A comprehensive study of Semi-Supervised Learning in Medical Imaging

Project report

Anna Banaszak, Cenk Eralp, Mert Sayar

Supervision: Tariq Bdair

Chair of Computer Aided Medical Procedures
Technical University of Munich

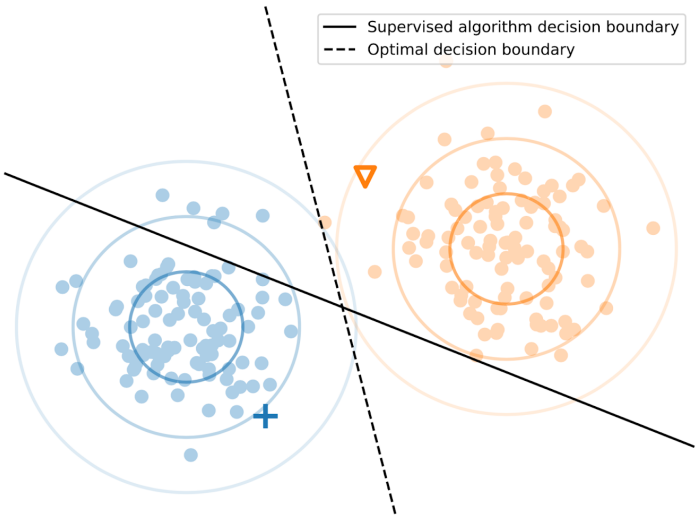
Introduction

The project focuses on the evaluation of several Semi-Supervised Deep Learning techniques (SSL) in realistic settings. Used SSL models are designed and trained for a classification task on medical data. The project might be a foundation of an evaluation framework targeting SSL methods used in the medical domain. The project consisted of two parts:

- implementation and evaluation of SSL methods on medical data classification problems,
- implementation of the models in a basic federated setting.

Motivation

Fully Supervised Deep Learning Neural networks require large amounts of labelled data which might be hard to obtain in the medical domain. Labelling requires a significant time investment of medical specialists, which is very expensive and sometimes simply not available. To overcome this limitation we explored the feasibility of Semi-Supervised Deep Learning methods to leverage information from unlabelled data. Adding unlabelled data to the training process aims to optimize the decision boundary based on the intrinsic class distribution to move it to the region with low sample density.



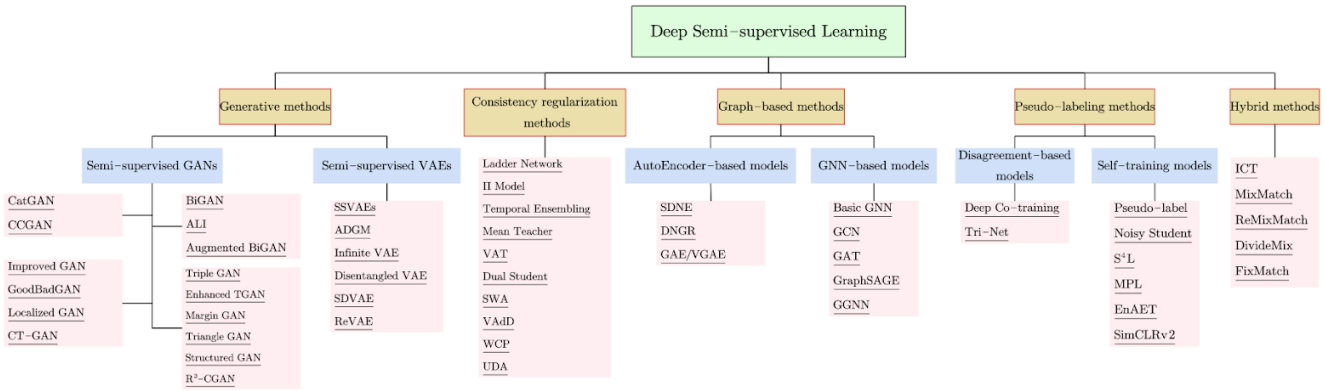
Visualization of a binary classification SSL method [1]

According to Oliver et al [2], SSL methods often do not follow unified benchmarking methods and might be evaluated in optimistic settings in favour of SSL algorithms. We propose several experiments based on real-life scenarios.

Background & Related Work

Semi-Supervised Learning

Deep semi-supervised methods are divided into five categories: Generative methods, Consistency regularization methods, Graph-based methods, Pseudo labelling methods and Hybrid methods[1]. The latter can be a combination of the first four categories.



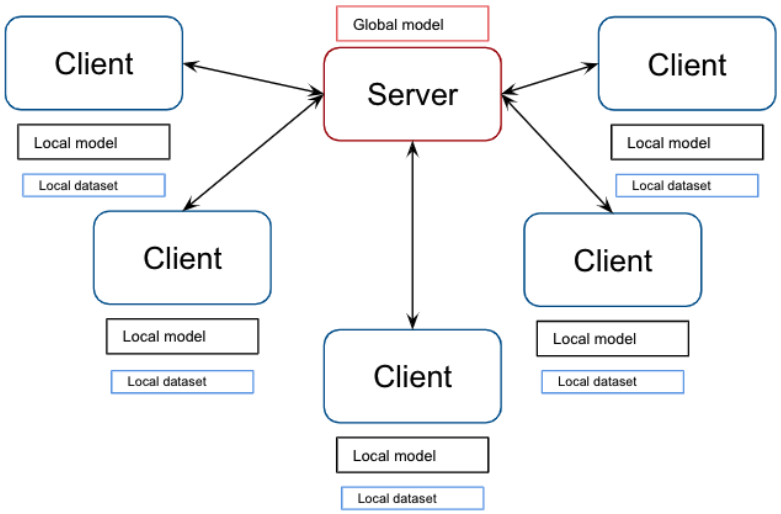
Taxonomy graph of SSL methods[1]

We proceeded with three different models chosen from three different categories. The first model is the basic Semi-supervised GAN[7] as part of the generative model category. Our second model is the Mean Teacher[8] from the consistency regularization category, and the third model is the Comatch[9] from the hybrid category. It is a combination of graph-based models and pseudo labelling methods. Comatch can be considered the State of the Art algorithm in Semi-Supervised Learning studies.

Federated Learning

Federated Learning (FL) is a method to train Deep Learning models on distributed data[3]. It is based on Client-Server architecture with one Global model stored on the Server and multiple Local models with corresponding Local datasets in Clients. The Global model and Local models have the same architecture.

Architecture diagram of the FL method



Training is performed in rounds. Each round corresponds to one Global model update. In the beginning, Clients participating in the round of training are sent the weights of the Global model. Weights are loaded to Local models. Then, Local models are trained for a defined number of epochs with the use of the Local datasets. After completion, obtained weights of Local models are sent back to the server where they are combined according to a chosen algorithm, e.g. FedAvg [3] to result in new weights of the Global model. This operation finalizes the round.
In each round, only a fraction of all clients is participating in the training process to simulate network connection disruptions. Participating clients are chosen randomly.
Final evaluation is performed on the Global model.
Health records and patient data are subject to strict data protection regulations. Therefore, Federating Learning is a promising method of training Deep Learning models in the medical domain, because it does not require moving the data to one centralized server. Patient data can be stored securely in the institution of its origin. Incorporating additional mechanisms like Differential Privacy or Homomorphic encryption allows for training the model in a privacy-preserving manner while leveraging data from multiple sources.

Evaluation framework

We based our evaluation methodology on a paper by Oliver et. al[2]. The study focuses on the realistic evaluation and a comparison of the SSL methods. The authors suggest that many SSL methods are evaluated under optimistic settings which do not reflect reality. It might result in comparing underperforming fully supervised models against SSL models trained favourably. To prevent this from happening authors introduce six principles of proper SSL benchmarking. The first principle is shared implementation, which is crucial for a fair comparison. The baseline model and the used hyperparameters should be the same. The second principle is the usage of a high-quality supervised baseline to obtain meaningful results. The third principle explains the importance of comparing the performance of the SSL model with Transfer Learning on the baseline model. Transfer Learning might result in better predictions on the fully-supervised model than the SSL, but this case is often overlooked. The fourth principle is to consider a class distribution mismatch between labelled and unlabelled data. This principle focuses on a real-life problem when samples corresponding to some of the classes are underrepresented in the unlabelled dataset. The fifth principle is to experiment with varying amounts of labelled and unlabelled data. With these experiments, two realistic scenarios can be simulated: a large unlabelled dataset and a relatively small unlabelled dataset. The last principle is to use a realistic size of the validation set. The idea is that, since labelled data is a small part of the training data, the original validation set can often be larger than the labelled data, which is against the nature of train, validation and test set ratios.

Methodology

SSL methods evaluation

Following principles proposed by Oliver et al[1] we conducted the following experiments:

1. Baseline model implementation
2. Transfer Learning on the baseline model
3. Varying amounts of labelled data
4. Varying amounts of unlabelled data
5. Class distribution mismatch in labelled and unlabelled data
6. Small validation set
7. Federated Learning on varying labelled data amounts

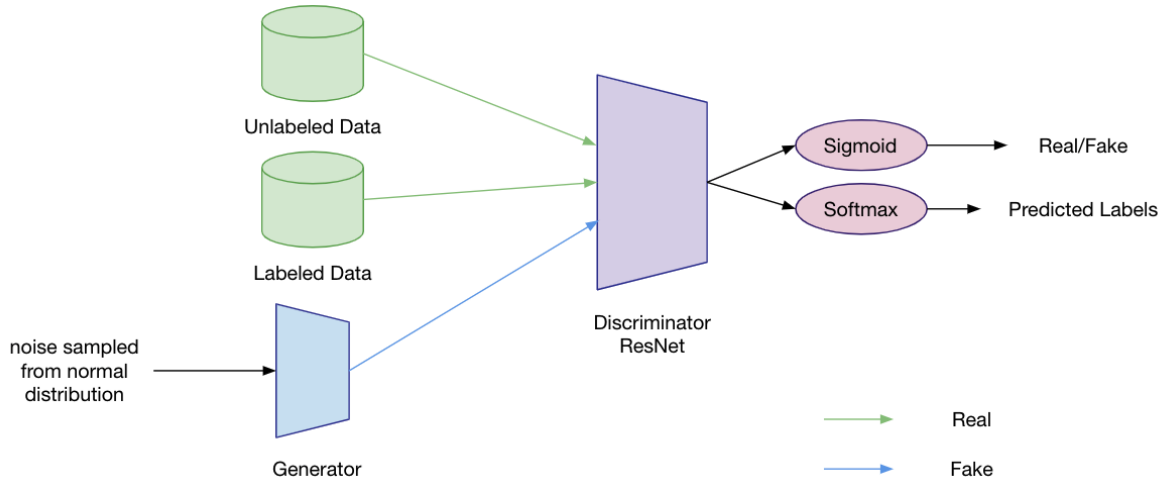
The last experiment is an extension to the evaluation framework proposed by Oliver et al. Federated Learning is becoming more and more relevant in training DL models in the medical domain. Evaluating SSL models trained on distributed data seems to be one more important step in measuring the realistic performance of SSL models.

Dataset

The PathMNIST[3] dataset used in the project is a 9-Class Colon Pathology data with 107,180 samples: 89,996 for training, 10,004 for validation and 7,180 for testing. It consists of 3-channel RGB images of size 28x28. In the implementation, we used the MedMNIST[4] library, which provides a convenient interface to access the data. For purposes of SSL, we created two datasets: labelled and unlabelled data. The labelled data is a class-balanced random sample from PathMNIST training data. In the case of unlabelled samples, the label has been discarded.

Semi-supervised models

S-GAN

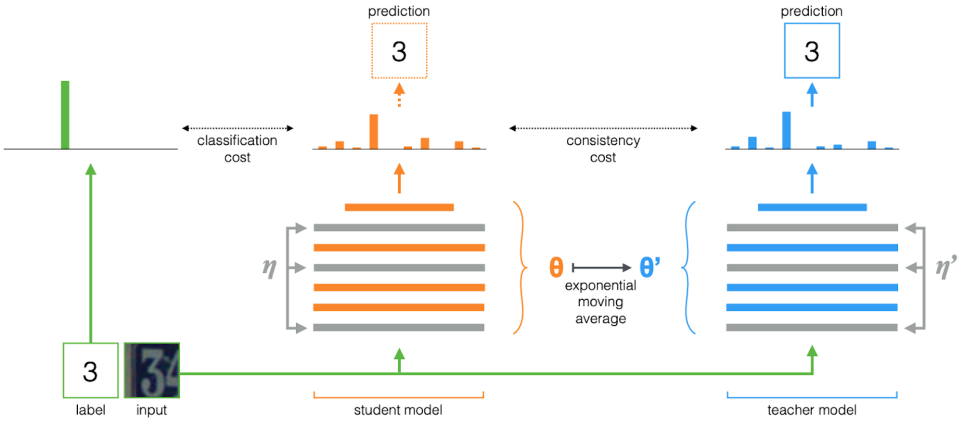


Architecture diagram of the S-GAN model

Semi-Supervised GAN (S-GAN)[7] is the most basic SSL method we used in our project. It is based on the GAN model but also contains the labelled input. The Discriminator, besides differentiating between fake and real samples, also learns the classification. The Discriminator is the baseline ResNet18 for purposes of fair comparison between methods. The Discriminator is used later for inference in the validation and testing phases. The Generator consists of two convolution layers with upsampling, batch normalization and leaky ReLU activation function and one convolution layer with Tanh activation function as the output of the model.

Training of the Discriminator consists of two parts. One is learning the classification where we use the output with softmax activation function and cross-entropy loss. The second part is learning the discrimination between fake and real samples, where fake come from the Generator and real from the unlabelled part of the dataset. We use here the output of the sigmoid function and binary cross entropy loss. It allows us to incorporate the intrinsic structure of the training data into the Discriminator.
The last part is the Generator training. Based on the sigmoid output and binary cross entropy loss of the Discriminator we try to make the generated samples better to fool the Discriminator.

Mean Teacher



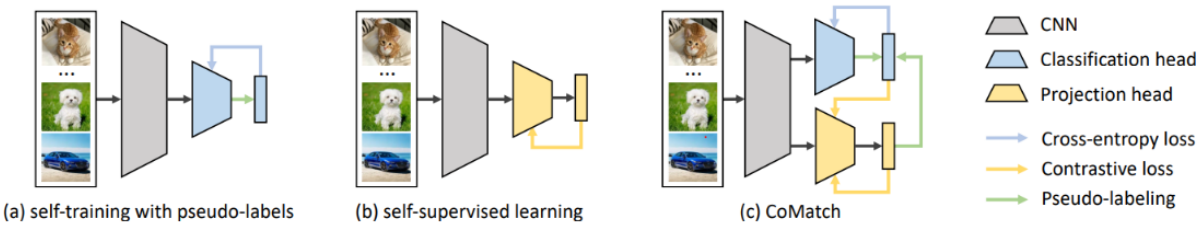
Architecture diagram of Mean Teacher method[8]

Mean Teacher [8] consists of two identical models: Student and Teacher. Student is the main model and is used for inference in validation and test. Student and Teacher models are fed slightly different inputs varying in added noise level during data augmentation. Teacher's weights are a moving average of the Student's weights. Loss function, used to train the Student model, is a weighted sum of cross entropy loss and consistency loss $J(\theta)$ between predictions of student and teacher. It is based on the assumption that similar input instances should lie close to each other on the output space. We take an unlabelled sample, apply a small random gaussian noise differently and feed these slightly different inputs into the student and teacher.

$$J(\theta) = \mathbb{E}_{x,\eta',\eta}\left[\|f(x,\theta',\eta') - f(x,\theta,\eta)\|^2\right]$$

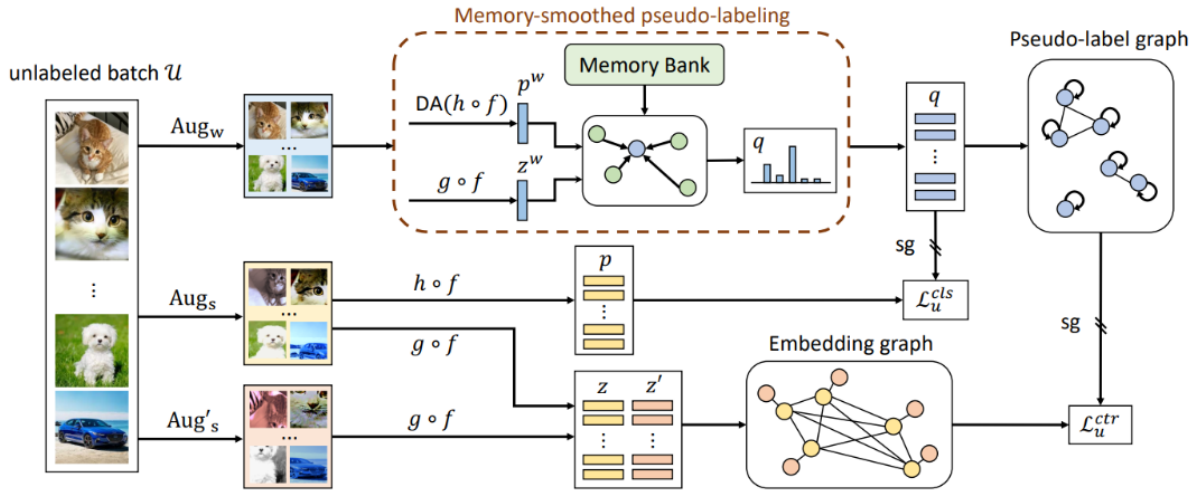
Comatch

Comatch is the SOTA Semi-Supervised learning algorithm. It combines graph-based and pseudo labelling methods. In Comatch, each image has two compact representations. One of the representations is the classification head, and the other is the low-dimensional embedding produced by the projection head. These representations interact with each other and jointly evolve in a co-training framework.



Methods that leverage unlabelled data[9]

As a detailed explanation, batches of unlabelled images are used as their weakly augmented images are used to produce the memory-smoothed labels which are later used as targets to train the class prediction on strongly augmented images. Also, in order to measure the similarity between samples, a pseudo-label graph is created. This pseudo-label graph will train the embedding graph in the way that images with similar pseudo-labels have similar embeddings.



Architecture diagram of Comatch method[9]

As part of the memory smoothed labeling, following objective is optimized:

$$J(q_b) = (1-\alpha)\sum_{k=1}^{K} a_k \left\|q_b - p_k^w\right\|_2^2 + \alpha \left\|q_b - p_b^w\right\|_2^2$$

Where the first term is the smoothness constraint and the second term attempts to maintain its original class prediction.

Results

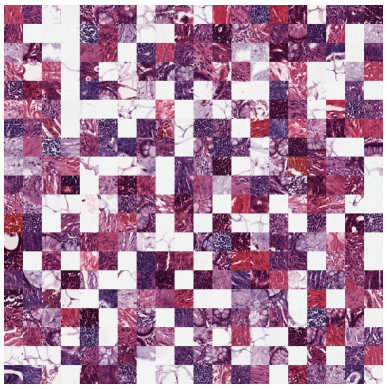
In the following experiments we compare SSL methods with a fully supervised baseline ResNet18 model. We used common hyperparameters of the models and accuracy(%)↑ as the evaluation metric. Accuracy results are presented in the tables.

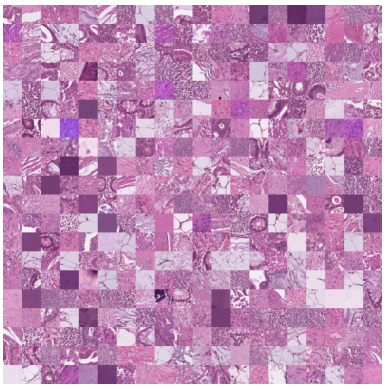
Experiment 0 - Baseline model implementation

To ensure comparability between results we trained a baseline model - ResNet18 with 5 residual blocks, which was later used for tuning common hyperparameters. The ResNet model was incorporated into SSL models which allows for a fair comparison between all methods. Chosen hyperparameters are learning rate 0.001, batch size 64, patience 30 epochs, max number of epochs 100, learning rate scheduler after 10 and 50 steps with gamma 0.1.

Experiment 1 - Transfer Learning on the baseline model

SSL models require more computational resources and longer training time compared to ResNet. In terms of the explainability of the results and overall interpretability of the models, SSL methods are far more complicated. We performed a transfer learning experiment to investigate whether it might be a feasible solution to reduce computational time and model complexity. We choose another multi-class colorectal cancer histology dataset for pertaining[5]. It is an 8-class dataset with 4000 training samples and 1000 samples used for validation. Later we fine-tuned the model on 1% of the PathMnist dataset.

a) 

b) 

a) 100 exemplary samples from dataset used for pre-training[5]
b) 100 exemplary samples from PathMNIST dataset[6]

SSL models were trained on 1% of Training Data as labelled data and the remainder as unlabelled data.

ResNet	ResNet with pre-training	Comatch	S-GAN	Mean Teacher
61.59	68.44	82.09	64.15	74.61

Pretrained fully-supervised model ResNet performed better than the S-GAN but worse compared to Mean Teacher and especially Comatch. We conclude that Transfer Learning might be a good option in cases where using complex models is not possible due to computational or model complexity constraints.

Experiment 2 - Varying amounts of labelled data

In the second experiment, we investigate the influence of the varying ratio of labelled training samples on accuracy of the models. Three ratios of labelled data are experimented: 1%, 5% and 10% of the training data.

Labelled data	ResNet	Comatch	S-GAN	Mean Teacher
1% of training data	61.59	82.09	64.15	77.95
5% of training data	78.55	85.79	79.26	83.41
10% of training data	82.45	89.86	84.02	85.17

As expected, all semi-supervised models exceed the fully-supervised ResNet18 model with the corresponding labelled data ratios. As the labelled data amount increased, the performance of every model also increased.

Experiment 3 - Varying amounts of unlabelled data

In this experiment, we fixed the amount of labelled data to 10% of the training set and changed the amount of unlabelled data used to train SSL models. The accuracy of the baseline ResNet model trained on 10% of training data is 82.45%.

Unlabelled data	Comatch	S-GAN	Mean Teacher
25% of training data	84.43	79.42	82.70
50% of training data	86.27	83.00	82.99
100% of training data	89.86	84.02	85.17

The only model with lower accuracy than the baseline is S-GAN trained on 25% of training data as the unlabelled data. Important conclusion of this experiment is that using an SSL model with not enough unlabelled data can hurt the performance. Case of varying amounts of unlabelled data should be evaluated to objectively assess the performance of the SSL model.

Experiment 4 - Class distribution mismatch

In the fourth experiment, we considered the class distribution mismatch between labelled and unlabelled data. In previous experiments, we used the remainder of the training set with discarded labels as the unlabelled data. However, in real life, some of the classes might be underrepresented in the unlabelled data. It is impossible to simply verify the class distribution of the unlabelled dataset because of the lack of labels. Therefore, SSL models should be tested against class distribution mismatch between labelled and unlabelled data.
In this experiment, we use a class-balanced sample of the training data as labelled data. Unlabelled data consists of samples from the remaining part of the training dataset except for samples with labels 6,7 and 8.

Case	ResNet	Comatch	S-GAN	Mean Teacher
Distribution mismatch	82.45	80.28	78.52	78.13
Similar distribution	82.45	89.86	84.02	85.17

In case of class distribution mismatch performance of SSL models drops significantly in comparison to the case of similar distribution between labelled and unlabelled data. The performance drop is so big that the SSL models have lower accuracy than the fully-supervised baseline. Experiment shows that it is especially important to test the SSL models under various pessimistic constraints. Well performing methods like Comatch might degrade quickly in less optimistic settings.

Experiment 5 - Small validation set

In the Semi-Supervised learning the main idea is to extract as much as information we can with very limited labelled data. Therefore keeping the validation set small would be highly beneficial. Then, more labelled data could be used in the training process. In order to simulate this case we trained SSL models using 20% of the original validation dataset and compared the results with 100% case.

Validation Set	ResNet	Comatch	S-GAN	Mean Teacher
20 % of the original validation set	81.78	87.94	82.85	84.23
100 % of the original validation set	82.45	89.86	84.02	85.17

The drop in performance of the SSL methods is not significant yet noticeable. We conclude that when evaluating SSL methods one should adjust the training set as well as the validation set to reflect realistic settings of SSL model development.

Experiment 6 - Federated Learning

The experiment is based on the FedAvg algorithm [5]. In each round, weights coming from clients are averaged in order to obtain the weights of the Global model. Federated Learning implementation of the SSL methods is performed on 1% and 10% of labelled data. The rest of the training set is the unlabelled dataset. We run the algorithm for 100 rounds, 1 epoch each, with 2 out of 5 clients participating in each round. Data is distributed among Clients in i.i.d. fashion. We used S-GAN and Mean Teacher models in this experiment because the Comatch model turned out to be too complex to transform the learning process into the federated setting in a limited amount of time.

Labelled data	ResNet	S-GAN	Mean Teacher
1% of training data	69.43	54.23	77.88
10% of training data	83.44	84.04	85.81

Fully-supervised baseline performance is slightly better than the performance of corresponding models in non-federated settings. SSL methods generally perform better than the ResNet except for S-GAN trained on 1%. The reason for that might be insufficient labelled data amount in each client. In the case of 1% of training data as labelled data, each client receives only 180 labelled samples.

Discussion

In general, SSL methods perform better than fully supervised baselines. Although before deciding on the model one should evaluate alternatives, like Transfer Learning, especially in cases with limited computational resources. Moreover, the accuracy of SSL methods is sensitive to changes in class distribution between labelled and unlabelled data. Additionally, a too small amount of unlabelled data might hurt the performance of some SSL methods. These constraints should be monitored throughout the development and deployment phases of creating SSL models.

Another observation is the difficulty in hyperparameter tuning of SSL models. This process requires significant computational resources as well as a careful approach to data augmentation techniques. Mean-teacher model turned out to be very sensitive to noise levels during augmentation on the PathMNIST dataset.

Federated Learning allows unlocking the potential of the data stored in a distributed manner among multiple institutions. This solution might be especially promising in healthcare to combine data and knowledge across multiple parties. Simple SSL models like S-GAN or Mean-Teacher can be implemented in the FL setting fairly quickly while preserving good results accuracy. Based on the outcomes of Experiment 6 SSL method should be compared against fully-supervised in case of very limited data stored in participating Clients. More basic SSL methods can be simply transformed into the federated learning setting with promising outcomes. More complex models like Comatch might require tailored FL algorithms.

Conclusion

We showed the applicability of SSL methods in the medical domain and proposed several experiments to comprehensively evaluate SSL methods in realistic settings. Results are consistent with the ones obtained by Oliver et al on natural images.
As future work, we propose an extension of the study by adding more datasets, baseline fully-supervised models, SSL models as well as different objectives such as segmentation or detection. Also, the scope of already conducted experiments can be broadened. For example, different label amounts for the second and third experiments, different cases of class distribution mismatch in the fourth experiment, and different ratios for the size of the validation set for the fifth experiment.

References

1. Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373-440
2. Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. Advances in neural information processing systems, 31.
3. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics
4. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., ... & Ni, B. (2021). Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification
5. Kather, J. N., Weis, C. A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., ... & Zöllner, F. G. (2016). Multi-class texture analysis in colorectal cancer histology. Scientific reports
6. Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., ... & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine
7. Odena, A. (2016). Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583
8. Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30.
9. Li, J., Xiong, C., & Hoi, S. C. (2021). Comatch: Semi-supervised learning with contrastive graph regularization. In Proceedings of the IEEE/CVF International Conference on Computer Vision