Understanding Clinical Collaborations Through Federated Classifier Selection

Sebastian Caldas[1] SCALDAS@CMU.EDU
Joo Heung Yoon[2] YOONJH@PITT.EDU
Michael R. Pinsky[2] PINSKY@PITT.EDU
Gilles Clermont[2] CLER@PITT.EDU
Artur Dubrawski[1] AWD@CS.CMU.EDU

[1] *Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
[2] *School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA*

Abstract

Deriving true clinical utility from models trained on multiple hospitals' data is a key challenge in the adoption of Federated Learning (FL) systems in support of clinical collaborations. When utility is equated to predictive power, population heterogeneity between centers becomes a key bottleneck in training performant models. Nevertheless, there are other aspects to clinical utility that have frequently been overlooked in this context. Among them, we argue for the importance of understanding how a collaboration may be affecting the quality of a center's predictions. Insights into how and when external knowledge is being useful can lead to strategic decisions by stakeholders, such as better allocation of local resources or even identifying best practices outside of the current organization. We take a step towards deriving such utility through FedeRated CLassifier Selection (FRCLS, pronounced "freckles"): an algorithm that reuses classifiers trained in outside institutions. It identifies regions of the feature space where the collaborators' models will outperform the local center's classifier, and can provide interpretable rules to describe these regions of beneficial expertise. We apply FRCLS to a sepsis prediction task in two different hospital systems, demonstrating its benefits in terms of understanding the types of patients for which the collaboration is useful and reasoning about the strategic decisions that may stem out of these analyses.

1. Introduction

When training Machine Learning (ML) models for healthcare applications, previous studies have shown the advantages of using multiple centers' data. This strategy augments the sample size available for data-intensive models, increases the availability of rare and new events, and potentially enhances the generalizability of model predictions (Lee et al., 2012; Wiens et al., 2014; Sheller et al., 2018; Curth et al., 2019; Li et al., 2019). This collaborative approach, however, faces several obstacles that can prevent it from delivering true clinical utility:

- **Limits on data sharing**: Hospitals are responsible for protecting their patients' privacy. As such, legal constraints, organizational policies, and ethical barriers often impede centers from sharing patient-level data (Van Panhuis et al., 2014).
- **Population heterogeneity**: Data in different centers is inherently heterogeneous. They collect data in different ways, have different laboratory procedures, and have varying care styles and organizational cultures. Of most interest to us, different hospitals serve different populations. As a consequence of this natural population diversity, data from different centers is not identically distributed. However, sharing intelligence across centers, in spite of these misalignments, is a worthwhile effort that can bring improvements in quality of care and reductions in its cost (Lee et al., 2012; Curth et al., 2019). Naive collaborative models will fail to recognize and leverage this variation.
- **Overemphasis on Predictive Power**: ML engineering has traditionally focused on optimizing predictive power, measured through some metric such as empirical accuracy. Nonetheless, there are other aspects of clinical utility that are often neglected. In the context of clinical collaborative strategies, we are interested in explaining how the collaboration itself is affecting a center's predictions, e.g., whether a decision is being made based on knowledge from an external center. Concrete rationale of this type can incentivize further cooperation, identify local bottlenecks and inform local resource allocation, or even help identify external best practices.

In this work, we introduce Federated CLassifier Selection, or FRCLS (pronounced "freckles"), a Federated Learning (FL) classification algorithm that tackles all three obstacles outlined above. Similarly to other FL algorithms, FRCLS overcomes the data sharing obstacle by training distributed models through the exchange of their associated parameters instead of the data they are being trained on (Kairouz et al., 2019; Li et al., 2020a; Rieke et al., 2020). However, unlike other FL algorithms that train a single global model for all participating collaborators (McMahan et al., 2017; Li et al., 2020b; Karimireddy et al., 2020), FRCLS is designed to adapt application of models to the data distribution of a particular hospital. In addition, FRCLS is able to identify groups of patients for which the collaboration is particularly useful.

FRCLS is driven by the intuition that inter-center population heterogeneity makes each hospital an expert on different patient subpopulations, just as we illustrate in Figure 1. It follows that each center could be an expert in a different region of the feature space. FRCLS leverages this diversity of competence among classifiers and dynamically picks the model that is best for each incoming instance. Other works on FL for healthcare that recognize the challenge of heterogeneity of data address it through techniques such as domain adaptation (Curth et al., 2019; Andreux et al., 2020a) and clustering (Huang et al., 2019). However, they all suffer from the third obstacle above: clinicians cannot judge the utility of the collaboration itself beyond the predictive performance of the resulting models.

FRCLS addresses this last obstacle by explicitly recognizing when it is leveraging knowledge from an external center. It can also produce rules that clearly delineate the regions of the feature space where external centers are more competent than the local center, providing an interpretable rationale for decisions made by local stakeholders. By optimizing for both accuracy and interpretability simultaneously, and by targeting a deeper understanding of the collaborative predictions, we hope to optimize for our actual goal: clinical utility.

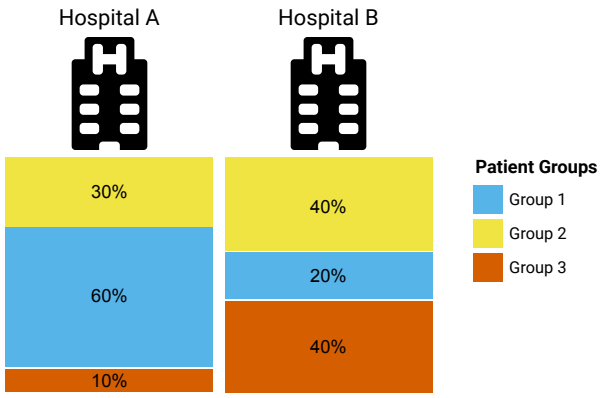

Figure 1: Illustration of inter-center population heterogeneity. Our hypothesis is that a model trained on the data from Hospital A will be an expert on patients from Group 3, while a model trained on Hospital B may underperform for that group. It will thus be beneficial to use model A on Group 3 in Hospital B.

We demonstrate the effectiveness of FRCLS on an early sepsis prediction task (Reyna et al., 2019), where we train and exchange classifiers between two different hospital systems. We show how each system yields benefits from selectively using the external classifier along two axes: in terms of a significant number of predictions flipped from incorrect, when using the local classifier, to correct; and in terms of understanding for which types of patients these flips happen.

Generalizable Insights about Machine Learning in the Context of Healthcare

Recent advances in Federated Learning (FL) are a promising step towards identifying the strengths of collaborations between clinical centers, allowing models to be built while data remain siloed. Nevertheless, before FL can be deployed successfully in clinical collaborations, we must develop algorithms that (i) are robust to population heterogeneity, and (ii) allow clinical centers to understand when they are making use of external knowledge, providing them with utility beyond an increase in predictive power. To address these issues, we propose FRCLS, an algorithm that selectively exploits population heterogeneity, finds regions of the feature space where models trained at external centers can outperform a local model, and describes these regions of expertise through simple rules. Thus, FRCLS characterizes types of patients where it is advantageous to use external knowledge, codified in the form of an external machine learning model.

2. Related Work

We review relevant work in three key directions: (i) Dataset shifts in healthcare, which is a crucial motivation for our work, (ii) Federated Learning (FL), to situate our work in the broader context of this field, and (iii) Dynamic Classification, the underlying technique behind FRCLS.

Dataset Shifts in Healthcare. At a high level, dataset shift refers to a scenario in which a machine learning model is tested on data drawn from a different distribution than the one it was trained on (Subbaswamy et al., 2021). Usual performance guarantees don't apply to this setting, as they refer to how the model will perform on data drawn from the training distribution. Dataset shift may cause the deterioration of clinical prediction models when they are validated on an external distribution and when a collaboratively trained model is applied to individual centers (Lee et al., 2012; Zech et al., 2018; AlBadawy et al., 2018; McKinney et al., 2020). Different works have tried to prevent this degradation by using techniques such as domain adaptation (Curth et al., 2019) and transfer learning (McKinney et al., 2020; Mustafa et al., 2021). FRCLS exploits the model heterogeneity that such a shift causes through its dynamic classifier approach.

On the clinical side, recent efforts to offset dataset shifts include developing and implementing common data models across collaborating institutions. Nevertheless, these are only partially successful, as other considerations, such as local workflows and treatment protocols, are also contributing factors (Matheny et al., 2019).

Federated Learning. Traditional FL algorithms coordinate the training of machine learning models through a central server, which iteratively broadcasts the current model to participating collaborators and aggregates the updates it receives in return (Li et al., 2020a; Kairouz et al., 2019). Different algorithms vary in terms of how they perform the local updates and the central aggregation (McMahan et al., 2017; Li et al., 2020b; Karimireddy et al., 2020). Previous work has shown the feasibility of using federated techniques of this kind on healthcare problems (Sheller et al., 2018; Li et al., 2019; Andreux et al., 2020b,a; Caldas et al., 2020), others have focused particularly on FL's fairness challenges (Li et al., 2020c; Mohri et al., 2019). Algorithmically, FRCLS's approach is different in one crucial way: it performs one single exchange between all collaborators instead of several exchanges with a central server.

Dynamic Classification. Our method is related to the Dynamic Classifier Selection (DCS) and the Regression-based Informative Projection Recovery (RIPR) frameworks (Cruz et al., 2018; Fiterau and Dubrawski, 2012), both of which make use of a heterogeneous pool of classifiers and serve a different model for each test instance. Both frameworks select the classifier to be served by estimating the competence of each candidate model on the region of the feature space where the new instance resides. DCS methods estimate the models' performance on the instance's k-nearest neighbors, while RIPR derives a local entropy measure. FRCLS takes a similar approach to the one used by DCS methods.

3. Method

In this section, we first present the proposed algorithm, Federated CLassifier Selection or FRCLS (pronounced as "freckles") in Section 3.1, before detailing its dynamic classifier component in Section 3.2.

3.1. Federated Classifier Selection

At a high level, FRCLS proceeds in three stages: the local training of classifiers, the exchange of fully trained classifiers between centers, and the dynamic selection of classifiers

in each center. We first give a high-level overview of each stage before further detailing the third one, which is the focus of our contributions.

1. **Training of local classifiers**: Each clinical center can independently choose its own type of model, hyperparameter tuning strategy, etc.
2. **Exchange of classifiers**: In this stage, clinical centers exchange both their fully trained classifiers and the local imputation/standardization parameters used during training. We assume an honest-but-curious threat model and thus consider it safe to share this information among clinical centers. A central authority may also anonymize the exact source of the external classifiers. In the end, each hospital is left with a local classifier c_L and a pool of candidate external classifiers $C = \{c_1, \ldots, c_M\}$.
3. **Dynamic selection of candidate classifiers**: This stage takes place at each center independently. The ultimate goal is, for each new incoming instance, to select the most competent classifier among all candidates. We explain the details on how FRCLS does this in Section 3.2.

3.2. Dynamic Selection of Candidate Classifiers

We are given a local classifier c_L and a set of M external classifiers $C = \{c_m\}_{m=1}^{M}$. Our objective is to determine, for each new instance x, whether to use c_L or one of the elements of C. To make this decision, we set out to quantify the utility of the external classifiers in C relative to c_L. Define

$$L_c(x,k) = \frac{1}{k} \sum_{j \in nn(x,k)} \ell(c(x_j), y_j),$$
$$\rho_m(x,k) = \frac{L_{c_L}(x,k)}{L_{c_m}(x,k)},$$

where $nn(x,k)$ returns indices of the k-nearest neighbors of x, ℓ is the cross-entropy loss, and $c(x)$ is the score that classifier c assigns to x.

Notice that $L_c(x,k)$ estimates classifier c's competence on a given point x by averaging c's loss on the point's k-nearest *known* neighbors. Because it looks at the classifier's loss, L_c is inversely related to c's competence. Meanwhile, $\rho_m(x,k)$ takes the ratio of L_c for the local c_L and an external c_m. Because of L_c's inverse relation to competence, a higher value of ρ_m translates into a higher competence for the external classifier c_m.[1]

We now construct the greedy external classifier c_E which solves

$$c_E(x) = \arg\max_{c_m \in C} c_m(x)$$

for each new instance x. The final step is to pick between c_L and c_E. A naive strategy would choose c_E whenever $\rho_E(x,k) > 1$. However, this strategy won't be necessarily optimal, as ρ_E quantifies relative competence in terms of loss, which is just a proxy for actual clinical utility. Instead, we propose two data-driven strategies that we illustrate in Figure 2.

1. For computational stability, the quantity we use in our experiments is $\rho'_m(x,k) = \log \frac{L_{c_L}(x,k)+\epsilon}{L_{c_m}(x,k)+\epsilon}$ for some small ϵ.

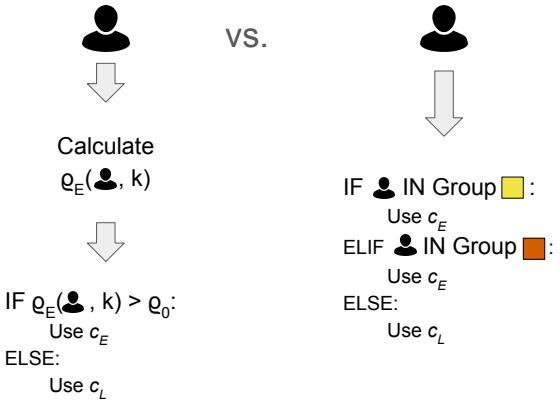

Figure 2: Illustration of FRCLS's strategies to select between the local and the external classifiers once a new instance arrives: competence threshold and decision list. By construction, the decision list strategy is interpretable, which adds to its clinical utility.

3.2.1. Competence Threshold

Our first strategy finds a threshold ρ_0 such that c_E will be used whenever $\rho_E(x, k) > \rho_0$. We optimize this threshold by minimizing the p-value of a statistical test whose null hypothesis states that c_L's utility is greater than c_E's, where we measure utility in terms of a classifier's correct predictions.

More precisely, we define

$$F(\rho_0) = |\{x_i : \rho_E(x_i, k) > \rho_0, c_L^*(x_i) \neq c_E^*(x_i), c_E^*(x_i) \neq y_i\}|,$$
$$S(\rho_0) = |\{x_i : \rho_E(x_i, k) > \rho_0, c_L^*(x_i) \neq c_E^*(x_i).c_E^*(x_i) = y_i\}|$$

where $c^*(x)$ is the label that classifier c assigns to x. Notice that S is the number of instances where using c_E actually changes the prediction made by c_L, and the new prediction is correct. Meanwhile F is the number of instances that changed predictions to an incorrect one.

Having these quantities, we perform a one-tailed binomial test to check for the statistical significance of $\frac{S(\rho_0)}{S(\rho_0)+F(\rho_0)} < 0.5$. In our experiments, we use a simple grid search strategy to look for the threshold ρ_0 that lets us reject this null hypothesis with the most confidence.

3.2.2. Decision Lists

Our second strategy uses c_E if the instance satisfies a set of interpretable rules, and otherwise defaults to c_L. To build these rules, FRCLS uses a rule learning algorithm to create a decision list that maximizes a lower bound on the mean of ρ_E. Then, we iterate over the list and choose the rule that minimizes the p-value of our binomial test when applied to the instances selected by all rules so far on the list. This will be the last rule that prescribes the use of c_E. Our approach is agnostic to the implementations of the rule learning algorithm. The one we use in our experiments is the one proposed by Moore and Schneider (2002).

Limitations

After we find the threshold or rule that maximizes c_E's utility relative to c_L's, the optimal p-value of our binomial test may still be higher than a satisfying confidence level. In this scenario, both of FRCLS strategies would default to the local model for all instances and there would be no gains in predictive power due to the collaboration. However, the knowledge that the external models do not outperform the local one for any patient subpopulation is still a valuable insight into the limited utility of the collaboration. Additionally, our k-nearest neighbors estimator for L_c is known to suffer from the curse of dimensionality. We also need to tune an extra hyperparameter for it: the number of neighbors or k. We propose a heuristic to perform this tuning in Appendix B.

4. Results and Discussion

In this section, we first present the details of our experimental setup (Section 4.1), before presenting and discussing the results of the local classifiers (Section 4.2), and FRCLS's competence threshold and decision list strategies (Sections 4.3 and 4.4, respectively).

4.1. Experimental Setup

We demonstrate our method on an early sepsis prediction task. Sepsis is linked with high mortality, morbidity, and cost of care in hospitalized patients. To mitigate this burden, early identification of risk for sepsis and timely treatment are recommended (Angus et al., 2001). It follows that systems for early and accurate identification of sepsis are of crucial interest for the community (Nemati et al., 2018; Reyna et al., 2019).

Data Source. We use the data shared by Reyna et al. (2019) as part of the 2019 PhysioNet/Computing in Cardiology Challenge. The released data corresponds to ICUs in two geographically distinct hospital systems with different Electronic Medical Record systems. In the rest of the paper, we refer to these as hospital system A and hospital system B, matching the nomenclature used by Reyna et al. (2019). The public data accounts for $40,336$ patients and over 1.5 million instances.

Machine Learning Task. Our machine learning task is to predict sepsis 6 hours before its onset time, according to the definition used by Reyna et al. (2019). Due to the nature of this task, the label distribution is skewed: only 1.80% of the given labels correspond to the positive class. To facilitate the predictive task and to focus our study on its federated aspects, we randomly undersample the negative labels in order to match the number of negative and positive labels. We are left with 55.8 thousand instances from $20,779$ patients. Table 5 shows the number of instances per hospital system.

Feature Choices. We use the features provided by the 2019 PhysioNet Challenge. These consist of a mixture of hourly vital signs, laboratory values, and patient descriptors. Table 6 and Table 7 describe the numerical and categorical features provided, respectively. We also use the standard deviations of the mean arterial pressure and the respiration rate (Nemati et al., 2018).

Data Splits. We split each hospital system's data into three disjoint sets. First, a training set used for training and tuning the local classifier. Second, a validation set used to either find the optimal ρ_0 or to train FRCLS's decision list. This set is used to measure

L_c for all candidate classifiers, i.e., the operator $nn(x, k)$ is restricted to returning instances from this set. Third, we have a test set for estimating FRCLS's performance. We perform the split in a 40/30/30 fashion.

Predictive Models. Our local classifiers are logistic regression models with ridge penalty. We tune the regularization hyperparameter using 5-way cross-validation to optimize for loss. We locally impute missing values, using the mean for numerical features and the mode for categorical ones. Finally, we locally standardize numerical features.

These models are not state-of-the-art for a sepsis prediction task, but they are sufficient for evaluating our method, as our purpose is to selectively use c_E to successfully change local predictions and not to achieve the best possible accuracy on the task.

Evaluation Metrics. We are interested in comparing the utility of classifiers c_E and c_L on the instances where FRCLS decides to use c_E over c_L. We measure this relative utility as we did in Section 3: by looking at the instances (x, y) for whom $c_L^*(x) \neq c_E^*(x)$. We refer to these predictions as *flipped*, and we consider them as *successful flips* if $c_E^*(x) = y$. If the number of successful flips represents more than 50% of the total number of flips, then we consider the method successful.

Our metric requires crisp predictions from the classifiers. As such, we focus on two constraints that optimize for complementary objectives and allow us to obtain crisp classifiers: holding the true positive rate at 90% (@90% TPR), and holding the false positive rate at 10% (@10% FPR). Given our sepsis prediction task, a center would likely prefer guaranteeing a high recall (@90% TPR), but we show both constraints for the sake of completeness.

Finally, this metric only measures c_E's performance on those instances in which c_E's predictions differ from c_L's. To quantify c_E's performance both when it agrees and disagrees with c_L, we also measure the accuracy of the crisp classifiers on all instances where FRCLS uses c_E over c_L.

4.2. Results of Local Classifiers

We show the performance of our sepsis prediction models trained independently in Hospitals Systems A and B in Figure 3. We plot the Receiver Operating Characteristic (ROC) curve for both models when evaluated in the test data of each hospital system. In both cases, c_L either outperforms or matches the performance of c_E, a consistent behaviour throughout the curve. Judging by these results, both hospital systems may have deemed c_E's utility as limited. However, in Section 4.3 and Section 4.4, we'll show that, by comparing the models' local behaviour, these hospital systems do yield utility from their external models c_E, doing so selectively on a subset of their instances.

4.3. Results of Competence Threshold Strategy

Table 1 presents the results for our competence threshold strategy. For three out of the four scenarios we consider, we obtained an optimal p-value lower than 0.05 in both our validation and test sets, a confidence level we consider satisfying for our experiments. These results speak to FRCLS's generalization ability. We note that, in the one scenario in which FRCLS did not generalize (Hospital System A @90% TPR), the validation set p-value was several orders of magnitude higher than for the other scenarios.

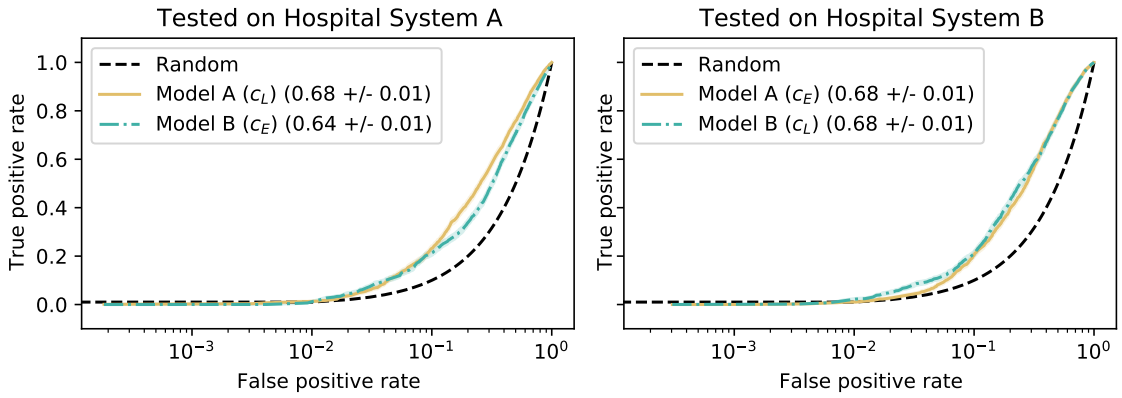

Figure 3: AUC ROC for the classifiers when tested on each hospital system. Confidence bands reflect Wilson scores. We plot the false positive rate in logarithmic scale for better visibility of models' performance at clinically relevant low error settings.

Table 1: Results for our competence threshold strategy for one run. All results are measured on the test set unless specified. In bold, p-values higher than 0.05.

Hospital System	**Val p-value**	**p-value**	**Instances handled by c_E**	**Successful Flips (% of flips)**	**Local Accuracy**	**External Accuracy**
A (@90% TPR)	1.69e−3	**9.99e−1**	1058	26 (32.91%)	63.71%	61.15%
A (@10% FPR)	3.21e−42	1.52e−9	1525	280 (64.22%)	53.64%	61.77%
B (@90% TPR)	7.26e−31	9.04e−11	1243	195 (68.90%)	49.64%	58.25%
B (@10% FPR)	1.51e−7	1.90e−2	2548	128 (57.14%)	57.50%	58.75%

Given our data splits, our test sets consist of over 9.5 and 7.3 thousand instances for hospital systems A and B, respectively. With this strategy, FRCLS ends up using c_E on 15−35% of the data. Out of this data, 5−18% correspond to successful flips. This translates into the hospital systems reaping real utility out of c_E in most cases, as evidenced in the p-values of our binomial test. Finally, in all cases in which we observe utility in terms of successful flips, we also observe utility in terms of an increase in accuracy.

4.4. Results of Decision List Strategy

To generate our decision lists, we use an implementation of the computationally efficient algorithm proposed by Moore and Schneider (2002). We limit each rule to have a maximum of two features in order to make their interpretation easy, and restrict the minimum support of each rule to at least 1.5% of the validation sample size in each hospital system, to safeguard against overfitting. These hyperparameter settings work well in the presented examples, but they may need to be optimized for other applications.

Table 2: Results for our decision list strategy. Notice that, when *Val p-value* is greater than 0.05 (bolded), no instances are handled by c_E.

Hospital System	Val p-value	p-value	Instances handled by c_E	Successful Flips (% of flips)	Local Accuracy	External Accuracy
A (@90% TPR)	**2.52e−1**	-	0	-	-	-
A (@10% FPR)	1.12e−6	3.07e−5	1925	299 (58.97%)	57.92%	62.65%
B (@90% TPR)	1.35e−3	1.25e−8	700	134 (70.16%)	51.43%	62.43%
B (@10% FPR)	**8.04e−2**	-	0	-	-	-

Hospital System A

```
IF PTT > 84.55 AND Phosphate <= 8.42:
    Use c_E
ELIF BUN > 83.56 AND Calcium <= 9.66:
    Use c_E
ELIF Hct > 40.31:
    Use c_E
ELIF Calcium > 9.66:
    Use c_E
ELIF Hgb > 12.14:
    Use c_E
ELSE:
    Use c_L
```

Hospital System B

```
IF BaseExcess > -0.92:
    Use c_E
ELIF FiO2 > 0.65:
    Use c_E
ELSE:
    Use c_L
```

Figure 4: Rules learned by FRCLS's decision list strategy for our early sepsis prediction task. Instances satisfying these rules will use c_E instead of c_L.

Just as in the previous section, we show our results in Table 2. Meanwhile, in Figure 4, we illustrate the decision lists learned for each of the hospital systems. This time, we encounter two situations in which the optimal p-value in the validation set is lower than 0.05. In these cases, FRCLS defaults to c_L. In the other two situations, Hospital System A @10% FPR and Hospital System B @90% TPR, FRCLS generalizes well, using c_E on $10 - 20\%$ of the test instances, out of which $15 - 19\%$ are successful flips. Compared to our competence threshold strategy, we observe a degradation in our p-values, which are now tens of orders of magnitude greater. This translates into lower ratios of successful flips. This is an expected trade-off, as we can now easily interpret FRCLS's decisions. Conversely, with respect to the previous strategy, the difference in p-values between the validation and test sets has decreased by several orders of magnitude. This is also expected, as the use of a decision list with short rules prevents overfitting.

Finally, we look at the relation between strategies in terms of the instances FRCLS selects to use c_E. We argue that our competence threshold strategy selects two types of

Table 3: Percentage of instances selected by our competence threshold strategy that are successfully explained by our decision list strategy.

Hospital System	**Set**	**% Explained by Rules**
A (@10% FPR)	Val	40.54%
	Test	38.75%
B (@90% TPR)	Val	21.23%
	Test	20.68%

Table 4: Instances whose predictions get flipped with the use of c_E. We present one instance per hospital system. The complete list of features for both instances is shown in Table 8.

Variable	**Patient A**
Age	69
Gender	Male
BUN	13
Calcium	8.6
Hct	38.5
Hgb	13.1
PTT	28.2
Phosphate	2.5
True Label	1

Variable	**Patient B**
Age	54
Gender	Male
Base Excess	-3.2
FiO2	1
True Label	0

instances: some that can be explained with the type of rules we desire, with at most two attributes, and some that are not. Our decision list strategy picks out the former group. In Table 3, we show that these easy-to-explain instances correspond to $20 - 40\%$ of the instances selected through competence thresholding.

4.4.1. Using FRCLS's rules

We turn our attention into demonstrating how clinical centers can use the rules learned by FRCLS's decision learning strategy, and into showing possible ways to derive strategic utility from them.

In Table 4 we show two instances corresponding to two different patients, one from each hospital system. We refer to them as Patient A and Patient B. These are instances for whom the rules of their respective hospital system apply and for whom the use of c_E proves beneficial when holding the appropriate constraints, i.e., @10% FPR for hospital system A and @90% TPR for hospital system B, as shown in Table 2. In the case of Patient A, FRCLS uses c_E because his hemoglobin is greater than the found threshold. For patient B, it's because his fraction of inspired oxygen is higher than 0.65.

We propose a simple argument to explain why c_E does better than c_L for these two cases: the external hospital center sees relatively more instances that satisfy the relevant rules, i.e., Hgb > 12.14 for hospital system A, and FiO2 > 0.65 for hospital system B. To check for this, we construct a one-tailed z-test to compare the proportion of instances that satisfy the rule in the local and external hospital systems. We call these p_L and p_E, respectively. Our null hypothesis states that $p_L - p_E > 0$. For the two rules in question we obtain p-values of 5.60e$-$4 and 4.38e$-$10, meaning we have enough evidence to conclude that $p_E > p_L$ in these cases[2].

We recognize that the utility of c_E for a group of patients may be due to different factors, with sample size being just one of many. However, to derive strategic utility, clinical centers do have to explore why others are doing better. If it is just an issue of differing subpopulation sizes, then strengthening institutional cooperation would make obvious sense. However, it could also be due to issues in data capturing, or even differences in the consistency and quality of the clinical practices themselves. Different conclusions could lead to different decisions by stakeholders.

5. Conclusions

We proposed an approach to derive clinical utility from Federated Learning (FL) systems that goes beyond an increase in predictive power. Compared to previous works in FL for healthcare applications, we argued for a deeper understanding of potential benefits of the clinical collaborations supported by these systems, particularly of when and why external knowledge was affecting local predictions, as this understanding can lead to strategic decisions by stakeholders.

We used a dynamic classification framework to contextually leverage models trained at different clinical institutions, and produced simple rules to clearly outline regions of the feature space where one model outperformed the others. We tested our proposed approach on a benchmark sepsis prediction task in two hospital systems, showing that it was capable of providing both a boost in predictive power and interpretable insights into the types of patients most benefited by the collaboration. Such insights can be used to motivate follow up investigations into specifics of clinical practice that may lead to such differences in model performance. These investigations could help identify the most effective practices for industry-wide proliferation, as well as create awareness of potential inefficiencies of organizational culture or processes that may be addressable at local institutions.

Additional research can improve our work. First, designing a feedback mechanism between collaborators could result in further gains: specializing external models to local needs. This mechanism could be inspired by popular boosting techniques. A more immediate next step is to hybridize both of FRCLS's current strategies, using a competence threshold on those instances not picked out by learned rules. Finally, further exploration into why external models sometimes do better is crucial for making strategic decisions in clinical institutions. An ambitious step in this direction would be the development of frameworks to compare the consistency and quality of practices through model outputs.

2. We note that this z-test shares more information across hospital systems than mere model parameters. However, no raw data is being shared.

Acknowledgments

This work was partially supported by the Defense Advanced Research Projects Agency award FA8750-17-2-0130, and by the National Institutes of Health award R01HL144692.

References

Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

Mathieu Andreux, Jean Ogier du Terrail, Constance Beguier, and Eric W Tramel. Siloed federated learning for multi-centric histopathology datasets. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 129–139. Springer, 2020a.

Mathieu Andreux, Andre Manoel, Romuald Menuet, Charlie Saillard, and Chloé Simpson. Federated survival analysis with discrete-time Cox models. *arXiv preprint arXiv:2006.08997*, 2020b.

Derek C Angus, Walter T Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R Pinsky. Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001.

Sebastian Caldas, Vincent Jeanselme, Gilles Clermont, Michael R. Pinsky, and Artur Dubrawski. A case for federated learning: Enabling and leveraging inter-hospital collaboration. In *American Thoracic Society International Conference*. American Thoracic Society, 2020.

Rafael MO Cruz, Robert Sabourin, and George DC Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018.

Alicia Curth, Patrick Thoral, Wilco van den Wildenberg, Peter Bijlstra, Daan de Bruin, Paul Elbers, and Mattia Fornasa. Transferring clinical prediction models across hospitals and electronic health record systems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 605–621. Springer, 2019.

Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In *Advances in Neural Information Processing Systems*, pages 3023–3031, 2012.

Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *2012 IEEE 12th international conference on data mining workshops*, pages 57–63. IEEE, 2012.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020b.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020c.

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141. Springer, 2019.

Michael Matheny, S Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher. Artificial intelligence in health care: The hope, the hype, the promise, the peril. *NAM Special Publication. Washington, DC: National Academy of Medicine*, page 154, 2019.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577 (7788):89–94, 2020.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Andrew Moore and Jeff Schneider. Real-valued all-dimensions search: Low-overhead rapid searching over subsets of attributes. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 360–369. Morgan Kaufmann Publishers Inc., 2002.

Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Alan Karthikesalingam, Neil Houlsby, and Vivek Natarajan. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical care medicine*, 46(4):547, 2018.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617. JMLR Workshop and Conference Proceedings, 2011.

Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104. Springer, 2018.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021.

Willem G Van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. A systematic review of barriers to data sharing in public health. *BMC public health*, 14(1):1–9, 2014.

Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

Appendix A. Additional Tables

We present additional tables that complement the main body of our paper. Table 5 shows the dataset size for each hospital system. Tables 6 and 7 describe the features in our data. Finally, Table 8 provides a complete description of the instances we use to exemplify how to use FRCLS's rules.

Table 5: Number of instances in each hospital system.

Hospital System	**Number of Instances**
A	$31,253$
B	$24,579$

Appendix B. Tuning the Number of Neighbors

We expand on how to tune the number of neighbors k used to estimate $L_c(x, k)$. Through k we control how useful our estimates are: too high and we loose the local information on which FRCLS relies, too low and we overfit to noise. To avoid these scenarios, we propose a heuristic to tune k on our validation set.

For this tuning, we use $\ell_c(x, y) = \ell_c(c(x), y)$, where ℓ is the cross-entropy loss, as a surrogate for the utility of classifier c. This way, it is easy to differentiate two groups of instances: those for which $\ell_{c_E} < \ell_{c_L}$, and those for which $\ell_{c_E} \geq \ell_{c_L}$. If k is chosen correctly, then we expect the distribution of ρ_E in each one of these groups to not change drastically between validation and test sets. Our heuristic aims to minimize this change but instead uses two disjoint splits of the validation set.

To measure changes in distribution, we use the Rényi divergence (Póczos and Schneider, 2011). For each group of instances defined above, we explore a grid of possible values for k and plot the resulting divergences. Finally, we find the *knee* of each curve and choose the highest k between them. In our experiments we explore the range of values $k = \{2, 7, 15, 50, 100, 150, 500, 1000\}$ and end up using $k = 100$.

Table 6: Numerical features in our data. We show the median (Q1/Q3) for each one of our hospital systems.

Variable	**System A**	**System B**
AST	43.0 (24.0/91.0)	33.0 (21.0/67.0)
Age	65.1 (52.4/75.8)	62.0 (50.0/72.0)
Alkalinephos	78.0 (56.0/116.0)	70.0 (53.0/99.0)
BUN	19.0 (13.0/32.0)	19.0 (12.0/32.0)
Base Excess	0.0 (-2.0/3.0)	-3.5 (-5.7/-0.8)
Bilirubin Total	0.7 (0.4/1.5)	0.9 (0.6/1.5)
Calcium	8.3 (7.8/8.7)	8.3 (7.5/8.8)
Chloride	106.0 (102.0/109.0)	106.0 (103.0/110.0)
Creatinine	0.9 (0.7/1.4)	1.0 (0.8/1.7)
DBP	59.0 (52.0/68.0)	64.0 (55.5/74.0)
FiO2	0.5 (0.4/0.5)	0.4 (0.4/0.6)
Glucose	125.0 (105.0/152.0)	123.0 (104.0/150.0)
HCO3	24.0 (22.0/27.0)	22.1 (20.1/24.7)
HR	87.0 (75.0/100.0)	86.0 (74.0/99.0)
Hct	30.5 (27.7/34.0)	30.8 (26.5/35.8)
Hgb	10.4 (9.3/11.6)	10.0 (8.6/11.7)
Lactate	1.4 (1.1/2.1)	1.6 (1.2/2.3)
MAP	77.0 (68.0/87.0)	83.0 (73.0/96.0)
Magnesium	2.0 (1.8/2.2)	2.0 (1.9/2.3)
O2Sat	98.0 (96.0/99.0)	98.0 (95.0/99.5)
PTT	31.0 (27.0/38.2)	31.6 (28.2/38.1)
PaCO2	40.0 (36.0/45.0)	38.0 (34.0/44.0)
Phosphate	3.3 (2.7/4.1)	3.4 (2.7/4.2)
Platelets	193.0 (137.0/267.0)	180.0 (123.0/248.0)
Potassium	4.0 (3.7/4.4)	4.0 (3.7/4.4)
Resp	19.0 (16.0/23.0)	18.0 (16.0/22.0)
RR	0.7 (0.6/0.8)	0.7 (0.6/0.8)
SBP	118.0 (104.0/135.0)	122.0 (106.0/141.5)
SaO2	97.0 (93.0/98.0)	97.5 (95.7/98.8)
Temp	37.1 (36.6/37.6)	36.8 (36.4/37.5)
WBC	11.3 (8.3/15.0)	10.3 (7.5/14.1)
pH	7.4 (7.4/7.4)	7.4 (7.3/7.5)

Table 7: Categorical features in our data. We show the percentage of the specified value in each one of our hospital systems. Unit1 is an administrative reference to a medical ICU (as opposed to a surgical one). A Gender of 0 refers to female.

Variable (Value)	**System A**	**System B**
Gender (0)	40.48%	44.47%
Gender (1)	59.52%	55.53%
Unit1 (0)	18.85%	36.88%
Unit1 (1)	29.29%	34.42%

Table 8: Complete list of features for instances whose predictions get flipped with the use of c_E. We present one instance per hospital system.

Variable	**Patient A**	**Patient B**
Age	69	54
Gender	1	1
AST	-	23
Alkalinephos	-	58
BUN	13	13
Base Excess	-	-3.2
Bilirubin total	-	2.1
Calcium	8.6	4.39
Chloride	110	113
Creatinine	1	0.8
DBP	43	60
FiO2	-	1
Glucose	140	116
HCO3	27	22.4
HR	51	80
Hct	38.5	31.7
Hgb	13.1	10.9
Lactate	-	1.25
MAP	63	75
Magnesium	2.4	1.9
O2Sat	97	98
PTT	28.2	38.4
PaCO2	-	42
Phosphate	2.5	2.1
Platelets	107	140
Potassium	4	4.4
Resp	16	18
RR	1.17647	0.75
SBP	123	126
SaO2	-	95.9
Temp	36.06	37.4
Unit1	-	1
WBC	25.2	14.9
pH	-	7.34
True Label	1	0