PhD thesis defense

Leveraging Domain Adaptation methods for Federated Learning applied to 2D mammography image classification

Gonzalo Iñaki Quintana

PhD advisors: Mathilde Mougeot, Agnès Desolneux
GEHC advisors: Laurence Vancamberg, Vincent Jugnon

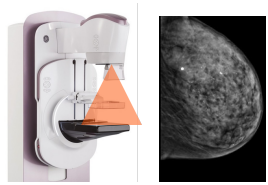
Centre Borelli, ENS Paris-Saclay
Gif-sur-Yvette, France

GE HealthCare
Buc, France

December 6, 2024

Computer Aided Detection (CAD) in mammography

- Improve radiologists image reading efficiency in 2D (FFDM) and 3D (DBT) mammography.
- Data challenges in developing Deep Learning-based CAD:
  - Data collection & privacy concerns
  - Expert annotations
  - Data variability & heterogeneity
- Federated Learning
  - Decentralized learning paradigm
  - Train models locally on-site, aggregate trainings to obtain final model
  - Reduce data collection cost & access to data with large variability



(a) GE Senographe Pristina ™. (b) FFDM image.

Figure: example FFDM image acquisition.

Figure: Federated Learning [1].

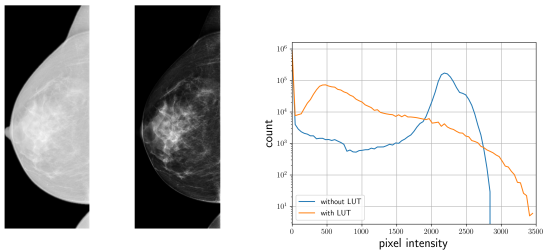
Investigate Federated Learning for addressing the data challenges in DL-based CADs.

[1] Image from S. Pouriyeh et al., published under CC BY 4.0, MDPI Applied Sciences, 2022. DOI: https://doi.org/10.3390/app12188980.

Datasets and heterogeneity

- Full mammography images (GEHC private dataset & public CBIS-DDSM [4])
- Mammography patches
- Synthetic mammography patches

Heterogeneity: image style due to post-processing differences (LUT function as proxy transformation)



(a) w/o LUT (b) w/ LUT (c) pixel intensity histogram

Figure: image-style heterogeneity in mammography.

Table of Contents

Baseline Deep Learning model

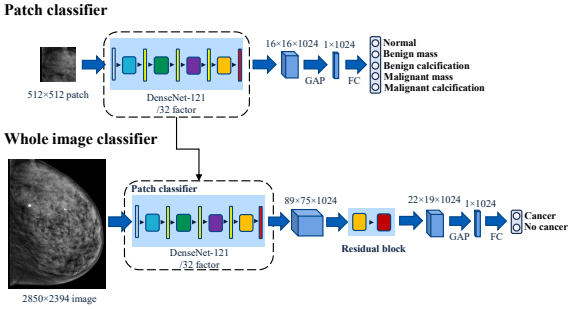


Figure: mammography image patch-based classifier. GAP: Global Average Pooling, FC: Fully Connected layer.

Associated publication

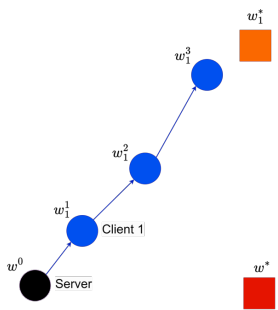
- Impact of patch-size and resolution.
- Proposed multi-patch size & multi-resolution model.

Quintana, G. I., et al. (2023). Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification. Bioengineering, 10(5), 534.

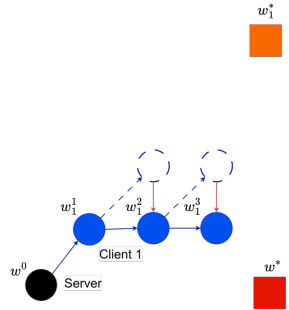
Federated Learning with strong data heterogeneity

Client drift in Federated Learning

- Standard Federated Averaging (FedAvg) algorithm performs well when data is **homogeneous** across clients.
- In **heterogeneous** settings, performance strongly degrades due to client drift.
- Different kinds of algorithms:
  - Regularization (FedProx [6], FedDANE [5])
  - Server-level (adaptive aggregation, weight matching)
  - Variance reduction (SCAFFOLD [3], MIME [2])



(a) FedAvg client drift.



(b) SCAFFOLD client drift correction.

Figure: client drift, where w^* is the global optimal model, w_1^* is the optimal model of the client, and w_1^i is client's model at the i-th iteration.

Client drift in Batch Normalization (BN) layers

- In BN the intermediate features are normalized with empirical mini-batch $\mathcal{B}$ statistics $s_{\mathcal{B}}^t$ (mean and variance).
- As each client normalizes with its own statistics, this introduces an **additional bias** in the gradient updates, not controlled by SOTA FL algorithms (SCAFFOLD, FedProx, etc.):

$$\nabla_w F_i(w, s_{\mathcal{B},i}) = \frac{1}{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \nabla_w \mathcal{L}(w, s_{\mathcal{B},i}; z_j), \quad \forall i \in [1, N]. \tag{1}$$

- Controlling this additional drift is an **open research question**.
- Some algorithms have been proposed but they either lack theoretical guarantees (FedBN [7], SiloBN [1], FixBN [10]) or are computationally prohibitive (FedTAN [9]).

The BN-SCAFFOLD algorithm

Define **statistics control variates** k_i^r for each client i, used with the SCAFFOLD gradient control variates c_i^r.
At each **global step** r and **local step** t, each **client** i controls the **client drift**:

1) Correct statistics during forward pass: $\tilde{s}_i^{r,t} = s_i^{r,t} - k_i^{r-1} + k^{r-1}$,
2) Calculate gradient with corrected statistics: $g_i^{r,t} = g_i(w_i^{r,t}; \tilde{s}_i^{r,t})$,
3) Correct gradient with gradient control variates: $\tilde{g}_i^{r,t} = g_i^{r,t} - c_i^{r-1} + c^{r-1}$,
4) Update the weights with corrected gradient: $w_i^{r,t+1} = w_i^{r,t} - \gamma \tilde{g}_i^{r,t}$.

At the end of the global step, the server averages local control variates to update **global control variates** c^r and k^r:

$$\{c^r; k^r\} \leftarrow \sum_{i=1}^{N} P_i \{c_i^r; k_i^r\}, \tag{2}$$

The BN-SCAFFOLD algorithm

As in SCAFFOLD, two options are provided for calculating the statistics control variates

- **option I:** use full gradients to obtain a good estimation of the gradients and statistics, i.e., $c_i^r = \nabla_w F_i(w_i^{r,0}; S_i^{r,0})$ and $k_i^r = S_i^{r,0}$.
- **option II:** use SGD during E local steps and estimate the control variates from initial and final states of the trainable parameters and the running statistics:

$$\begin{cases} c_i^r = c_i^{r-1} - c^{r-1} + \frac{1}{E\gamma}(w_i^{r,0} - w_i^{r,E}) \\ k_i^r = k_i^{r-1} - k^{r-1} + \frac{1}{1-\rho^E}(\hat{s}_i^{r,E-1} - \rho^E \hat{s}_i^{r,0}) \end{cases} \tag{3}$$

It can be shown that Equation (3) is equivalent to $c_i^r = \frac{1}{E}\sum_{t=1}^{E} g_i(w_i^{r,t}, \tilde{s}_i^{r,t})$ and $k_i^r = \frac{1-\rho}{1-\rho^E}\sum_{t=0}^{E-1} \rho^{E-1-t} s_i^{r,t}$.

Convergence analysis

Define generic family of algorithms using **generic update** functions:

- ϕ: statistics correction.
- Ψ: gradient control variates update.
- Φ: statisitcs control variates update.

Definition

The variance reduction family of algorithms is defined by the following local updates:

$$\begin{cases} \tilde{s}_i^{r,t} = \phi(s_i^{r,t}, k_i^{r-1}, k^{r-1}) \\ w_i^{r,t} = w_i^{r,t-1} - \gamma \left[g_i(w_i^{r,t-1}; \tilde{s}_i^{r,t}) + c^{r-1} - c_i^{r-1} \right] \\ \hat{s}_i^{r,t} = (1-\rho)\hat{s}_i^{r,t-1} + \rho\ \tilde{s}_i^{r,t} \\ c_i^r = \Psi(\{g_i(w_i^{r,t}; \tilde{s}_i^{r,t})\}_{t=1}^E, c_i^{r-1}, c^{r-1}) \\ k_i^r = \Phi(\{\tilde{s}_i^{r,t}\}_{t=1}^E, k_i^{r-1}, k^{r-1}). \end{cases} \tag{4}$$

Convergence analysis

Standard assumptions for convergence theorem:

- Lower-bounded global loss function: $F(w;S) \geq \underline{F} > -\infty$
- Lipschitz continuity: $\|\nabla_w F_i(w';S_i') - \nabla_w F_i(w;S_i'')\| \leq L\|w'-w\| \quad \forall \quad i \in [1,N]$
- Bounded stochastic gradient variance:
  $\mathbb{E}\left[\|g_i(w;s_i) - \nabla_w F_i(w;S_i)\|^2\right] \leq \sigma^2 = \sigma_0^2/|\mathcal{B}|$
- Bounded gradient dissimilarity: $\begin{cases} \sum_{i=1}^N P_i \ \|\nabla_w F_i(w;S_i) - \nabla_w F_i(w;S)\|^2 \leq B^2 \\ \sum_{i=1}^N P_i \ \|\nabla_w F_i(w;S) - \nabla_w F(w;S)\|^2 \leq V^2 \end{cases}$

Additional assumptions for BN-SCAFFOLD:

- Bounded stochastic statistics: $\mathbb{E}\left[\|s_i^{r,t} - S_i^{r,t}\|^2\right] \leq \sigma_s^2 = \sigma_{s,0}^2/|\mathcal{B}|$,
- Lipschitz continuity for the statistics:
  $\begin{cases} \|\nabla_w F_i(w;S_i') - \nabla_w F_i(w;S_i'')\| \leq J\|S_i' - S_i''\| \\ \|S_i' - S_i''\| \leq M\|w' - w''\| \end{cases}$

Convergence analysis

Theorem

The convergence rate of the variance reduction family of algorithms defined in Definition 1 for a general non-convex objective function F, and under the previous assumptions, is given by

$$\min_r \mathbb{E}\left[\|\nabla_w F(\bar{w}_{r-1}; S^{r,0})\|^2\right] \leq \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\|\nabla_w F_i(\bar{w}_{r-1}; S^{r,0})\|^2\right] \leq \frac{\mathcal{T}}{1-2\delta_{E,\gamma,L}}, \tag{5}$$

where $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4 + \mathcal{T}_5 - \mathcal{T}_6$ with

$$\begin{cases} \mathcal{T}_1 = \frac{2}{\gamma RE}\left[F(\bar{w}_0; S^{1,0}) - \underline{F}\right] \\ \mathcal{T}_2 = \frac{8}{R}\delta_{E,\gamma,L}\sum_{j=1}^{N} P_j \|\nabla_w F_j(\bar{w}_0; \tilde{S}_j^{1,0}) - c_j^0\|^2 \\ \mathcal{T}_3 = \frac{2}{R}(1+2\delta_{E,\gamma,L})\sum_{r=1}^{R}\sum_{j=1}^{N} P_j \mathbb{E}\left[\|\nabla_w F_j(\bar{w}_{r-1}; S^{r,0}) - \nabla_w F_j(\bar{w}_{r-1}; \tilde{S}_j^{r,0})\|^2\right] \\ \mathcal{T}_4 = \frac{16}{R}\delta_{E,\gamma,L}\sum_{r=1}^{R}\sum_{j=1}^{N} P_j \mathbb{E}\left[\|\nabla_w F_j(\bar{w}_{r-1}; \tilde{S}_j^{r,0}) - c_j^r\|^2\right] \\ \mathcal{T}_5 = (1+2\delta_{E,\gamma,L})\sigma^2 \\ \mathcal{T}_6 = \frac{1}{RE}(1-\gamma LE - 16L^2E^2\gamma^2\delta_{E,\gamma,L})\sum_{r=1}^{R}\sum_{t=1}^{E}\mathbb{E}\left[\|\sum_{j=1}^{N} P_j\ g_j(w_j^{r,t-1}; \tilde{s}_j^{r,t-1})\|^2\right], \end{cases} \tag{6}$$

and with $\delta_{E,\gamma,L} \coloneqq \frac{4E^2\gamma^2L^2}{1-8E^2\gamma^2L^2}$ and $\gamma < \frac{1}{\sqrt{12}LE}$.

Convergence analysis: application

FL algorithm	Convergence rate	Communication rounds	Communication overhead	Gradients computed
FedAvg	$\mathcal{O}(\frac{L}{\sqrt{R}}[F_0 - \underline{F}] + \frac{\nabla_w F_0^2}{R^2} + \frac{V^2}{R} + B^2 + \frac{\sigma_0^2}{\lvert\mathcal{B}\rvert})$	$2N/E$	$\lvert W\rvert+\lvert S\rvert$	$N\lvert\mathcal{B}\rvert$
SCAFFOLD				
option I	$\mathcal{O}\left(\frac{L}{\sqrt{R}}[F_0 - \underline{F}] + \frac{(\Delta c^0)^2}{R^2} + B^2 + \frac{\sigma_0^2}{\lvert\mathcal{B}\rvert}\right)$	$2N/E$	$2\lvert W\rvert+\lvert S\rvert$	$N\lvert\mathcal{B}\rvert+\lvert\mathcal{D}\rvert/E$
option II	$\mathcal{O}\left(\frac{L}{\sqrt{R}}[F_0 - \underline{F}] + \frac{(\Delta c^0)^2}{R^2} + B^2 + (1 + \frac{1}{R} + \frac{1}{RE})\frac{\sigma_0^2}{\lvert\mathcal{B}\rvert}\right)$	$2N/E$	$2\lvert W\rvert+\lvert S\rvert$	$N\lvert\mathcal{B}\rvert$
BN-SCAFFOLD				
option I	$\mathcal{O}\left(\frac{L}{\sqrt{R}}[F_0 - \underline{F}] + \frac{(\Delta c^0)^2}{R^2} + \frac{\sigma_0^2}{\lvert\mathcal{B}\rvert}\right)$	$2N/E$	$2\lvert W\rvert+2\lvert S\rvert$	$N\lvert\mathcal{B}\rvert+\lvert\mathcal{D}\rvert/E$
option II	$\mathcal{O}\left(\frac{L\Omega}{R}[F_0 - \underline{F}] + \frac{J^2+\Omega}{R}(\Delta k^0)^2 + \frac{1+\Omega}{R}(\Delta c^0)^2 + (\frac{1}{E} + \Omega)\frac{\sigma_0^2}{\lvert\mathcal{B}\rvert} + J^2\Omega\frac{\sigma_{s,0}^2}{\lvert\mathcal{B}\rvert}\right)$	$2N/E$	$2\lvert W\rvert+2\lvert S\rvert$	$N\lvert\mathcal{B}\rvert$
FedTAN	$\mathcal{O}(\frac{L}{\sqrt{R}}[F_0 - \underline{F}] + \frac{\nabla_w F_0^2}{R^2} + \frac{V^2}{R} + \frac{\sigma_0^2}{\lvert\mathcal{B}\rvert})$	$(2 + 6W_D)\frac{N}{E}$	$2\lvert W\rvert+4\lvert S\rvert$	$N\lvert\mathcal{B}\rvert$

Table: comparison of the different algorithms with N clients, E local updates, R global steps, W_D the model depth, and $\Omega := \frac{L^2 - M^2J^2}{L^2 + M^2J^2} > 0$.

Numerical experiments: natural images

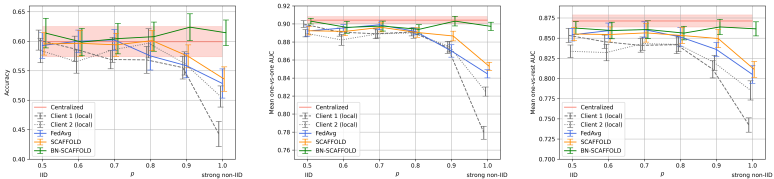
Data heterogeneity: skewed distribution of data labels.

	MNIST (2 clients)	MNIST (5 clients)	CIFAR-10 (2 clients)
Centralized	0.993 ± 0.002 (n.s.)	0.987 ± 0.003 (< 0.01)	0.820 ± 0.012 (< 0.001)
FedAvg [8]	0.983 ± 0.002 (< 0.001)	0.931 ± 0.011 (< 0.001)	0.724 ± 0.015 (< 0.001)
FedBN [7]	0.985 ± 0.003 (< 0.001)	0.936 ± 0.022 (< 0.001)	0.719 ± 0.021 (< 0.001)
SiloBN [1]	0.984 ± 0.002 (< 0.001)	0.935 ± 0.019 (< 0.001)	0.716 ± 0.006 (< 0.001)
FixBN [10]	0.980 ± 0.003 (< 0.001)	**0.982 ± 0.003 (n.s.)**	0.754 ± 0.017 (< 0.01)
FedTAN [9]	**0.992 ± 0.003 (n.s.)**	**0.984 ± 0.003 (n.s.)**	**0.778 ± 0.010 (n.s.)**
SCAFFOLD [3]	0.982 ± 0.004 (< 0.001)	0.934 ± 0.007 (< 0.001)	0.685 ± 0.027 (< 0.001)
FedBN+SCAFFOLD	0.983 ± 0.003 (< 0.001)	0.942 ± 0.011 (< 0.001)	0.697 ± 0.009 (< 0.001)
SiloBN+SCAFFOLD	0.983 ± 0.004 (< 0.001)	0.940 ± 0.011 (< 0.001)	0.693 ± 0.027 (< 0.001)
FixBN+SCAFFOLD	0.979 ± 0.007 (< 0.001)	0.934 ± 0.018 (< 0.001)	0.718 ± 0.043 (< 0.001)
BN-SCAFFOLD (ours)	**0.992 ± 0.002 (n.a.)**	**0.983 ± 0.004 (n.a.)**	**0.776 ± 0.021 (n.a.)**

Table: 5-fold test accuracy, 95% CI, and Welch's t-test p value. n.s.: not significant ($p > 0.05$).

Numerical experiments: mammography patches[2]

- Federated Learning setting with $N = 2$ clients and image style heterogeneity (w/ LUT vs. w/o LUT).
- DenseNet-121 for 5-class classification (lesion and pathology).



(a) accuracy. (b) one-vs-one AUC. (c) one-vs-rest AUC.

Figure: classification performance for different heterogeneity degree p

[2]Similar results were obtained with full mammography images and synthetic patches.

Contrastive-based Domain Adaptation

Domain Adaptation

- **Domain Adaptation (DA):** transfer a model trained on a source domain to a different, but related, target domain.
- Exploit DA to adapt two different domains, $\mathcal{D}_1^*$ and $\mathcal{D}_2^*$, and learn a **domain invariant classifier**.
- the Class-wise Mean Maximum Discrepancy (CMMD) is a widely used measure of Domain Adaptation, and is minimized to increase DA.

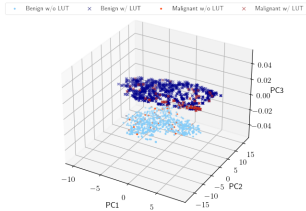


Figure: t-SNE plot of the last features of the whole image classifier, in centralized setting.

Definition (Class-wise Mean Maximum Discrepancy - CMMD)

Given two labeled domains $\mathcal{D}_1^* = \{\mathcal{X} \times \mathcal{Y}, \pi_1\}$ and $\mathcal{D}_2^* = \{\mathcal{X} \times \mathcal{Y}, \pi_2\}$, and the non-linear mapping $\phi : \mathcal{X} \to \mathcal{Z}$. The Class-wise Mean Maximum Discrepancy (CMMD) is defined as:

$$\text{CMMD}(\mathcal{D}_1^*, \mathcal{D}_2^*, \phi) = \mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\left\| \mathbb{E}_{X \sim \pi_{1,C}^{\mathcal{X}|\mathcal{Y}}}[\phi(X)] - \mathbb{E}_{X \sim \pi_{2,C}^{\mathcal{X}|\mathcal{Y}}}[\phi(X)] \right\|_{\mathcal{Z}} \right]. \tag{7}$$

Class-separability

- Separability of classes in the feature space is commonly associated to higher classification performance.
- If not controlled, Domain Adaptation can be attained at the expense of **reducing class-separability**.
- We introduce an MMD-based measure of class-separability that considers class-separability within individual domains, and across domains.

Definition (Different-class Mean Maximum Discrepancy - DCMMD)

Given two labeled domains $\mathcal{D}_1^* = \{\mathcal{X} \times \mathcal{Y}, \pi_1\}$ and $\mathcal{D}_2^* = \{\mathcal{X} \times \mathcal{Y}, \pi_2\}$, a mixed domain $\mathcal{D}_{m,p}^* = \{\mathcal{X} \times \mathcal{Y}, \pi_{m,p}\}$, and the non-linear mapping $\phi : \mathcal{X} \to \mathcal{Z}$. The Different-class Mean Maximum Discrepancy (DCMMD) is defined as:

$$\text{DCMMD}(\mathcal{D}_1^*, \mathcal{D}_2^*, \phi) = \mathbb{E}_{C_1, C_2 \sim \pi^{\mathcal{Y}}_{m,1/2}; C_1 \neq C_2; D_1, D_2 \sim Ber(1/2)} \left[\left\| \mathbb{E}_{X \sim \pi^{\mathcal{X}|\mathcal{Y}}_{D_2, C_1}} [\phi(X)] - \mathbb{E}_{X \sim \pi^{\mathcal{X}|\mathcal{Y}}_{D_1, C_2}} [\phi(X)] \right\|_{\mathcal{Z}} \right], \tag{8}$$

where $\pi^{\mathcal{Y}}_{m,p}$ is the marginal probability measure on the labels of the mixed domain, and $Ber(p)$ is the Bernoulli distribution.

Contrastive Learning

- **Contrastive Learning:** representation learning paradigm, widely used in Self-supervised Learning.
  - Drag together semantically similar features.
  - Pull apart semantically different features.
- We investigate using Supervised Contrastive Learning for learning a **domain-invariant** feature extractor, while increasing **class-separability**.

Definition (Supervised Contrastive loss)

Given a batch of instances $\mathcal{B}$, where $z_i = \phi(x_i)$ is the feature representation of instance x_i, and $\mathcal{P}(i) = \{j \in \mathcal{A}(i) : y_j = y_i\}$ is the set of indices of the positive counterparts of feature z_i, the Supervised Contrastive loss is defined as:

$$\mathcal{L}_{SupContr} = -\frac{1}{|\mathcal{B}|} \sum_{i \in |\mathcal{B}|} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{e^{z_i \cdot z_j / \tau}}{\sum_{l \in \mathcal{A}(i)} e^{z_i \cdot z_l / \tau}}. \tag{9}$$

Contrastive Learning and Domain Adaptation

Lemma

In a high temperature regime, the Supervised Contrastive loss and the NT-Xent loss can be expressed in terms of the CMMD by the following equation:

$$\begin{aligned}\tau \mathcal{L}_{Contr} \approx \frac{1}{4} CMMD^2(\mathcal{D}_1, \mathcal{D}_2, \phi) + & \underbrace{\mathbb{E}_{X,X' \sim \pi^{\mathcal{X}}_{m,p=1/2}} \left[k(X, X')\right]}_{\textbf{Similarity between all pairs of features}} \\ & - \frac{1}{2} \underbrace{\mathbb{E}_{C \sim \pi^{\mathcal{Y}}} \left[\mathbb{E}_{X,X' \sim \pi^{\mathcal{X}|\mathcal{Y}}_{1,C}} \left[k(X, X')\right] + \mathbb{E}_{X,X' \sim \pi^{\mathcal{X}|\mathcal{Y}}_{2,C}} \left[k(X, X')\right] \right]}_{\textbf{Similarity between features with the same class and domain}} \\ & + \frac{1}{2\tau} \underbrace{\mathbb{E}_{X \sim \pi^{\mathcal{X}}_{m,p=1/2}} \left[\mathbb{V}ar_{X' \sim \pi^{\mathcal{X}}_{m,p=1/2}} \left[k(X, X')\right] \right]}_{\textbf{Variance term}} \\ & + \underbrace{\tau \log\left(|\mathcal{B}|-1\right)}_{\textbf{Constant, does not affect optimization}} + \underbrace{\mathcal{O}\left(\frac{\mathbb{E}_{X \sim \pi^{\mathcal{X}}_{m,p=1/2}} \left[\mathbb{V}ar_{X' \sim \pi^{\mathcal{X}}_{m,p=1/2}} \left[k(X, X')\right]^2 \right]}{\tau^3} \right)}_{\to 0 \text{ as } \tau \text{ increases}}. \end{aligned} \tag{10}$$

Minimizing the contrastive losses decreases the CMMD, increasing Domain Adaptation.

Contrastive Learning and class-separability

Lemma

By assuming that the kernel over X is bounded $|k(x,x')| < k^{max}$, $\forall x, x'$, and that the kernel over Y $l(y_i, y_j) = \Delta l \; \mathbb{1}_{\{y_i = y_j\}} + l_0$ satisfies $\Delta l = M$ for the Supervised Contrastive loss and $\Delta l = N$ for the NT-Xent loss, the contrastive losses bound the inter-class MMD:

$$-\frac{1}{\alpha}\underbrace{\mathbb{E}_{C_1,C_2\sim\pi^{\mathcal{Y}}_{m,1/2}}\left[\left\|\mathbb{E}_{X\sim\pi^{\mathcal{X}|\mathcal{Y}}_{m,1/2,C_1}}[\phi(X)] - \mathbb{E}_{X\sim\pi^{\mathcal{X}|\mathcal{Y}}_{m,1/2,C_2}}[\phi(X)]\right\|^2\right]}_{\text{inter-class MMD} \geq 0} \tag{11}$$
$$+ \underbrace{\gamma HSIC(X,X)}_{\text{Hilbert-Schmidt Independence Criterion}} + \underbrace{\mathcal{O}\left(\mathbb{V}ar\left[k\left(X,X'\right)\right]\right)}_{\text{Variance term}} \leqslant \mathcal{L}_{Contr},$$

where $\gamma \in \mathbb{R}$ is a constant satisfying $\min\{-2, -2k^{max}\} = -(1+\sqrt{1-4\gamma})/(2\gamma)$.

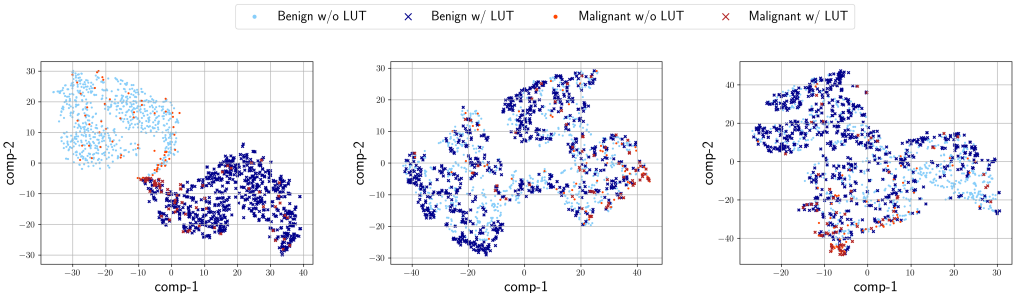
Minimizing the contrastive losses:

- Increases inter-class MMD, and thus **increases class-separability**.
- Decreases the HSIC between all the features, i.e., the level of cross-covariance.

Application: full mammography images

- Cancer vs. no cancer classification:
- Image style heterogeneity (w/ LUT vs w/o LUT).
- Compare models trained with (a) Cross Entropy (CE) loss, (b) Supervised Contrastive (SupContr) loss, (c) SupContr and CE fine-tuning (SupContr+CE).

Qualitative evaluation

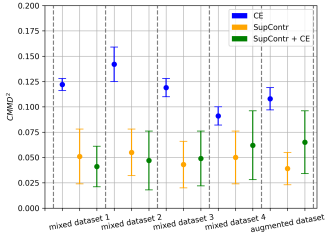


(a) CE. (b) SupContr. (c) SupContr+CE.

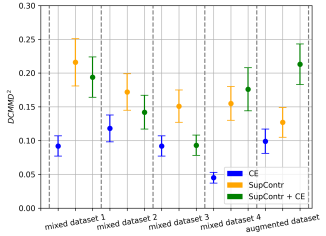
Figure: t-SNE plots of the features from the whole image classifier, indicating class and domain.

Application: full mammography images

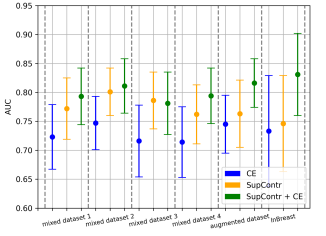
Quantitative evaluation



(a) Domain Adaptation (CMMD).



(b) class-separability (DCMMD).



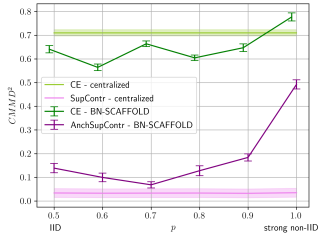
(c) classification (AUC).

Figure: full mammography classification results, for different training splits.

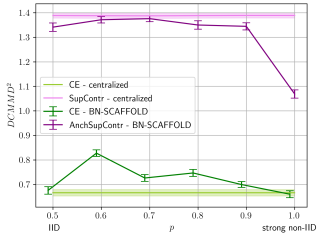
Contrastive Learning increases Domain Adaptation, class-separability, and classification performance.

Domain Adaptation in federated setting

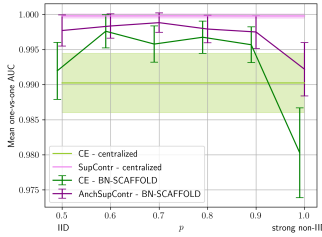
- A direct extension of Contrastive Learning to the federated setting threatens **data privacy**, as it involves the transfer of features.
- We propose the **Anchored Supervised Contrastive (AnchSupContr)** loss:
  - Calculate similarities between features and per-class means of the features.
  - Per-class means can be transferred out of the clients with less risk on data privacy.



(a) Domain Adaptation (CMMD).



(b) class-separability (DCMMD).



(c) classification (one-vs-one AUC).

Figure: performance of the patch-classifier with synthetic patches and $N = 2$ clients.

Main contributions and future research directions

Main contributions

Mammography image classification

- Studied the impact of patch-size and resolution in mammography image classifier.
- Proposed a multi-patch size and a multi-resolution approach, achieving SOTA performance on CBIS-DDSM.

Federated Learning

- Proposed BN-SCAFFOLD, which achieves the best trade-off between convergence rate, communication cost, and computational complexity.
- Developed a unified theoretical framework for obtaining the convergence rate of different FL algorithms.
- Showed that BN-SCAFFOLD outperforms other state-of-the-art methods when heterogeneity is strong, in both natural images and in mammography images (synthetic and clinical).

Main contributions

Domain Adaptation

- Theoretically showed that minimizing contrastive losses increases DA and class-separability.
- Leveraged Contrastive Learning to obtain domain-invariant models, increasing Domain Adaptation, class-separability, and classification performance (synthetic and clinical images).

Federated Learning and Domain Adaptation

- Modified the Supervised Contrastive loss to avoid transferring the client's features.
- Showed increased Domain Adaption, class-separability, and classification performance in synthetic images. Additional research is needed for extending these results to clinical images.

Future research directions

- Federated Learning in other clinical applications (e.g., lesion detection or segmentation).
- Data privacy guarantees and robustness to attacks (poisoning, inference, model manipulation, free-rider, etc.) in Federated Learning.
- Federated Learning and Self-supervised Learning.
- Federated Learning and Foundation Models.

Associated publications

[1] Quintana, G. I., Li, Z., Vancamberg, L., Mougeot, M., Desolneux, A., & Muller, S. (2023). Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification. Bioengineering, 10(5), 534.

[2] Quintana, G. I., Jugnon, V., Vancamberg, L., Desolneux, A., & Mougeot, M. (2024, May). Contrastive learning: an efficient Domain Adaptation strategy for 2D mammography image classification. In 2024 IEEE 21st International Symposium on Biomedical Imaging (ISBI).

[3] Quintana, G. I., Vancamberg, L., Jugnon, V., Mougeot, M., & Desolneux, A. (2024). BN-SCAFFOLD: controlling the drift of Batch Normalization statistics in Federated Learning. arXiv preprint arXiv:2410.03281.

[4] Garin, M., & Quintana, G. I. (2023). Incidence of the sample size distribution on one-shot federated learning. Image Processing On Line, 13, 57-64.

References I

[1] Mathieu Andreux et al. "Siloed federated learning for multi-centric histopathology datasets". In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer. 2020, pp. 129–139.

[2] Sai Praneeth Karimireddy et al. "Mime: Mimicking Centralized Stochastic Algorithms in Federated Learning". In: *CoRR* abs/2008.03606 (2020). arXiv: 2008.03606. URL: https://arxiv.org/abs/2008.03606.

[3] Sai Praneeth Karimireddy et al. "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 5132–5143. URL: https://proceedings.mlr.press/v119/karimireddy20a.html.

[4] Rebecca Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific Data* 4 (Dec. 2017), p. 170177. DOI: 10.1038/sdata.2017.177.

[5] Tian Li et al. "Feddane: A federated newton-type method". In: *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2019, pp. 1227–1231.

[6] Tian Li et al. "Federated optimization in heterogeneous networks". In: *Proceedings of Machine learning and systems* 2 (2020), pp. 429–450.

[7] Xiaoxiao Li et al. "FedBN: Federated Learning on Non-IID Features via Local Batch Normalization". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=6YEQUn0QICG.

[8] Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[9] Yanmeng Wang, Qingjiang Shi, and Tsung-Hui Chang. "Why Batch Normalization Damage Federated Learning on Non-IID Data?" In: *IEEE Transactions on Neural Networks and Learning Systems* PP (Nov. 2023), pp. 1–15. DOI: 10.1109/TNNLS.2023.3323302.

[10] Jike Zhong, Hong-You Chen, and Wei-Lun Chao. "Making Batch Normalization Great in Federated Deep Learning". In: *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*. 2023.

Thank You
for Listening.