2022)).

Acknowledgments: Thanks to Mo Jiaqing for his valuable advice and professional guidance, and to the Natural Science Foundation of Xinjiang Uygur Autonomous Region and the Education Department of Guizhou Province for their financial support of our research work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dhruva, S.S.; Ross, J.S.; Akar, J.G.; Caldwell, B.; Childers, K.; Chow, W.; Ciaccio, L.; Coplan, P.; Dong, J.; Dykhoff, H.J.; et al. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *npj Digit. Med.* **2020**, *3*, 60. [CrossRef] [PubMed]
2. Silva, S.; Gutman, B.A.; Romero, E.; Thompson, P.M.; Altmann, A.; Lorenzi, M. Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019.
3. Razzak, M.I.; Naz, S.; Zaib, A. Deep Learning for Medical Image Processing: Overview, Challenges and Future. In *Classification in BioApps: Automation of Decision Making*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
4. Dong, N.; Voiculescu, I. *Federated Contrastive Learning for Decentralized Unlabeled Medical Images*; Springer: Berlin/Heidelberg, Germany, 2021; p. 12903. [CrossRef]
5. Dou, Q.; So, T.Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H.H.C.; Yu, K.; et al. Author Correction: Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study. *npj Digit. Med.* **2022**, *5*, 56. [CrossRef] [PubMed]
6. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. *arXiv* **2021**, arXiv:2102.07623. [CrossRef]
7. Roth, H.R.; Chang, K.; Singh, P.; Neumark, N.; Li, W.; Gupta, V.; Gupta, S.; Qu, L.; Ihsani, A.; Bizzo, B.C.; et al. *Federated Learning for Breast Density Classification: A Real-World Implementation*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12444, pp. 181–191.
8. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R.; et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 12598. [CrossRef] [PubMed]
9. Wu, Y.; Zeng, D.; Wang, Z.; Shi, Y.; Hu, J. *Federated Contrastive Learning for Volumetric Medical Image Segmentation*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12903, pp. 367–377.
10. Chang, Q.; Qu, H.; Zhang, Y.; Sabuncu, M.; Chen, C.; Zhang, T.; Metaxas, D.N. Synthetic Learning: Learn From Distributed Asynchronized Discriminator GAN Without Sharing Medical Image Data. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
11. Li, D.; Kar, A.; Ravikumar, N.; Frangi, A.F.; Fidler, S. *Federated Simulation for Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 12261, pp. 159–168.
12. Liu, Q.; Yang, H.; Dou, Q.; Heng, P.-A. *Federated Semi-Supervised Medical Image Classification via Inter-Client Relation Matching*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12903, pp. 325–335.
13. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2022**, *3*, 119. [CrossRef] [PubMed]
14. Gyawali, P.K.; Ghimire, S.; Bajracharya, P.; Li, Z.; Wang, L. *Semi-Supervised Medical Image Classification with Global Latent Mixing*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12261, pp. 604–613.
15. Shi, X.; Su, H.; Xing, F.; Liang, Y.; Qu, G.; Yang, L. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Med. Image Anal.* **2020**, *60*, 101624. [CrossRef] [PubMed]
16. Liu, Q.; Yu, L.; Luo, L.; Dou, Q.; Heng, P.A. Semi-supervised Medical Image Classification with Relation-driven Self-ensembling Model. *IEEE Trans. Med. Imaging* **2020**, *39*, 3429–3440. [CrossRef] [PubMed]
17. Amyar, A.; Modzelewski, R.; Vera, P.; Morard, V.; Ruan, S. Weakly Supervised Tumor Detection in PET Using Class Response for Treatment Outcome Prediction. *J. Imaging* **2020**, *8*, 130. [CrossRef]

18. Jiang, M.; Yang, H.; Li, X.; Liu, Q.; Heng, P.-A.; Dou, Q. *Dynamic Bank Learning for Semi-Supervised Federated Image Diagnosis with Class Imbalance*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13433, pp. 196–206.
19. Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P.M.; Rueckert, D. *Semi-Supervised Learning for Network-Based Cardiac MR Image Segmentation*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10434, pp. 253–260.
20. Lee, H.; Shin, S.; Kim, H. ABC: Auxiliary Balanced Classifier for Class-imbalanced Semi-supervised Learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 7082–7094.
21. Bdair, T.; Navab, N.; Albarqouni, S. Semi-Supervised Federated Pe-er Learning for Skin Lesion Classification. *arXiv* **2021**, arXiv:2103.03703. Available online: https://arxiv.org/pdf/2103.03703.pdf (accessed on 17 December 2022).
22. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. *arXiv* **2016**, arXiv:1602.05629.
23. Oerton, E.; Roberts, I.; Lewis, P.S.; Guilliams, T.; Bender, A. Understanding and predicting disease relationships through similarity fusion. *Bioinformatics* **2018**, *35*, 1213–1220. [CrossRef]
24. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *arXiv* **2019**, arXiv:1911.09785. [CrossRef]
25. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, C.K.; Marchetti, M.; et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A ChallengeHosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368. [CrossRef]
26. Cao, X.; Chen, B.C.; Lim, S.N. Unsupervised Deep Metric Learning via Auxiliary Rotation Loss. *arXiv* **2019**, arXiv:1911.07072. [CrossRef]
27. Zhai, X.; Oliver, A.; Kolesnikov, A.; Beyer, L. S4L: Self-Supervised Semi-Supervised Learning. *arXiv* **2019**, arXiv:1905.03670. [CrossRef]
28. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef]
29. Yuan, Z.; Yan, Y.; Sonka, M.; Yang, T. Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 17 October 2021.
30. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. RandAugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
31. Amyar, A.; Modzelewski, R.; Vera, P.; Morard, V.; Ruan, S. Multi-Task Multi-Scale Learning For Outcome Prediction in 3D PET Images. *Comput. Biol. Med.* **2022**, *151*, 106208. Available online: https://arxiv.org/pdf/2203.00641.pdf (accessed on 17 December 2022). [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.