EVALUATING THE EFFICACY OF FEDERATED SCORING SYSTEMS WITH HETEROGENEOUS ELECTRONIC HEALTH RECORDS

Qimimg Wu, Siqi Li, Di Miao, Yuqing Shang, Xin Li & Nan Liu
Center for Quantitative Medicine, Duke-NUS Medical School, Singapore

ABSTRACT

Federated learning in healthcare research has primarily focused on black-box models, leaving a notable gap in interpretability crucial for clinical decision-making. While scoring systems, acknowledged for their transparency, are widely employed in clinical science, there are notably limited privacy-preserving solutions for scoring system generators. FedScore, an example of such a solution, has been demonstrated using artificially partitioned data. In this study, we further improve FedScore and conduct empirical experiments utilizing real-world heterogeneous clinical data.

1 INTRODUCTION

Clinical scoring systems have been widely used in various medical domains, including emergency medicine (Oprita et al., 2014), cardiology (Dag et al., 2016), and neurology (Petersen et al., 2022) for patient risk stratification due to their model interpretability (Fleig et al., 2011). Existing data-driven scoring systems, such as the Supersparse Linear Integer Model (Ustun & Rudin, 2015) and the Interval Coded Scoring (Billiet et al., 2018), were mostly designed for data as a single dataset. With clinical practices becoming increasingly digitized, cross-institutional collaboration has been more prevalent in recent years to develop more robust and transferable models (Brisimi et al., 2018). Data sharing, however, is often impractical due to various privacy concerns (Antunes et al., 2022). As a result, federated learning (FL) has been widely employed to address privacy concerns (Sheller et al., 2020), enabling distributed model training without collecting or sharing patient data across institutions (Rieke et al., 2020).

The FedScore framework, recently proposed for generating scoring systems using clinical data from multiple sites in a privacy-preserving way, was initially demonstrated using artificially partitioned datasets (Li et al., 2023b). Our work further improves FedScore by employing several engineering-based (Li et al., 2023a) FL strategies, and we evaluate the enhanced method using real-world heterogeneous datasets by comparing the performance of FL models with local and centralized models. Through empirical experiments, we demonstrate that the modified FedScore framework is more robust and generalizable, capable of handling cross-institutional data heterogeneity for future international collaborations. Our code is available at this GitHub repository.

2 METHODS

As illustrated in Figure 1, the FedScore framework comprises five modules, most of which are versatile (Li et al., 2023b) and can be adjusted based on the specific needs of the intended clinical questions. Further details are available in Appendix A.1. Our work primarily focuses on Modules 3, 4 and 5. The original FedScore employs ODAL2 (Duan et al., 2019), a one-shot statistics-based (Li et al., 2023a) FL algorithm for conducting federated logistic regression (LR). ODAL2 is model-specific and assumes identically and independently distributed data, which may be inapplicable to real-world heterogeneous clinical data. Therefore, we adopt commonly used engineering-based FL strategies: FedAvg (McMahan et al., 2017), FedAvgM (Hsu et al., 2019) and q-FedAvg (Li et al., 2020), to enhance FedScore. We conduct an empirical comparison of FL models to local and central models created from the baseline scoring method (Xie et al., 2023) using real-world heterogeneous electronic health records (EHR) datasets. Specifically, we use the public dataset MIMIC-IV-ED (Johnson et al., 2023) collected from the United States, and private data (Liu et al., 2022) collected from Singapore General Hospital (SGH). Further details about both datasets are provided in Appendix A.2.

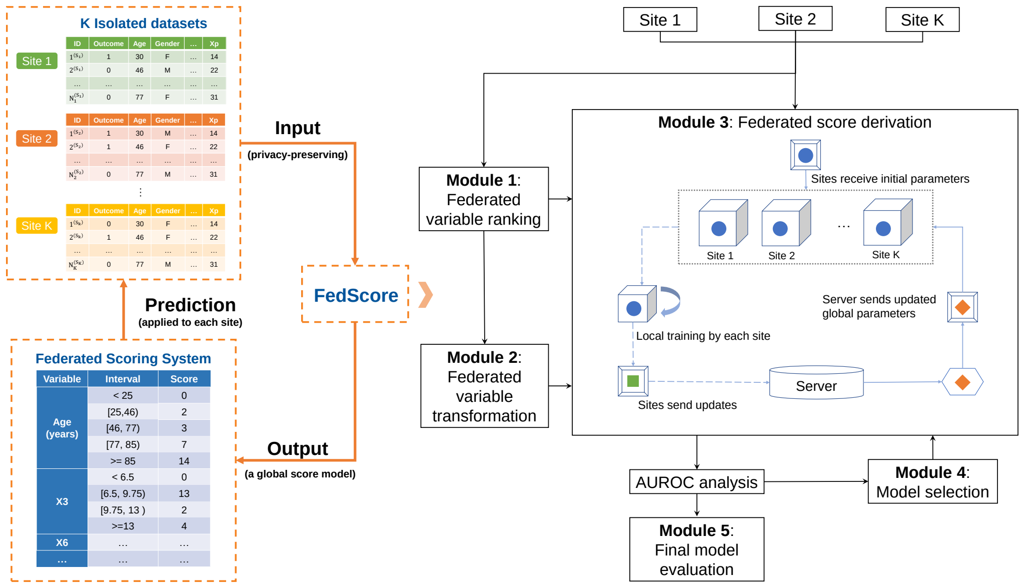

Figure 1: FedScore workflow.

3 Results & Conclusion

We conduct experiments using two real-world EHR datasets as two clients. The setup for hyperparameter fine-tuning is detailed in Appendix A.3. The best average area under the receiver operating characteristic curve (AUROC) values for FedScore using each FL strategy are reported in Table 1, alongside the baseline models. The trends in model performance for FedScore with different FL strategies, considering various fine-tuned hyperparameters are illustrated in Appendix A.4.

Table 1: Comparisons of FedScore and baseline models (AUROC).

FL / Local Model	MIMIC AUROC	SGH AUROC	Average AUROC
Local Model (MIMIC)	0.6751	0.8055	0.7403
Local Model (SGH)	0.6843	0.8229	0.7536
Central Model	0.6824	0.8232	0.7528
FedScore (FedAvg)	0.7024	**0.8243**	**0.7634**
FedScore (FedAvgM)	0.7025	0.8215	0.7620
FedScore (q-FedAvg)	**0.7056**	0.8210	0.7633

As displayed in Table 1 and Appendix A.4, with appropriate hyperparameters, all FedScore models outperform local and central models on average. These results demonstrate the stability and effectiveness of the enhanced FedScore, which consistently performs well across various FL algorithms.

In conclusion, we enhance the FedScore framework by integrating engineering-based FL algorithms for score derivation. The improved FedScore demonstrates increased robustness to real-world heterogeneity, highlighting its potential for future international clinical collaborations. Furthermore, the updated framework is highly adaptable, with the capacity to incorporate new score models beyond LR, thus laying the foundation for ongoing exploration across different types of clinical outcomes.

URM Statement

The authors acknowledge that all authors of this work meet the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology*, 13(4):1–23, May 2022. ISSN 2157-6912. doi: 10.1145/3501813. URL `http://dx.doi.org/10.1145/3501813`.

Lieven Billiet, Sabine Van Huffel, and Vanya Van Belle. Interval coded scoring: a toolbox for interpretable scoring systems. *PeerJ Computer Science*, 4:e150, April 2018. ISSN 2376-5992. doi: 10.7717/peerj-cs.150. URL `http://dx.doi.org/10.7717/peerj-cs.150`.

Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, April 2018. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2018.01.007. URL `http://dx.doi.org/10.1016/j.ijmedinf.2018.01.007`.

Ali Dag, Kazim Topuz, Asil Oztekin, Serkan Bulur, and Fadel M. Megahed. A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems*, 86:1–12, June 2016. ISSN 0167-9236. doi: 10.1016/j.dss.2016.02.007. URL `http://dx.doi.org/10.1016/j.dss.2016.02.007`.

Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H Chang, Hua Xu, Haitao Chu, Christopher H Schmid, Christopher B Forrest, John H Holmes, Martijn J Schuemie, Jesse A Berlin, Jason H Moore, and Yong Chen. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, December 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz199. URL `http://dx.doi.org/10.1093/jamia/ocz199`.

V. Fleig, F. Brenck, M. Wolff, and M.A. Weigand. Scoring-systeme in der intensivmedizin: Grundlagen, modelle, anwendung und grenzen. *Der Anaesthesist*, 60(10):963–974, October 2011. ISSN 1432-055X. doi: 10.1007/s00101-011-1942-8. URL `http://dx.doi.org/10.1007/s00101-011-1942-8`.

Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. 2019. URL `https://arxiv.org/abs/1909.06335`.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Siqi Li, Pinyan Liu, Gustavo G Nascimento, Xinru Wang, Fabio Renato Manzolli Leite, Bibhas Chakraborty, Chuan Hong, Yilin Ning, Feng Xie, Zhen Ling Teo, Daniel Shu Wei Ting, Hamed Haddadi, Marcus Eng Hock Ong, Marco Aurélio Peres, and Nan Liu. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. *Journal of the American Medical Informatics Association*, 30(12):2041–2049, 08 2023a. ISSN 1527-974X. doi: 10.1093/jamia/ocad170. URL `https://doi.org/10.1093/jamia/ocad170`.

Siqi Li, Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Chuan Hong, Feng Xie, Han Yuan, Mingxuan Liu, Daniel M. Buckland, Yong Chen, and Nan Liu. Fedscore: A privacy-preserving framework for federated scoring system development. *Journal of Biomedical Informatics*, 146:104485, 2023b. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2023.104485. URL `https://www.sciencedirect.com/science/article/pii/S153204642300206X`.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ByexElSYDr`.

Nan Liu, Feng Xie, Fahad Javaid Siddiqui, Andrew Fu Wah Ho, Bibhas Chakraborty, Gayathri Devi Nadarajan, Kenneth Boon Kiat Tan, and Marcus Eng Hock Ong. Leveraging large-scale electronic health records and interpretable machine learning for clinical decision making at the emergency department: Protocol for system development and validation. *JMIR Research Protocols*, 11(3): e34201, March 2022. ISSN 1929-0748. doi: 10.2196/34201. URL http://dx.doi.org/10.2196/34201.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

B Oprita, B Aignatoaie, and DA Gabor-Postole. Scores and scales used in emergency medicine. practicability in toxicology. *Journal of medicine and life*, 7(Spec Iss 3):4, 2014.

Kellen K. Petersen, Richard B. Lipton, Ellen Grober, Christos Davatzikos, Reisa A. Sperling, and Ali Ezzati. Predicting amyloid positivity in cognitively unimpaired older adults: A machine learning approach using a4 data. *Neurology*, 98(24):e2425–e2435, June 2022. ISSN 1526-632X. doi: 10.1212/wnl.0000000000200553. URL http://dx.doi.org/10.1212/WNL.0000000000200553.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1), September 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00323-1. URL http://dx.doi.org/10.1038/s41746-020-00323-1.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), July 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69250-1. URL http://dx.doi.org/10.1038/s41598-020-69250-1.

Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, November 2015. ISSN 1573-0565. doi: 10.1007/s10994-015-5528-6. URL http://dx.doi.org/10.1007/s10994-015-5528-6.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1), October 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01782-9. URL http://dx.doi.org/10.1038/s41597-022-01782-9.

Feng Xie, Yilin Ning, Mingxuan Liu, Siqi Li, Seyed Ehsan Saffari, Han Yuan, Victor Volovici, Daniel Shu Wei Ting, Benjamin Alan Goldstein, Marcus Eng Hock Ong, Roger Vaughan, Bibhas Chakraborty, and Nan Liu. A universal autoscore framework to develop interpretable scoring systems for predicting common types of clinical outcomes. *STAR Protocols*, 4(2):102302, 2023. ISSN 2666-1667. doi: https://doi.org/10.1016/j.xpro.2023.102302. URL https://www.sciencedirect.com/science/article/pii/S2666166723002691.

A Appendix

A.1 Overview of FedScore Framework

The FedScore framework is designed to offer a user-friendly interface for collaborative clinical score generation, consisting of five modules. In Module 1, local variable importance rankings for the P predictors are initially computed independently at each site using common ranking methods, such as random forest. For a given variable X_m where $1 \leq m \leq P$, the rank at site j, denoted as $q_j \in N$, is determined and then aggregated to generate the global ranking. This is achieved by calculating the weighted average of local rankings for each variable: $\sum_{j=1}^{K} w_j q_j$, where each integer set $[1, P] \in Z$ provides its global ranking. The normalized weights for site j should satisfy $\sum_{j=1}^{K} w_j = 1$, and the default option is sample-size-based weights.

In Module 2, numerical variables are transformed into categories to model the nonlinear effects of predictors. Local cut vectors are computed at each site and then federated using sample-size-weighted means. A maximum number of categories (usually 5) is set, and if exceeded, categories are combined until the requirement is met. Specifically, the quantiles of continuous variables are specified as $0\%, k_1\%, k_2\%, k_3\%, k_4\%$, with default settings for k_1, k_2, k_3, k_4 being $20, 40, 60, 80$, respectively. The unified cutoff for each continuous variable is then calculated by weighting the k values acquired at each site, using weights w_j.

Modules 3 and 4 involve fitting federated logistic regression models, utilizing up to the top K variables from the global ranking obtained in Module 1. A parsimony plot is generated to illustrate the relationship between model complexity and prediction performance, aiding in model selection. Following user decisions on variable selection and threshold values for cutting continuous variables, Module 5 fits the final federated logistic regression model. The resulting score table is applied to the testing set for the final performance evaluation.

Choice of model: In both the original and updated frameworks, we prioritize simple logistic regression to model clinical questions with binary outcomes. This choice is driven by its transparency and interpretability, making it the most common and natural option for clinicians in binary classification tasks in FL clinical research (Li et al., 2023a).

Motivation: In the original framework, ODAL2 was used in Modules 3 and 5 for federated logistic regression due to its communication efficiency as a one-shot algorithm. Its model-specific nature served as a motivation for conducting this follow-up study to develop a more convenient, model-agnostic framework.

Summary of Contributions: The major contribution of this work is threefold. Firstly, we addressed the limitation that the original framework is model specific and can only be applied to logistic regression. The new framework is model-agnostic and applicable to different types of models. Secondly, we offer a new implementation of the framework in Python that can handle heterogeneous data with new FL frameworks, whereas the original implementation in R only works for homogeneous data, which restricts its real-world applications. Thirdly, this study provides empirical evidence regarding the effect of different FL strategies and hyperparameters on model performance. This may be particularly useful for researchers or users who wish to use FedScore using these FL frameworks.

A.2 Details of MIMIC-IV-ED and SGH-ED Datasets

MIMIC-IV-ED is an open-source dataset of ED admissions at Beth Israel Deaconess Medical Center between 2011 and 2019. We first construct a master dataset following the pipeline proposed by Xie et al. (2022). The dataset is then filtered to include only ED admissions of Asian patients who were at least 21 years old. Observations with missing values are removed, resulting in a final cohort of 9071 admissions.

The SGH-ED dataset is a private dataset collected in the ED of Singapore General Hospital and extracted from the SingHealth Electronic Health Intelligence System. A waiver of consent was granted for EHR data collection and retrospective analysis, and the study has been approved by the Singapore Health Services' Centralized Institutional Review Board, with all data deidentified. The dataset is filtered to include only ED admissions of adult Chinese patients in 2019. Observations with missing values are also removed, resulting in a final cohort of 81110 admissions.

For both datasets, the binary outcome of interest is inpatient mortality. The datasets contain 17 variables, including age, gender, pulse (beats/min), respiration (times/min), peripheral capillary oxygen saturation (SpO_2; %), diastolic blood pressure (mm Hg), systolic blood pressure (mm Hg), and comorbidities such as myocardial infarction, congestive heart failure, peripheral vascular disease, stroke, dementia, chronic pulmonary disease, rheumatic disease, peptic ulcer disease, paralysis and kidney disease.

A.3 HYPERPARAMETERS

Table 2: Hyperparameter values for fine-tuning.

Hyperparameter	FL Models	Candidate values
Server-side learning rate	FedAvgM, q-FedAvg	0.001, 0.005, 0.01, 0.05, 0.1, **0.5**, 1.0
Local epochs	FedAvg, FedAvgM, q-FedAvg	1, 3, 5, **10**, 20, 30, 50
Rounds of communication	FedAvg, FedAvgM, q-FedAvg	5, 10, 20, **30**, 40, 50
q parameter	q-FedAvg	0, 0.1, **0.2**, 0.3, 0.5, 0.8, 1.0
Momentum	FedAvgM	0, 0.1, 0.2, 0.5, 0.6, 0.8, **1.0**

A.4 IMPACT OF HYPERPARAMETERS ON FL MODEL PERFORMANCES

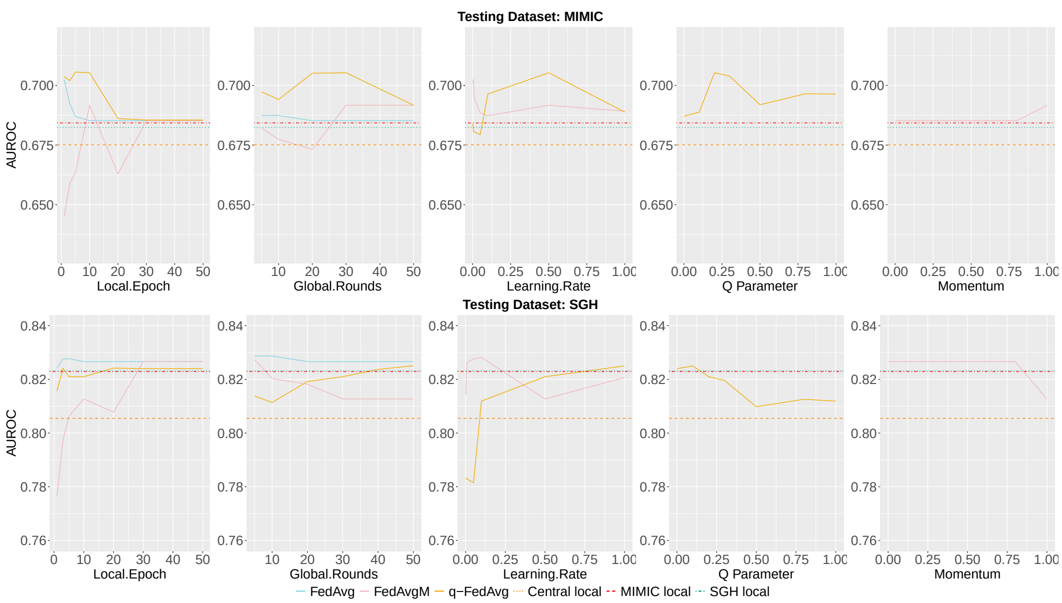

Figure 2: Performance of FL models with varying hyperparameters.