Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2025) 621-638

ECCOMAS
Proceedia

UNCECOMP 2025
6th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Science and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15-18 June 2025

TRAINING AND DEPLOYMENT OF A DEEP ABSTAINING CLASSIFIER UNDER FEDERATED LEARNING

Cristina Garcia-Cardona[1] **and Jamal Mohd-Yusof**[1]

[1]Los Alamos National Laboratory
Bikini Atoll Rd., SM 30, Los Alamos, NM 87545
e-mail: {cgarciac,jamal}@lanl.go v

Abstract. *In this work we extend the original deep abstaining classifier (DAC) model to be trained under a Federated Learning (FL) setup, with the goal of tackling a wider range of problems including distributed data silos that cannot be freely exchanged between different owner entities, e.g. electronic health records (EHR) from different registries, due to privacy considerations. We use typical FL aggregation strategies for the model parameters and devise new strategies for aggregating the parameters related to the abstention loss function. These are needed to allow the models to achieve their accuracy targets on disparate silos. We implement the DAC+FL training under Flower, a simple, open source federated learning framework. We apply the models to EHR data and compare vs. different baselines. We demonstrate that DAC can be trained under FL, obtaining a performance that is similar to the DAC in consolidated training, sometimes with a minimal drift towards increased abstention rate. We also find that the different aggregation strategies proposed produce similar results. Finally, we indicate how the DAC+FL model can be used for anomaly detection with the possibility of tuning at a local silo level.*

Keywords: Deep abstaining classifie , federated learning, anomaly detection.

ISSN:2623-3339
Peer-review under responsibility of the organizing committee of UNCECOMP 2025.
doi: 10.7712/120225.12382.21454

1 INTRODUCTION

Federated Learning (FL) was introduced in [1] as an alternative for training DNNs using privacy sensitive data that is distributed across separate devices (*silos*). Instead of storing data in a centralized location, FL learns a shared model via local aggregation of updates, each computed by a participating device (i.e. *client*) and communicated to a global coordinating *server*. This allows decoupling of the model training from the access to the raw data, collectively benefitin from the richness of the individual training datasets, while reducing privacy and security risks.

The basic FL approach usually involves performing local optimization, such as stochastic gradient descent (SGD), on each client and then computing an aggregated model on the server. McMahan et al. [1] demonstrate that FL is robust to unbalanced and non-independently and identically distributed (IID) data, and can operate under limited communication constraints. Further work has considered other open problems in FL [2], as well as general federated machine learning applications [3], and privacy issues [4, 5]. In medicine, in particular, where preserving privacy and reducing the security risks of training with data from electronic health records (EHR) are of utmost importance, FL constitutes an efficien and robust alternative [6, 7, 8] to consolidated training.

In this work we adapt the deep abstaining classifie (DAC) model [9] to be trained in a FL setting and demonstrate its application to EHR data. The DAC identifie data samples that are ambiguous given the input features and abstains from making predictions on them. Instead of discarding these ambiguous samples, they are kept during training to allow the deep neural network (DNN) model to learn useful information regarding the data distribution and associated noise characteristics. The DAC can be constructed on top of any DNN classifie by implementing minimal modification to the output layer and the training procedure.

The DAC+FL combination brings the fl xibility of the abstention framework to the efficien privacy-preserving distributed learning of FL. This has the potential to allow for more extensive applications of the DAC model to settings where privacy preserving considerations are crucial and, at the same time, enable a fine (i.e. local-level) control of the abstention performance by setting different operational targets per client, without significantl impacting overall performance. The main contributions of this work can be summarized as follows:

- Demonstrate that the DAC can be trained under FL setup.
- Show that the performance of the DAC+FL model is similar to the DAC in consolidated training, with a minimal drift towards increased abstention rate per client.
- Demonstrate the potential for the DAC+FL to be deployed as an anomaly detector.

This manuscript is organized as follows. Sec. 2 briefl reviews the DAC and describes how it is adapted to FL. Sec. 3 includes results for numerical experiments using EHR data. We evaluate performance with respect to consolidated (i.e. non-FL) models for homogeneous and heterogeneous datasets and demonstrate how the model can be deployed for anomaly detection with local tuning. Finally, Sec. 4 draws some conclusions about the combined model and suggests directions for future work.

2 DEEP ABSTAINING CLASSIFIER (DAC)

In a traditional classificatio problem, for a given set of input features $\boldsymbol{x}$, let y be the class to be predicted by the DNN. The probability of the ith class given $\boldsymbol{x}$, denoted $p_i = p_{\boldsymbol{w}}(y = i|\boldsymbol{x})$, can be define as the i^{th} output of the DNN, which implements the probability model $p_{\boldsymbol{w}}(y = i|\boldsymbol{x})$

(using a softmax function as its final layer) with $\boldsymbol{w}$ denoting a set of weight matrices that parameterize the DNN. For notational brevity, we use p_i to represent $p_{\boldsymbol{w}}(y = i|\boldsymbol{x})$.

The standard cross-entropy training loss for a DNN classifier with k classes can be expressed as:

$$\mathcal{L}_{\text{standard}}(\boldsymbol{x}) = -\sum_{i=1}^{k} t_i \log p_i \tag{1}$$

where $t_i \in \{0, 1\}$ is the target for the current sample i.

The DAC [9] modifies the DNN output layer to include an additional *abstention* class with a corresponding additional $k + 1^{\text{st}}$ output p_{k+1} which is the probability of abstention. This is achieved by training the DAC with the following modified version of the k-class cross-entropy per-sample loss:

$$\mathcal{L}_{\text{DAC}}(\boldsymbol{x}) = (1 - p_{k+1}) \left(-\sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha \log \frac{1}{1 - p_{k+1}} \tag{2}$$

where k is the number of classes excluding the abstention class, t_i is the true label of training data for class i (it is one when i is the true label, zero otherwise), $k + 1$ is the abstention class, p_{k+1} is the probability of the abstention class and $\alpha > 0$ weights the penalty term for abstention.

This loss function behaves like a regular cross-entropy loss on the original classes and adds an additional loss term, scaled by a tuning parameter α, that controls the propensity for abstention. Since the balance between these terms depends on the data and the sources of confusion, the optimal value of α cannot be determined a priori and must be tuned during the training process. Importantly, this parameter is tuned during training to guarantee a target accuracy while minimizing the abstention rate. A very high value of α means a high penalty for abstaining, driving the model towards no abstention. Conversely, a very low value of α may allow the model to abstain on everything.

2.1 Adaptive Abstention

To better understand how the α parameter is tuned, we describe the actual training of a DAC. The training is carried out by optimizing (2) while simultaneously tuning α to satisfy two complementary goals: (i) to achieve a target minimum accuracy while (ii) minimizing the total abstention.

Algorithm 1 Algorithm for Alpha Adaptation in Classificatio

Require: $r_{\text{max-abs}} \in (0, 1)$, $A_{\text{min-abs}} \in (0, 1)$ and $\alpha^{(k)} > 0$
Ensure: $\alpha^{(k+1)} > 0$
$e_A \leftarrow \max\{A_{\text{min-abs}} - A, 0\}$
$e_r \leftarrow \max\{r - r_{\text{max-abs}}, 0\}$
$a \leftarrow 1 - g_A e_A + g_r e_r$
$\alpha^{(k+1)} \leftarrow a\, \alpha^{(k)}$

Algorithm 1 describes the process for tuning α, a process that is executed after each training epoch. Let the abstention rate be defined as $r =$ # abstaining samples/# total samples, a fraction between 0 and 1. Also, let $r_{\text{max-abs}}$ denote the target maximum abstention rate and let $A_{\text{min-abs}}$ denote the target minimum accuracy. These target values are set at the beginning to reflect performance goals and are kept fixed during training. Note that the updating of a, the α

adaptive multiplier, depends on e_A, the difference between target and achieved minimum accuracy, and on e_r, the difference between target and achieved abstention rate. A positive e_A means that the accuracy is less than the target value set. As shown in the a update in algorithm 1, this nudges a to be less than one, which in turn means a smaller α, making it cheaper to abstain. This increases the abstention rate and potentially increases the accuracy by removing noisy samples from having any effect on the firs term of eq. (2). On the other hand, a positive e_r means that the abstention rate is greater than the target set. In the a update, this pushes a to be greater than one, which increments α, making it more expensive to abstain. This should decrease the abstention rate, albeit at the price of a possible decrement in accuracy.

Starting from an initial value of alpha $\alpha^{(0)} > 0$, this adaptive process should achieve a stable α value asymptotically. To avoid large α swings while training, the factor a can be confine to vary in a specifie range (e.g. $a \in [0.8, 1.25]$). The adaptation can be made more sensitive to one or the other target setting by using different multiplier factors (i.e. *gains* g_A, g_r) for e_A and e_r when updating a. Note that we use a validation set during training and base the α adaptation on the metrics computed over this set (not the training set). By this adaptive procedure, it has been shown [9] that the resulting DAC trained models are more robust to feature noise, while providing a powerful tool for identifying uncertain predictions at inference time.

The DAC has also shown to be efficien for multi-task problems [9], i.e. for simultaneously solving multiple non-overlapping classificatio tasks. For the multi-task case, the DAC poses each subclassificatio problem as a classificatio plus abstention, each requiring its own α and its own tuning during training. Earlier versions of the DAC were tuned using a combination of accuracy and abstention targets for each task. In the latest versions, used for this work, we have implemented training methods that, for each task, allow targets for either accuracy or abstention alone, in addition to the original mixed targets. Effectively this is equivalent to setting either $g_A = 0$ (pure abstention target) or $g_r = 0$ (pure accuracy target), but with a modifie definitio of e_A (respectively, e_r) aiming to preserve accuracy (or abstention rate) inside a given interval, i.e., for accuracy, if it is smaller than the lower limit, it triggers a decrease in alpha, while, if accuracy is greater than the upper limit, it triggers an increase in alpha. For this study, we train only for accuracy; specificall , in order to guarantee 97% accuracy, we choose accuracy targets of $97.5\% \pm 0.5\%$. It is possible to set tighter bounds on the accuracy (or abstention) targets, but this can produce significantly longer training times, especially in a multi-task setting. In a multi-task setting, the stopping criteria require that every task satisfy the desired minimum accuracy target (which can be set independently for each task), except when a task is able to exceed the target with zero abstention, in which case that task is considered to have satisfie its target.

2.2 DAC Training Under Federated Learning

For developing a DAC+FL training algorithm, we make use of FL methods typically employed to aggregate the model parameters and only introduce new methodologies to handle the α adaptation under the distributed FL setup.

The normal training of a model under FL is divided in multiple rounds, with each round composed of a number of epochs trained on individual silos, and local updates executed per client after each epoch, as in conventional DNN training. After a round is completed, the parameters of each silo are globally aggregated (typically, weighted mean values, with weights associated to the size of the silo) by the server, and an updated consolidated model is obtained. The server shares the consolidated model with the clients, so that they get back the updated parameters before a new training round starts.

To combine this round-based FL with the DAC training, we allow each round to be executed as in the normal FL training, with the α tuning proceeding independently in each client as described in Algorithm 1. What we need to consider then is how the client-wise tuned α's should be synchronized by the server to produce globally updated α parameters that will be used to optimize the abstention loss function (2) in the following round. For this synchronization we devise and compare two different strategies.

Local: This strategy allows the clients to proceed with the optimization using their *'local'* α parameter values without any global synchronization by the server, i.e. the α's evolve independently in each client, and the only global synchronization comes via the updates to the DNN model parameters performed by the server after each round. Hence, each client starts the next round with aggregated network weights and with the *'local'* α's saved from the previous round.

Shared: This strategy allows the server to aggregate both α parameter values and model weights after each round, following the same policy (e.g. weighted average across clients). This yields a model where each client starts the next round with aggregated network weights and *'shared'* α's across the different clients.

These strategies aim to capture the intuition that if the data silos are similar, then the shared strategy should produce the fastest convergence, while if the silos are dissimilar (in size, level or type of noise in the features, etc.) then allowing each one to adapt their α's independently may be the most robust and stable strategy.

We implement these two strategies for the α aggregation in the training of the FL+DAC model using the Flower Framework [10], a simple framework to bring existing machine learning workloads into a federated setting. This results in a scalable, well-tested, and open-source infrastructure.

3 NUMERICAL EXPERIMENTS

This section presents some of the results of training the DAC under a FL setting. We compare performance with baselines including no-abstention and non-federated learning DNN classifie models. First we contrast the performance of the DAC trained with the regular procedure (a.k.a. consolidated learning) vs. the performance based on a FL implementation, and demonstrate that the FL setup does not affect the main characteristics of the DAC. Then, we compare the performance of the two proposed alpha aggregation approaches, namely local and shared, both in experiments with homogeneous and heterogeneous data silos. We finaliz by illustrating how the DAC+FL can be used for silo-wise anomaly detection.

3.1 Dataset Description

In this work, we use text pathology reports from the US National Cancer Institute SEER (Surveillance, Epidemiology, and End Results) registries [11]. Each case of cancer (individual tumor), given by the case ID, is identifie by a combination of a patient ID and a tumor ID. Ground truth for each case of cancer is obtained from the manually abstracted and consolidated records in the cancer registries. There may be multiple reports for each case, each of which are identifie by a combination of a patient ID, a tumor ID, and a report ID, and has assigned values for five tasks: behavior (4 classes), histological type ($\sim$ 550 classes), laterality (7 classes), primary site ($\sim$ 70 classes) and primary subsite ($\sim$ 300 classes). In addition to the disparity in number of classes, there is a wide disparity in prevalence across classes. The ground truth is

consolidated for each case ID (i.e., each individual tumor), meaning all the reports pertaining to a particular tumor have the same ground truth regardless of the content of the text pathology report. Previous work to develop models for automated annotation of cancer pathology reports [12, 13, 14, 15] demonstrated the value of DAC methods in a consolidated training setting. For the purposes of this work, we focus on the mechanics of enabling DAC training in a FL setting rather than the details of the dataset or the specifi FL aggregation schemes.

3.2 Homogeneous Dataset Experiments

To assess the performance of DAC under consolidated and FL training modalities we selected the registry of Louisiana (LA), which is composed of 482,914 samples, divided into static train (338,249 samples), validation (72,391 samples) and test (72,274 samples) partitions. We used the registry to train one DAC model with a target accuracy of $97,5\% \pm 0.5\%$ for each task. We also simulated a federated environment by using two silos with the same dataset. By mimicking a FL environment using the same registry per silo and target configuratio we build an homogeneous dataset case where the results should be comparable with the consolidated training. Furthermore, by making homogeneous silos, we are more certain that any differences detected in the results are probably introduced by the different training modalities evaluated, and not caused by dataset issues.

We configure the FL for 4 rounds of training with a maximum of 25 epochs per round. We trained DAC+FL models using the proposed α aggregation strategies (i.e., local and shared). All models used analogous configurations with the same rounds $\times$ max. epochs and with the same abstention and accuracy targets. A maximum of 100 epochs was set for the non-FL training. We repeated the training for each model 5 times using different initializations.

Figure 1 displays the evolution of α for one instantiation of the training of the DAC+FL model. The left column shows the α evolution for local aggregation in each of the two clients. The righ column shows corresponding results for the shared aggregation. The 4 rounds are concatenated and the transition between rounds is indicated by vertical dotted lines. It can be seen that α's evolve smoothly in both aggregation strategies. We also observe that the shared strategy seems to require a few more iterations to converge, but, in the end, both strategies converge to very similar α values. Figures 2– 3 show analogous results for the evolution of the abstention rate and the accuracy in the validation set. Recall that is the trade-off between abstention rate and accuracy *in the validation set* which allows the dynamical tuning of the α values. These figure demonstrate that both strategies are able to satisfy the target accuracy, and that performance is similar in both strategies, albeit with a slightly slower convergence for the shared strategy as already mentioned.

To get a sense of the dispersion with respect to different training initializations, we report in Table 1 the number of training epochs used per round, per each repetition of the models evaluated. It can be seen that the base DNN classifie , that does not use abstention or FL ('No Abs, No FL' row in table), requires an average of 32 epochs to converge[1]. The same DNN classifie when trained in a FL setup ('No Abs' row in table) requires an average of about 41 epochs[2] (same convergence definitio as[1]), suggesting that FL tends to yield marginally slower convergence, which is usually justifie by the noise that is injected to maintain the privacy

[1]Convergence here is define as achieving the target minimum accuracy per task, or early stopping if no improvement is obtained after a patience of 5 epochs.

[2]The direct estimation of about 58 epochs on average to converge would include some extra patience epochs that are triggered due to the length of the rounds in FL. The average of 41 epochs for convergence is obtained after discarding the extra patience epochs, i.e. 5 patience epochs $\times$ 3 rounds $\simeq$ 15 extra epochs.

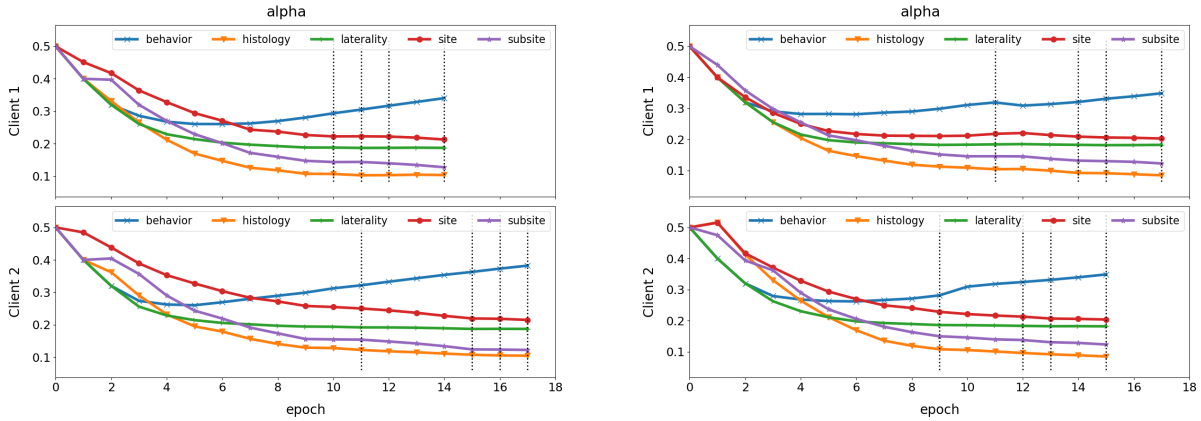

Figure 1: Evolution of alpha: one instantiation of DAC+FL training for for each of the 5 tasks (behavior, histology, laterality, site, subsite). Rounds are concatenated, vertical dotted lines indicate the different rounds. Left: local alpha aggregation. Right: shared alpha aggregation.

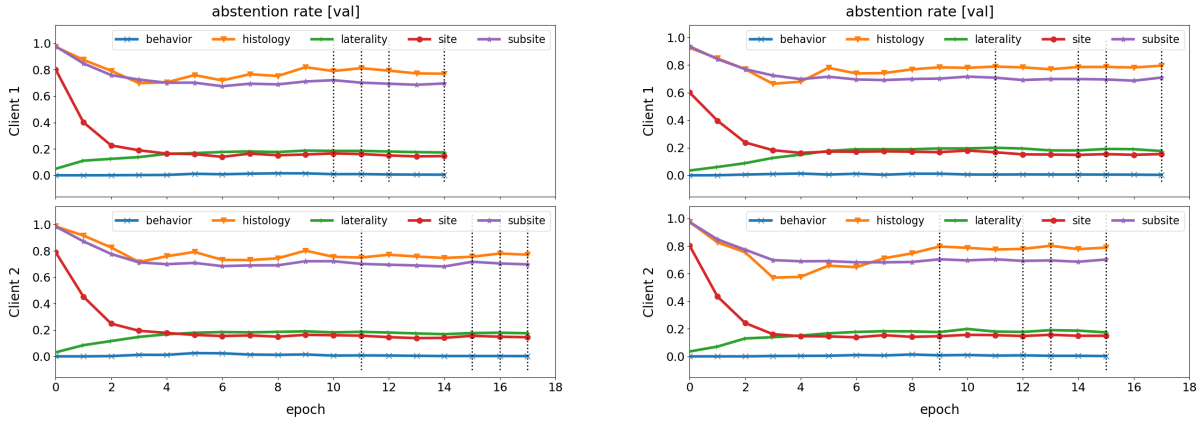

Figure 2: Evolution of abstention rate in validation set: one instantiation of DAC+FL training for each of the 5 tasks (behavior, histology, laterality, site, subsite). Left: local alpha aggregation. Right: shared alpha aggregation.

of the silo. Comparing with the DAC baseline ('No FL' row in table) which uses abstention training but no FL, it can be seen that convergence is much faster, requiring about 12 epochs to converge[3]. This illustrates the influenc of the abstention loss function (2) in regularizing the training and in helping to identify the more noisy samples. The proposed DAC+FL models exhibit a convergence (same convergence definitio as[3]) that is closer to the DAC baseline, with about 17 epochs on average for both local and shared strategies. Even though in one realization it seems that 'shared' was slower than 'local', both strategies have the same performance on average. This is somewhat expected, due to the fact that the silos are basically copies of each other in the homogeneous experiments, so there are no differences in data noise to prompt different local adjustments in α's or abstention rates.

Figures 4, 5 and 6 plot the α's, accuracies and abstention rates, respectively, obtained at the end of training for the validation set. It is evident that any of the abstention-based methods is able to achieve the set minimumaccuracy of 97%. This highlights the benefi of using abstention

[3]Convergence here is define as achieving the target minimumaccuracy per task, without surpassing the maximum target abstention rate per task, or early stopping if no improvement is obtained after a patience of 5 epochs.

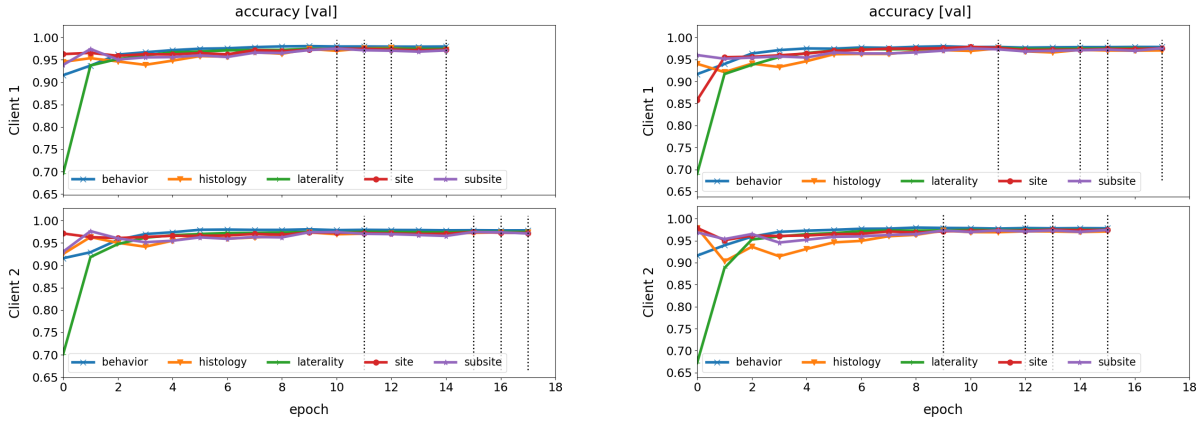

Figure 3: Evolution of accuracy in validation set: one instantiation of DAC+FL training for each of the 5 tasks (behavior, histology, laterality, site, subsite). Rounds are concatenated, vertical dotted lines indicate the different rounds. Left: local alpha aggregation. Right: shared alpha aggregation.

Method	Repetition 1	Repetition 2	Repetition 3	Repetition 4	Repetition 5
DAC + FL Local	11, 1, 1, 2	10, 3, 1, 1	13, 6, 1, 1	14, 1, 1, 1	11, 2, 2, 1
	12, 4, 1, 1	11, 1, 1, 1	15, 2, 2, 2	12, 2, 2, 1	11, 1, 2, 1
DAC + FL Shared	12, 3, 1, 2	14, 1, 1, 1	12, 1, 1, 1	12, 1, 2, 2	12, 2, 1, 1
	10, 3, 1, 2	13, 1, 3, 1	12, 2, 2, 2	14, 2, 2, 1	12, 2, 2, 1
No FL	10	11	12	13	12
No Abs	25, 16, 9, 6	25, 19, 6, 7	25, 17, 9, 6	25, 18, 6, 9	25, 17, 7, 9
	25, 16, 9, 7	25, 19, 6, 10	25, 17, 9, 6	25, 18, 9, 9	25, 17, 7, 7
No Abs, No FL	35	33	34	28	32

Table 1: Number of epochs in training per repetition. For FL-trained models this is reported by client and by round.

for detecting unreliable samples. It is also clear that both local and shared strategies converge to similar α values and accuracies. The major dispersions in accuracy observed are correlated with the tasks that exhibit lower abstention rates. This can be explained as follows. The abstention in the behavior task is zero, so the α parameter is really meaningless since it is being applied to a zero abstention term. Hence, the *apparent* differences in α between local and shared strategies for the behavior task are irrelevant. Laterality and site tasks show only minor divergences and low abstention rates. The histology and subsite tasks are much more tightly correlated in α's, reflectin the increased difficult posed by tasks with hundreds of classes, high class imbalance, and more ambiguity ([14]). This also results in higher abstention levels required to achieve the target minimum accuracy of 97%.

To finaliz the set of experiments with homogeneous silos, we present in Tables 2 and 3 the accuracy and abstention rate, respectively, evaluated in the testing set and averaged over all the trained models in each modality. These results show that the results in the testing set are consistent with the ones obtained in the validation set. Overall, both local and shared DAC+FL strategies match the DAC baseline, corroborating the statement that DAC can be trained under FL setup while achieving a performance that is very similar to the DAC in consolidated training for the homogeneous dataset, as expected.

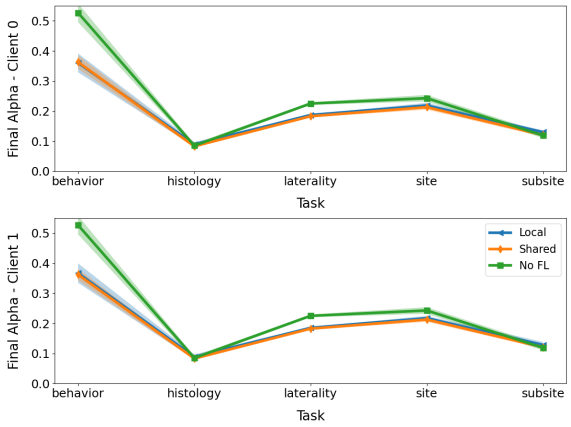

Figure 4: Values of alphas at the end of training for the compared modalities: DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation and DAC without FL.

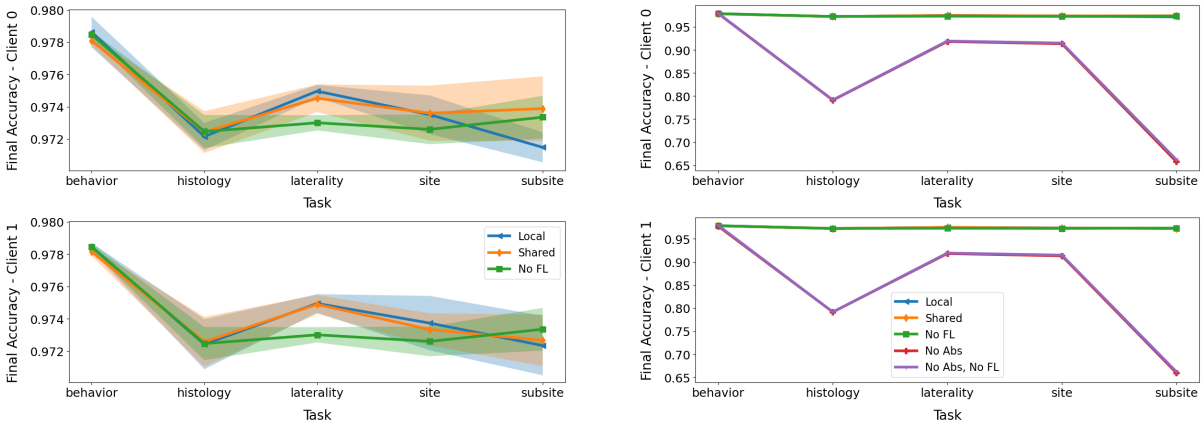

Figure 5: Accuracy at the end of training for the compared modalities. Plots correspond to the mean on 5 realizations and shaded region shows the standard deviation. **Left:** DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation and DAC without FL. **Right:** includes also the plots for FL without abstention and for baseline without abstention or FL.

3.3 Heterogeneous Dataset Experiments

Having confirme that homogeneous FL training essentially converges to the consolidated training result, we now address the more meaningful case of heterogeneous FL training. For these tests we use a combined dataset of Louisiana (LA) and Kentucky (KY). For the FL training, Client 0 uses the LA data and Client 1 the KY data. The KY dataset is comparable in size to LA (537,647 total) with a similar partitioning into train, validation and test sets.

We perform the same experiments as in the previous section, to confir that the 'local' and 'shared' aggregation modalities for the α coefficient converge similarly and are comparable to the results for the consolidated training on the combined LAKY dataset.

Figure 7 shows the distribution of fina α values for each task, across five training runs each for the 'local' and 'shared' FL training modes, and the same dataset trained as a single consolidated corpus. As with the homogeneous training, the values for the histology and subsite tasks are very tightly clustered across all modalities, reflectin the stringency required to obtain the desired accuracy bounds. Conversely, the values for behavior are quite disparate, since the

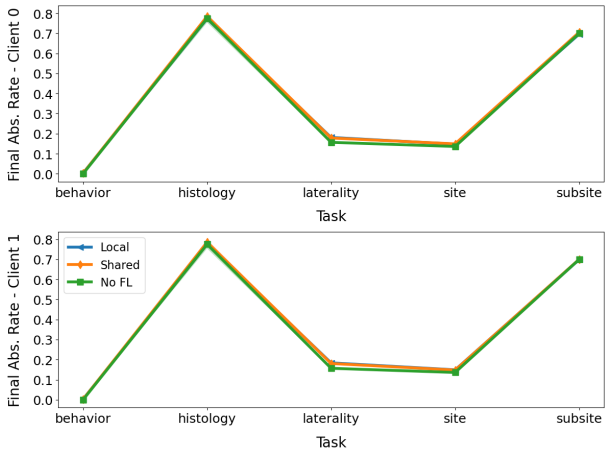

Figure 6: Abstention rate at the end of training for the compared modalities: DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation and DAC without FL.

Task	DAC	DAC + FL Local	DAC + FL Shared
Behavior	0.9794 ± 0.0001	0.9792 ± 0.0009	0.9789 ± 0.0003
		0.9791 ± 0.0004	0.9790 ± 0.0002
Histology	0.9750 ± 0.0015	0.9736 ± 0.0011	0.9751 ± 0.0007
		0.9739 ± 0.0019	0.9754 ± 0.0013
Laterality	0.9737 ± 0.0003	0.9756 ± 0.0006	0.9756 ± 0.0006
		0.9759 ± 0.0007	0.9758 ± 0.0006
Site	0.9756 ± 0.0009	0.9763 ± 0.0008	0.9764 ± 0.0014
		0.9766 ± 0.0015	0.9762 ± 0.0007
Subsite	0.9766 ± 0.0011	0.9738 ± 0.0013	0.9757 ± 0.0023
		0.9746 ± 0.0020	0.9744 ± 0.0017

Table 2: Mean accuracy $\pm$ standard deviation in test set per client and per task over the trained models.

accuracy target can be met with minimal or zero abstention (see Figures 8 and 9).

There is a visible difference in the fina α values for laterality and site between client 0 and client 1, consistent across all training runs, indicating that the disparate silos require a slightly different tuning to achieve the same nominal targets. This is also visible in Figure 9 where clear differences in abstention rate are visible between clients. Notably the abstention rates for the FL case are higher, due to both noise introduced in the FL workfl w and the need to satisfy accuracy targets on independent subsets of the data.

Overall, we demonstrate that the FL trained models converge to the consolidated training case, albeit with some slight variation between clients, primarily due to the data disparities and consequent divergence in the fina parameters.

3.4 Anomaly Detection

A previously observed feature of the DAC is the ability to signal the prevalence of anomalous samples through changes in the abstention rate. As evidenced in the heterogeneous training example, the actual abstention rate for nominally identical models on datasets with differing composition will vary. In this section we briefl demonstrate the variability of abstention rate in

Task	DAC	DAC + FL Local	DAC + FL Shared
Behavior	0.0003 ± 0.0002	0.0033 ± 0.0016	0.0027 ± 0.0013
		0.0029 ± 0.0014	0.0025 ± 0.0010
Histology	0.7732 ± 0.0202	0.7708 ± 0.0109	0.7822 ± 0.0125
		0.7728 ± 0.0105	0.7825 ± 0.0143
Laterality	0.1567 ± 0.0041	0.1808 ± 0.0074	0.1782 ± 0.0064
		0.1833 ± 0.0075	0.1803 ± 0.0057
Site	0.1329 ± 0.0036	0.1454 ± 0.0077	0.1464 ± 0.0060
		0.1470 ± 0.0081	0.1452 ± 0.0043
Subsite	0.6975 ± 0.0026	0.6949 ± 0.0059	0.7036 ± 0.0087
		0.6971 ± 0.0059	0.6984 ± 0.0065

Table 3: Mean abstention rate $\pm$ standard deviation in test set per client and per task over the trained models.

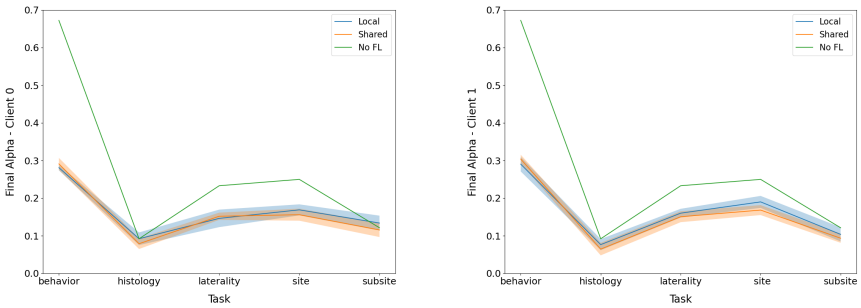

Figure 7: Values of alphas at the end of heterogeneous training for the compared modalities: DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation, and consolidated training. Left is client 0 (LA), right is client 1 (KY). Alpha values converge to similar values across multiple tests, with consistent differences between client 0 and client 1.

two scenarios, cross registry testing and variations over time. This is preliminary study which will be more fully explored in upcoming work.

3.4.1 Cross-Registry Performance

For this test, we use the five models from the heterogeneous training, trained on LA+KY, and test the performance on the 'test' partition of five separate registries. In addition to the LA and KY datasets, we include registries from New Mexico (NM), Seattle (SE) and Utah (UT). These registries differ substantially in total size, as well potentially having different compositions compared to the LA, KY registries, including differing class imbalance, quality of annotation (ground truth) and a variety of other factors. The full analysis of the variations between registries and the reasons for abstention are beyond the scope of this paper.

Figure 10 shows the summary of the accuracy and abstention variation across tasks for each of the registries. For accuracy, the 97% is maintained or exceeded for all tasks except behavior, since that task has effectively no abstention pattern learned and simply behaves as a non-abstainingclassifie in the presence of noise. The variation in abstention is obscured due to the large range of values in the plot so we display an alternate view in Figure 11, showing the variation of abstention across registries for each task. We exclude behavior since this remains

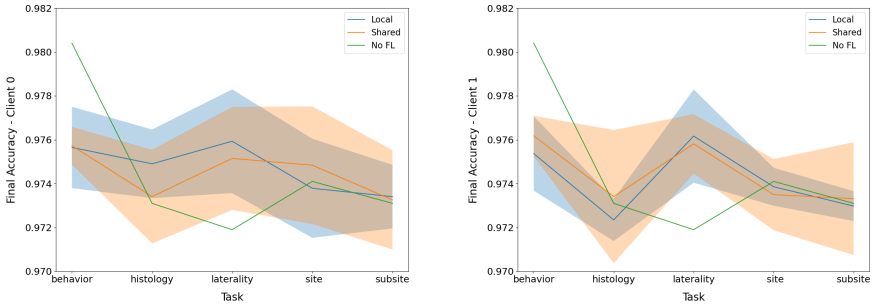

Figure 8: Accuracy at the end of training for the compared modalities: DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation and DAC without FL. Left is client 0 (LA), right is client 1 (KY)

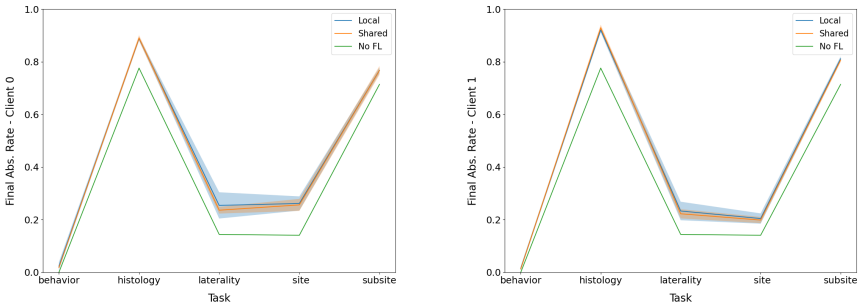

Figure 9: Abstention rate at the end of training for the compared modalities: DAC+FL with local alpha aggregation, DAC+FL with shared alpha aggregation and DAC without FL. Left is client 0 (LA), right is client 1 (KY)

effectively zero, as discussed above. Abstention rates vary across registries, but the trend is different for each task, reflectin the complexity of the dynamics and differences in number of classes, noise level and base abstention rates for each. Nevertheless, the ability of the DAC to signal variations in quality and the potential for abstention rates to vary, both positively and negatively, is demonstrated.

3.4.2 Time evolution inference tests

An alternate demonstration of the anomaly detection feature is to test the ability to discern changes in data over time. For this, we train the models on LA,KY data in an FL setting, but only using data up to and including the 2012 calendar year. The remaining LA,KY data, from 2013 to 2022, is partitioned into yearly increments and only the 'test' partition from each year is used.

Figures 12 and 13 show the abstention and accuracy performance for each client. Note that, as before, the trend for each task differs over time, but some consistent patterns emerge. Both clients display similar differences between LA and KY test data, despite being trained primarily on one or the other. For example, there is a clear separation between subsite accuracy rates, with LA (blue) always being more accurate than KY (orange), indicating a systematic difference in data quality between them. In contrast, the abstention rates for the same task vary more and converge over time.

The patterns learned by the DAC can be complex, and require an understanding of the details

Method	Repetition 1	Repetition 2	Repetition 3	Repetition 4	Repetition 5
DAC + FL Local	19, 1, 1, 1	21, 2, 6, 1	22, 9, 1, 2	24, 2, 2, 2	25, 9, 1, 1
	17, 2, 1, 1	17, 4, 1, 1	24, 2, 1, 1	16, 3, 3, 6	17, 14, 1, 1
DAC + FL Shared	25, 8, 2, 2	25, 1, 1, 1	23, 2, 2, 2	25, 10, 1, 3	23, 8, 2, 2
	21, 7, 2, 4	21, 2, 1, 7	17, 3, 1, 1	17, 3, 6, 2	17, 2, 7, 1

Table 4: Number of epochs in heterogeneous training per repetition. For each modality, top row is client 0 (LA) and bottom is client 1 (KY), epochs are reported by round. Note that in heterogeneous training the last round may require multiple epochs to converge, so that the fina model state, including both α and model weights, may be slightly different for each silo.

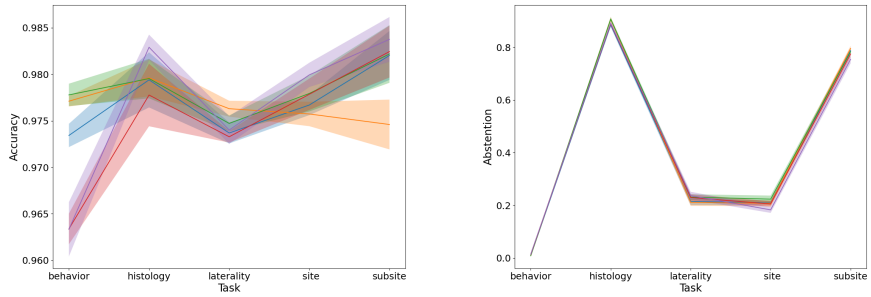

Figure 10: Performance summary, abstention and accuracy performance across registries. The models are trained on LA+KY data and tested on separate 'test' partitions from LA, KY, NM, SE, UT registries.

of the registry data. Over time, the cancer coding standards undergo periodic updates, and the quality of diagnostic tools may improve. Additional changes may be due to changes in population demographics within the catchment regions for the registries, due to a variety of factors. Again, the exact reasons for the variation are beyond the scope of this paper, where we only seek to demonstrate the potential for the DAC to signal (and separate, via the abstention flag the presence of anomalous samples in the data.

4 CONCLUSION

We have demonstrated the ability to implement the accuracy-targeted DAC in a FL setting, and tested two different aggregation modes for the α coefficien which are an essential component of the models. The resulting FL trained models are shown to converge to the model obtained via consolidated training, with small differences due to the noise introduced in the FL process (to preserve privacy) and the need to enforce the accuracy target on independent subsets of the data. We then demonstrate the ability of these models to signal changes in data quality, via testing on data from alternate registries and from variations over time from the same registry. This provides a jumping off point for further work to characterize and improve the sensitivity of the DAC as an anomaly detecter for deployment in real-world distributed data collection environments.

5 Acknowledgments

The authors would like to thank Chris Stanley, John Gounley and Heidi Hanson from ORNL for their guidance on federated learning, the Flower Framework in particular, assistance with

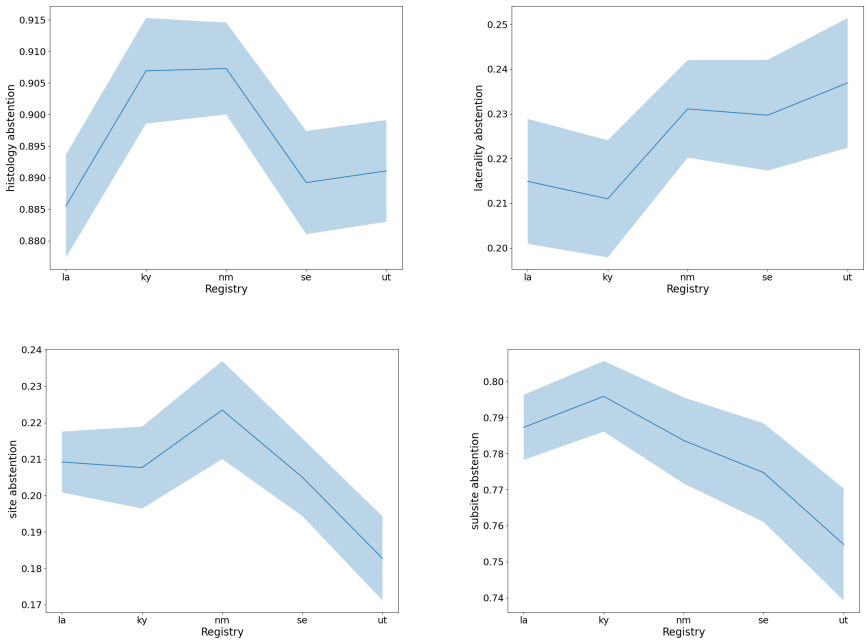

Figure 11: Abstention variation across registries for FL models trained on LA+KY. The behavior task is excluded since abstention remains effectively zero and the model has not learned an abstention pattern for this task.

data preparation and for valuable discussions.

This material is based upon work supported by the U.S. Department of Energy, Offic of Science, Offic of Advanced Scientifi Computing under Award Number DE-SC-ERKJ422.

The authors would like to acknowledge the contribution to this study from other staff in the participating central cancer registries. These registries are supported by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program, the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR), and/or state agencies, universities, and cancer centers.

The participating central cancer registries include the following:

- Fred Hutchinson Cancer Center working under contract numbers SEER: HHSN2612018000041
- Kentucky working under contract numbers SEER: HHSN261201800013I/HHSN26100001 and NPCR: NU58D007144
- Louisiana working under contract numbers SEER: HHSN261201800007I/HHSN26100002 and NPCR: NU58DP0063.
- New Mexico working under contract numbers SEER: HHSN261601800014l
- Utah working under contract numbers SEER: HHSN261201800016I and NPCR: NU58DP007131

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficien learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, and J. Gardner, "Advances and open problems in federated learning," in *Foundations and Trends in Machine Learning*, vol. 14, 2021.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, p. 1–19, 2019.

[4] C. Dwork, A. Roth, and et al., "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211 – 407, 2014.

[5] P. Bellavista, L. Foschini, and A. Mora, "Decentralised learning in federated deployment environments: A system-level survey," in *ACM Computing Surveys*, vol. 54, 2021.

[6] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, pp. 1 – 12, 2020.

[7] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickker, and A. N. Ahmad, "Swarm learning for decentralized and confidentia clinical machine learning," *Nature*, vol. 594, no. 7862, pp. 265–270, 2021.

[8] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, and M.-M. Steinborn, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473 – 484, 2021.

[9] S. Thulasidasan, T. Bhattacharya, J. Bilmes, G. Chennupati, and J. Mohd-Yusof, "Combating label noise in deep learning using abstention," in *Proceedings International Conference on Machine Learning*, 2019.

[10] Flower-Labs-GmbH, "Flower framework." [Online]. Available: https://flwer.ai/docs/framework/index.html

[11] "National Cancer Institute - Survellillance, Epidemiology, and End Results Program," https://seer.cancer.gov/.

[12] M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, and G. Tourassi, "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 89–98, 11 2019. [Online]. Available: https://doi.org/10.1093/jamia/ocz153

[13] S. Gao, M. Young, J. Qiu, H.-J. Yoon, J. Christian, P. Fearn, G. Tourassi, and A. Ramanthan, "Hierarchical attention networks for information extraction from cancer pathology reports," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 321–330, 2018.

[14] S. Dhaubhadel, J. Mohd-Yusof, K. Ganguly, G. Chennupati, S. Thulasidasan, N. W. Hengartner, B. J. Mumphrey, E. B. Durbin, J. A. Doherty, M. Lemieux, N. Schaefferkoetter, G. Tourassi, L. Coyle, L. Penberthy, B. H. McMahon, and T. Bhattacharya, "Why I'm not answering: Understanding determinants of classificatio of an abstaining classifie for cancer pathology reports," *Arxiv*, 2022.

[15] A. Peluso, I. Danciu, H.-J. Yoon, J. Mohd-Yusof, T. Bhattacharya, A. Spannaus, N. Schaefferkoetter, E. Durbin, X.-C. Wu, A. Stroup, J. Doherty, S. Schwartz, C. Wiggins, L. Coyle, L. Penberthy, G. Tourassi, and S. Gao, "Deep learning uncertainty quantificatio for clinical text classification " *Journal of Biomedical Informatics*, vol. 149, p. 104576, 2024.

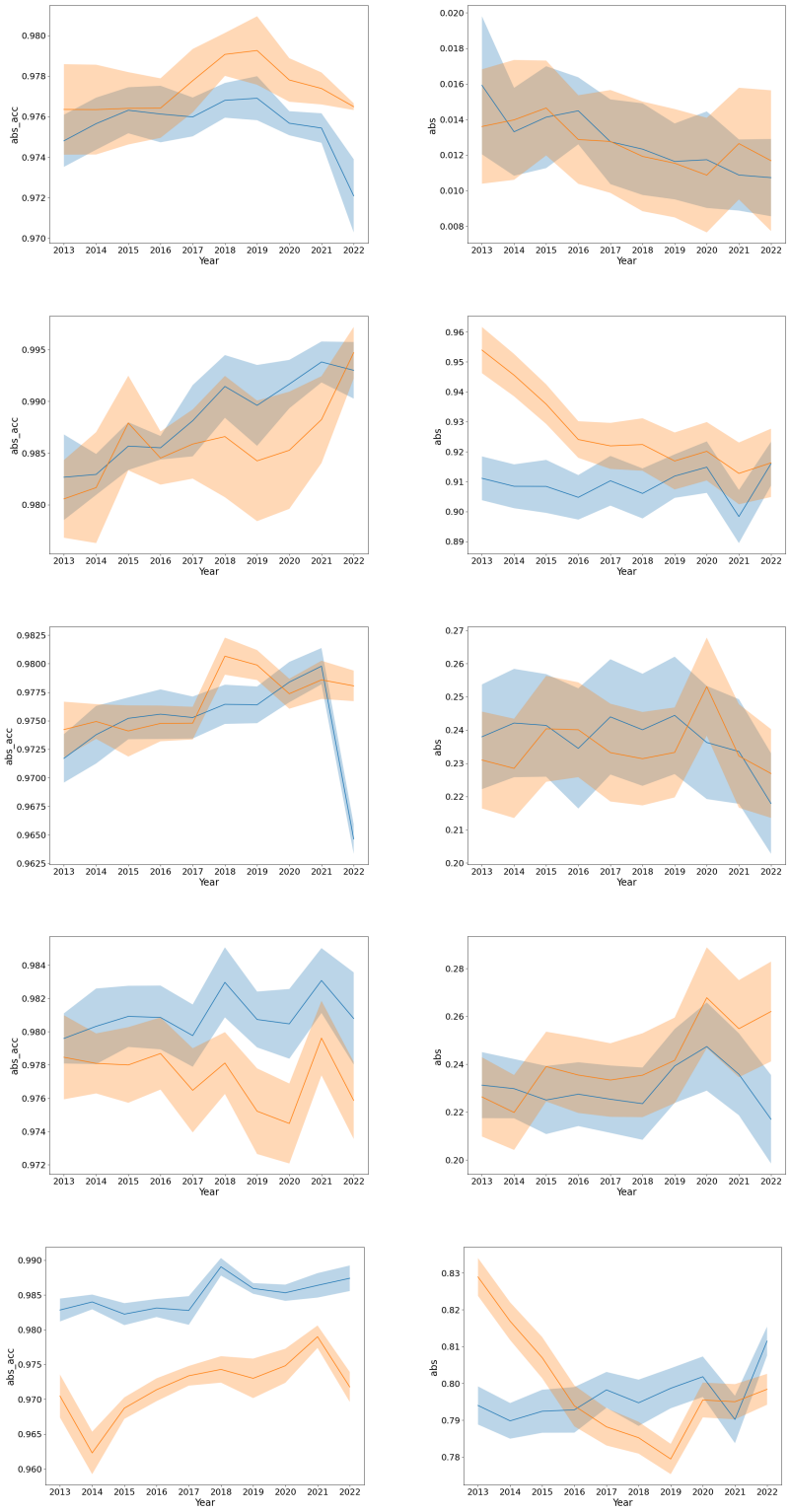

Figure 12: Inference performance of a DAC trained on pre-2012 data, tested on individual years from 2013 to 2022. Client 0 (trained on LA data in a heterogeneous LA+KY FL setting), tested on LA and KY data from 2013-2022. Tasks are (top to bottom) behavior, histology, laterality, site, subsite. Left column is accuracy, right column is abstention rate.

Figure 13: Inference performance of a DAC trained on pre-2012 data, tested on individual years from 2013 to 2022. Client 1 (trained on KY data in a heterogeneous LA+KY FL setting), tested on LA and KY data from 2013-2022. Tasks are (top to bottom) behavior, histology, laterality, site, subsite. Left column is accuracy, right column is abstention rate.