ARTIFICIAL INTELLIGENCE

original reports abstract

Cloud-Based Federated Learning Implementation Across Medical Centers

Suraj Rajendran, BS[1,2]; Jihad S. Obeid, MD[3]; Hamidullah Binol, PhD[4]; Ralph D`Agostino Jr, PhD[5]; Kristie Foley, PhD[6]; Wei Zhang, PhD[1]; Philip Austin, BS[7]; Joey Brakefield, BS[8]; Metin N. Gurcan, PhD[4]; and Umit Topaloglu, PhD[1,4,5]

PURPOSE Building well-performing machine learning (ML) models in health care has always been exigent because of the data-sharing concerns, yet ML approaches often require larger training samples than is afforded by one institution. This paper explores several federated learning implementations by applying them in both a simulated environment and an actual implementation using electronic health record data from two academic medical centers on a Microsoft Azure Cloud Databricks platform.

MATERIALS AND METHODS Using two separate cloud tenants, ML models were created, trained, and exchanged from one institution to another via a GitHub repository. Federated learning processes were applied to both artificial neural networks (ANNs) and logistic regression (LR) models on the horizontal data sets that are varying in count and availability. Incremental and cyclic federated learning models have been tested in simulation and real environments.

RESULTS The cyclically trained ANN showed a 3% increase in performance, a significant improvement across most attempts ($P < .05$). Single weight neural network models showed improvement in some cases. However, LR models did not show much improvement after federated learning processes. The specific process that improved the performance differed based on the ML model and how federated learning was implemented. Moreover, we have confirmed that the order of the institutions during the training did influence the overall performance increase.

CONCLUSION Unlike previous studies, our work has shown the implementation and effectiveness of federated learning processes beyond simulation. Additionally, we have identified different federated learning models that have achieved statistically significant performances. More work is needed to achieve effective federated learning processes in biomedicine, while preserving the security and privacy of the data.

JCO Clin Cancer Inform 5:1-11. © 2021 by American Society of Clinical Oncology

Recent advancements in artificial intelligence (AI) have demonstrated the potential to transform medicine[1] and are promising for improving outcomes while reducing the cost of patient care because of its capability for earlier, more accurate diagnosis and personalized patient-centered care. Image classification, speech recognition, and natural language processing have seen some noteworthy achievements.[2] Moreover, thanks to machine learning (ML), hospitals can accomplish more efficient clinical workflows by reducing unnecessary procedures, which leads to further cost reductions.[1]

The performance of an ML algorithm depends highly on the amount and quality of data it is trained on, particularly for more complex models.[3] In the era of precision medicine, the availability of complex multidimensional patient data sets requires larger population samples for generalization.[4] Furthermore, the scarcity of data in underrepresented populations may lead to biases when training data do not sufficiently reflect the attributes of these populations.[5] Healthcare data quality and algorithmic challenges are also known barriers for ML.[6]

Many approaches have been proposed to address the lack of data heterogeneity.[7,8] The most promising of these approaches requires multi-institutional collaborations that would increase not only the size of the training data but also its data diversity. Ideally, study data from each institution would be shared via a central data store where a single model can be trained on the combined multi-institutional data. However, there are several obstacles to implementing such a solution.[7-9] First, central storage and transferring large amounts of data over the network have an exorbitant associated cost.[10] The second major obstacle is the regulatory barrier surrounding patient data protection.

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on October 13, 2020 and published at ascopubs.org/journal/cci on January 7, 2021: DOI https://doi.org/10.1200/CCI.20.00060

CONTEXT

Key Objective

Machine learning (ML) models have the promise and potential of transforming health care from diagnosis to the treatment recommendations. However, lack of sufficient heterogeneous data because of patient privacy protections risks ML model generalizability. For this study, multiple ML models were implemented on highly heterogeneous data for a validated scientific question across medical centers without sharing data.

Knowledge Generated

Heterogeneous data across centers have improved the model performances compared with a simulation in a single institution. Additionally, cloud platforms have adequate tools and security controls to run federated learning implementations.

Relevance

Despite advancements, ML models are still not widely used at clinics because of a lack of sufficient and diverse data. This study has tested a platform on which many organizations can improve their models in a federated learning fashion.

Sharing patient data (with or without protected health information) requires several legal and regulatory approvals and interinstitutional agreements, which can be a cumbersome and lengthy process.

The aforementioned obstacles necessitate the development of various federated learning strategies to train ML models without sharing confidential patient data across institutional firewalls. Federated learning is an ML framework in which models are trained on data that reside at each institution.[7,9]

Federated Learning Models

Federated learning models can be divided mainly into two groups, parallel and nonparallel. Parallel training is developed with the intention of faster (and optimized) completion of the runs; however, it often poses a large logistical problem in certain applications, including the lack of uniformity in network connection speeds and computational resources. On the other hand, nonparallel training, although less efficient, can be implemented across nonhomogeneous computing environments without the need for synchronization of runs. Chang et al[9] have tested three nonparallel training structures: ensemble training, single weight training, and cyclical weight training.

Nonparallel models, including ensemble training, involve training separate models at each of the institutions on their respective data and subsequently gathering averages of the weights for each model toward a final model. In single weight training, the model is first trained on data from one institution until the training validation loss begins to plateau. The trained model is then transferred to the second institution where it is further trained on new data. The same process is continued across other institutions in the collaborative environment (Appendix Fig A1). It is important to note that the training order of the specific institutions often affects the final performance of the model.[9]

Cyclical weight training is very similar to single weight training in that the same model is transferred from one institution to the next with two main important differences. First, at each institution, the model is only trained for a preset number of epochs (generally, a lower number of epochs yield a better performing final model). Second, after the last institution trains the model with the initial preset number of epochs, the model is returned to the first institution to be retrained for the second group of the preset number of epochs. In essence, the model is trained by each institution multiple times before the final model is produced, hence the cyclical nature of this process. The process is summarized in the Appendix Figure A2.[9]

OBJECTIVE

To evaluate our federated ML approach, we trained ML models to predict the risks of diseases associated with tobacco and radon using data from electronic health records (EHRs) at two healthcare systems. Tobacco use is the leading modifiable risk factor for lung cancer. The majority of counties in the Carolinas have adult smoking rates that exceed the national average.[11] Radon is a colorless, odorless, radioactive gas. According to the Environmental Protection Agency (EPA), it is the most significant modifiable risk factor for lung cancer after tobacco use.[12,13] Radon is present in the ground as a byproduct of uranium decay, and it typically enters homes and buildings as it diffuses into the air.[14,15] At present, North and South Carolina have an average indoor radon screening level $>$ 4 pCi/L in many of its counties, and the EPA recommends radon mitigation measures at 4 pCi/L or greater.[15]

Prior studies have already demonstrated that smoking and radon have both independent and synergistic effects on lung cancer and chronic obstructive pulmonary disease (COPD) incidences. Given that these are well-established risk factors for lung cancer and COPD, they are optimal use

cases for establishing the capability of several federated ML models on disease outcomes. Moreover, because data are derived from both EHRs (patient-level data) and publicly available data (ecological data), they also allow us to test these models using data extracted from distinct sources.

MATERIALS AND METHODS

The Institutional Review Boards (IRBs) of the Wake Forest University Health Sciences and Medical University of South Carolina (MUSC) approved the study with protocol numbers IRB00056277 and Pro00090097, respectively. Upon IRB approval, patient EHR data from Wake Forest Baptist Medical Center (WFBMC) and the MUSC were used to test the efficacy and performance of federated learning models trained in a Databricks (Databricks Inc., San Francisco, CA) cloud environment. The deidentified data included each patient's sex, age (in years), race (Asian, Caucasian, African American, American Indian, other, or patient refused), smoking level (0, 1, or 2), radon exposure level (0, 1, or 2), and diagnosis of lung cancer and COPD (ICD-10: C34 and J44, respectively). Sex and race were integer encoded. For smoking, 0 is nonsmoker, 1 is former smoker (quit 5 + years ago), and 2 is coded to represent current smokers. For radon, 0 is used for an exposure of < 2 pCi/L, 1 is between 2 pCi/L and 4 pCi/L, and 2 is used for cases that have an exposure of > 4 pCi/L. Radon data for both North Carolina and South Carolina were gathered from public websites and from private companies that perform radon measurements.[16]

Data Security

To avoid sharing the data across institutions, we employed several MS Microsoft Azure (Microsoft, Redmond, WA) security toolkits.[17] First, data from each institution had been loaded to the respective Azure storage account, which is equipped with its own access control mechanism. Moreover, we needed to avoid the option of in-code access secret codes, usernames, and passwords, being stored in shared Jupyter Notebooks, thus compromising security. To circumvent this issue, we enabled the Azure Key Vault to house the storage account's access secrets. The Key Vault is configured for each study team through the Azure Active Directory (AD), which relies on institutional AD-based user management. Finally, we had to create a scope in Databricks with Key Vault's Domain Name System and Directory ID to point notebooks to the correct Key Vault. The configuration is outlined on Microsoft's public Microsoft Docs repository.[18]

Study Design

In the first part of this study, we attempted to simulate single weight and cyclical weight training in a local environment based on smoking and radon data using WFBMC data. Subsequently, actual implementation involves the two-institution federated learning processes with separate Azure subscriptions (and data) that were accessible only by the respective teams. We then tested the use of separate Databricks resources in combination with GitHub as a method of sharing model weights with various institutions without sharing data. We hypothesized that both methods would improve model performance compared with a model trained on only one institution's data.

Preprocessing. The study data have three diagnosis categories for prediction: lung cancer, other, and COPD that were integer encoded. To reduce the problem into a binary classification task, we created two outcome classes: class 1, which is a disease state with either lung cancer or COPD and class 0, no disease state (ie, no cancer or COPD). The distributions of the data sets are shown in the Appendix Figure A3 and Table A1. Conducting a test of homogeneity between the two data sets resulted in a χ^2 statistic of 19,008 ($P < .0001$), suggesting that the two data sets are extremely heterogeneous. Data were split (67/33) between training and testing data for each institution.

Because of a heavy imbalance in the Wake Forest data set (most of the labels were class 0), the synthetic minority oversampling technique (SMOTE)[19] was used to reduce the imbalance in WFBMC data. SMOTE effectively increases the count of the minor class in a set of data.

Model construction. In this study, we constructed an artificial neural network (ANN) to model our data. The model's weights were initialized with a Keras initializer. The learning rate was 0.001. The model consisted of two dense layers and used Adam as its optimization function.[20]

For single weight training, the models implemented an early stopping algorithm dependent on validation loss. For the cyclical weight training, each institution trained the model for five cycles with 10 epochs each. Additionally, the model was trained for 10 epochs and then transferred to the next institution. A total of five models were built: base 1 (model trained on institution 1's data), base 2 (model trained on institution 2's data), single weight model A (institution 1 trains the model first), single weight model B (institution 2 trains the model first), and cyclical weight model. Furthermore, two single weight models were necessary to capture performance changes because of the ordering of institutions in single weight training. Test data from each institution were run through the five models, and model performances were captured. For each of these measurements, 10 trials were conducted and then averaged to reach a final performance metric. A student's t test was used to determine significance between different models.

Logistic regression (LR) models were also constructed to test whether federated learning methods proved efficient when applied on traditional ML methods. It is important to note that because LR is not an epoch-based learning algorithm, only single weight federated learning was conducted in addition to base tests. Data aggregation was conducted as was done for the ANN. Model training took an average of 6.35 seconds. Transfer mechanism to and from GitHub took 3.02 and 3.05 seconds on average, respectively.

TABLE 1. ANN Model Performances on Institution 1 Test Data: Base 1 (Model Trained on Institution 1's Data), Base 2 (Model Trained on Institution 2's Data), Single Weight Model A (Institution 1 Trains the Model First), and Single Weight Model B (Institution 2 Trains the Model First); ANN Model Performances on Institution 2 Test Data: Base 1 (Model Trained on Institution 1's Data), Base 2 (Model Trained on Institution 2's Data), Single Weight Model A (Institution 1 Trains the Model First), and Single Weight Model B (Institution 2 Trains the Model First)

Model	F1 Score	Precision	Recall	Accuracy
ANN tested on institution 1				
Base (trained on institution 1 data)	0.4374 ± 0.0086	0.2962 ± 0.0077	0.8365 ± 0.0219	0.6797 ± 0.0121
Base (trained on institution 2 data)	0.4375 ± 0.0115	0.2899 ± 0.0091	0.8914 ± 0.0240	0.6587 ± 0.0137
Single weight model A	0.4473 ± 0.0106	0.3022 ± 0.0084	0.8624 ± 0.0368	0.6829 ± 0.0141
Single weight model B	0.4426 ± 0.0137	0.3039 ± 0.0118	0.8161 ± 0.0364	0.6939 ± 0.0163
Cyclical weight	0.4524 ± 0.0108	0.3046 ± 0.0079	0.8796 ± 0.0201	0.6832 ± 0.0088
ANN tested on institution 2				
Base (trained on institution 1 data)	0.4752 ± 0.0156	0.3301 ± 0.0134	0.8482 ± 0.0236	0.6888 ± 0.0136
Base (trained on institution 2 data)	0.4686 ± 0.0163	0.3171 ± 0.0169	0.9009 ± 0.0211	0.6558 ± 0.0269
Single weight model A	0.4875 ± 0.0074	0.3459 ± 0.0078	0.8267 ± 0.0304	0.708 ± 0.0116
Single weight model B	0.4936 ± 0.0152	0.3532 ± 0.0140	0.8210 ± 0.0323	0.7168 ± 0.0159
Cyclical weight	0.4987 ± 0.0090	0.3512 ± 0.0080	0.8600 ± 0.0090	0.7094 ± 0.0086

Abbreviation: AAN, artificial neural network.

Simulation. As the correlation of radon and tobacco with lung cancer and COPD is previously established, we sought to demonstrate that ML models resulted in the same correlation when using federated learning processes. For the simulation, we created two mock institutions with two unique training sets and one shared test set. To create these two unique training sets, data from the WFBMC were randomly shuffled and divided into three parts: two training sets and one test set, each of equal size. To ensure replicability and objectiveness, all trials for both cyclical weight training and single weight training were performed on the same splits of data.

Implementation on Azure Databricks. To set up the federated learning environment on Databricks, it was essential to develop a method to save ML models so that models can be transferred across institutions, as shown in the Appendix Figure A4. This allows training in an asynchronous fashion. Each institution's data were located in respective institutional Azure storage accounts. The Azure Key Vault was configured to limit access to only study staff via Azure AD. The data were accessed from the Jupyter Notebook through the Key Vault with an appropriate access scope. The ANN and LR models were implemented in the Jupyter environment.

To perform the federated learning, the trained model was saved on a shared GitHub repository, which was accessible by either institution. GitHub version control was required for model upload.[21] The model could then be shared with other

TABLE 2. LR Model Performances on Institution 1 Test Data: Base 1 (Model Trained on Institution 1's Data), Base 2 (Model Trained on Institution 2's Data), Single Weight Model A (Institution 1 Trains the Model First), and Single Weight Model B (Institution 2 Trains the Model First); LR Model Performances on Institution 2 Test Data: Base 1 (Model Trained on Institution 1's Data), Base 2 (Model Trained on Institution 2's Data), Single Weight Model A (Institution 1 Trains the Model First), and Single Weight Model B (Institution 2 Trains the Model First)

Model	F1 Score	Precision	Recall	Accuracy
LR tested on institution 1				
Base (trained on institution 1 data)	0.4485 ± 0.0085	0.7634 ± 0.0152	0.3175 ± 0.0061	0.7206 ± 0.0048
Base (trained on institution 2 data)	0.4570 ± 0.0083	0.7183 ± 0.0150	0.3352 ± 0.0061	0.7460 ± 0.0043
Single weight model A	0.4579 ± 0.0131	0.7141 ± 0.0151	0.3373 ± 0.0140	0.7422 ± 0.0058
Single weight model B	0.4547 ± 0.0087	0.7763 ± 0.0166	0.3215 ± 0.0063	0.7229 ± 0.0052
LR tested on institution 2				
Base (trained on institution 1 data)	0.4889 ± 0.0062	0.7686 ± 0.0168	0.3585 ± 0.0042	0.7300 ± 0.0039
Base (trained on institution 2 data)	0.4725 ± 0.0050	0.6952 ± 0.0110	0.3579 ± 0.0040	0.7392 ± 0.0034
Single weight model A	0.4666 ± 0.0086	0.6994 ± 0.0172	0.3502 ± 0.0105	0.7378 ± 0.0059
Single weight model B	0.4884 ± 0.0037	0.7771 ± 0.0049	0.3560 ± 0.0040	0.7264 ± 0.0043

Abbreviation: LR, logistic regression.

TABLE 3. ANN Model Performances on WF's Test Data: Base 1 (Model Trained on WF's Data), Base 2 (Model Trained on MUSC's Data), Single Weight Model A (WF Trains the Model First), and Single Weight Model B (MUSC Trains the Model First); ANN Model Performances on MUSC's Test Data: Base 1 (Model Trained on WF's Data), Base 2 (Model Trained on MUSC's Data), Single Weight Model A (WF Trains the Model First), and Single Weight Model B (MUSC Trains the Model First)

Model	F1 Score	Precision	Recall	Accuracy
ANN tested on WF				
Base (trained on WF's data)	0.4612 ± 0.0120	0.3285 ± 0.0103	0.7682 ± 0.0146	0.7022 ± 0.0112
Base (trained on MUSC's data)	0.3800 ± 0.0076	0.2986 ± 0.0119	0.7342 ± 0.0213	0.5990 ± 0.0047
Single weight model A	0.3675 ± 0.0145	0.2355 ± 0.0035	0.6425 ± 0.0314	0.6269 ± 0.0113
Single weight model B	0.4716 ± 0.0158	0.3876 ± 0.0150	0.6467 ± 0.0132	0.7481 ± 0.0167
Cyclical weight	0.4656 ± 0.0045	0.3567 ± 0.0104	0.7148 ± 0.0069	0.7352 ± 0.0096
ANN tested on MUSC				
Base (trained on WF's data)	0.5267 ± 0.0166	0.6896 ± 0.0138	0.4265 ± 0.0240	0.5540 ± 0.0166
Base (trained on MUSC's data)	0.6685 ± 0.0044	0.8228 ± 0.0038	0.5630 ± 0.0072	0.6748 ± 0.0024
Single weight model A	0.6749 ± 0.0022	0.8239 ± 0.0040	0.5715 ± 0.0049	0.6792 ± 0.0008
Single weight model B	0.5512 ± 0.0228	0.6884 ± 0.0154	0.4603 ± 0.0301	0.5641 ± 0.0144
Cyclical weight	0.6704 ± 0.0062	0.8053 ± 0.0055	0.5693 ± 0.0097	0.6816 ± 0.0040

Abbreviations: ANN, artificial neural network; MUSC, Medical University of South Carolina; WF, Wake Forest.

collaborators who had access to the shared GitHub repository. The GitHub application programming interface was accessed via the PyGitHub Python library to effectively implement this system.[22] Each institution had a unique and personal access token to this repository that was saved as a Databricks secret in their respective Azure Account. Upon training the model, each institution had access to the shared model, which was saved as a pickle file in their Jupyter Notebook. The shared repository could be accessed asynchronously, and the pertinent model could be extracted.

RESULTS

Four performance metrics were captured: F1 score, precision, recall, and accuracy. The primary metric that was used for model improvement was the F1 score, given the class imbalance. The F1 score is the harmonic mean of the precision and recall.

The mean metrics along with their respective standard deviations up to four significant digits are shown in Tables 1-4. T1-T4

Simulation

Table 1 represents the performance of the ANN models across each institution's test data. All three federated learning models depicted in Table 1 had a significant increase in accuracy and F1 score over the base models when tested on both institutions' test data ($P < .05$). The single weight model showed the most significant improvement over the base values.

TABLE 4. LR Model Performances on WF's Test Data: Base 1 (Model Trained on WF's Data), Base 2 (Model Trained on MUSC's Data), Single Weight Model A (WF Trains the Model First), and Single Weight Model B (MUSC Trains the Model First); LR Model Performances on MUSC's Test Data: Base 1 (Model Trained on WF's Data), Base 2 (Model Trained on MUSC's Data), Single Weight Model A (WF Trains the Model First), and Single Weight Model B (MUSC Trains the Model First)

Model	F1 Score	Precision	Recall	Accuracy
LR tested on WF				
Base (trained on WF's data)	0.4784 ± 0.0018	0.7501 ± 0.0049	0.3512 ± 0.0022	0.7381 ± 0.0026
Base (trained on MUSC's data)	0.4128 ± 0.0142	0.5939 ± 0.0531	0.3204 ± 0.0368	0.7278 ± 0.0380
Single weight model A	0.4110 ± 0.0071	0.5814 ± 0.0186	0.3184 ± 0.0133	0.7329 ± 0.0157
Single weight model B	0.4778 ± 0.0028	0.7511 ± 0.0048	0.3503 ± 0.0030	0.7402 ± 0.0035
LR tested on MUSC				
Base (trained on WF's data)	0.5220 ± 0.0066	0.3992 ± 0.0075	0.7540 ± 0.0036	0.5741 ± 0.0035
Base (trained on MUSC's data)	0.6746 ± 0.0050	0.5766 ± 0.0025	0.8130 ± 0.0041	0.6760 ± 0.0010
Single weight model A	0.6750 ± 0.0006	0.5774 ± 0.0025	0.8123 ± 0.0049	0.6760 ± 0.0015
Single weight model B	0.5182 ± 0.0047	0.3946 ± 0.0066	0.7550 ± 0.0054	0.5726 ± 0.0014

Abbreviations: LR, logistic regression; MUSC, Medical University of South Carolina; WF, Wake Forest.

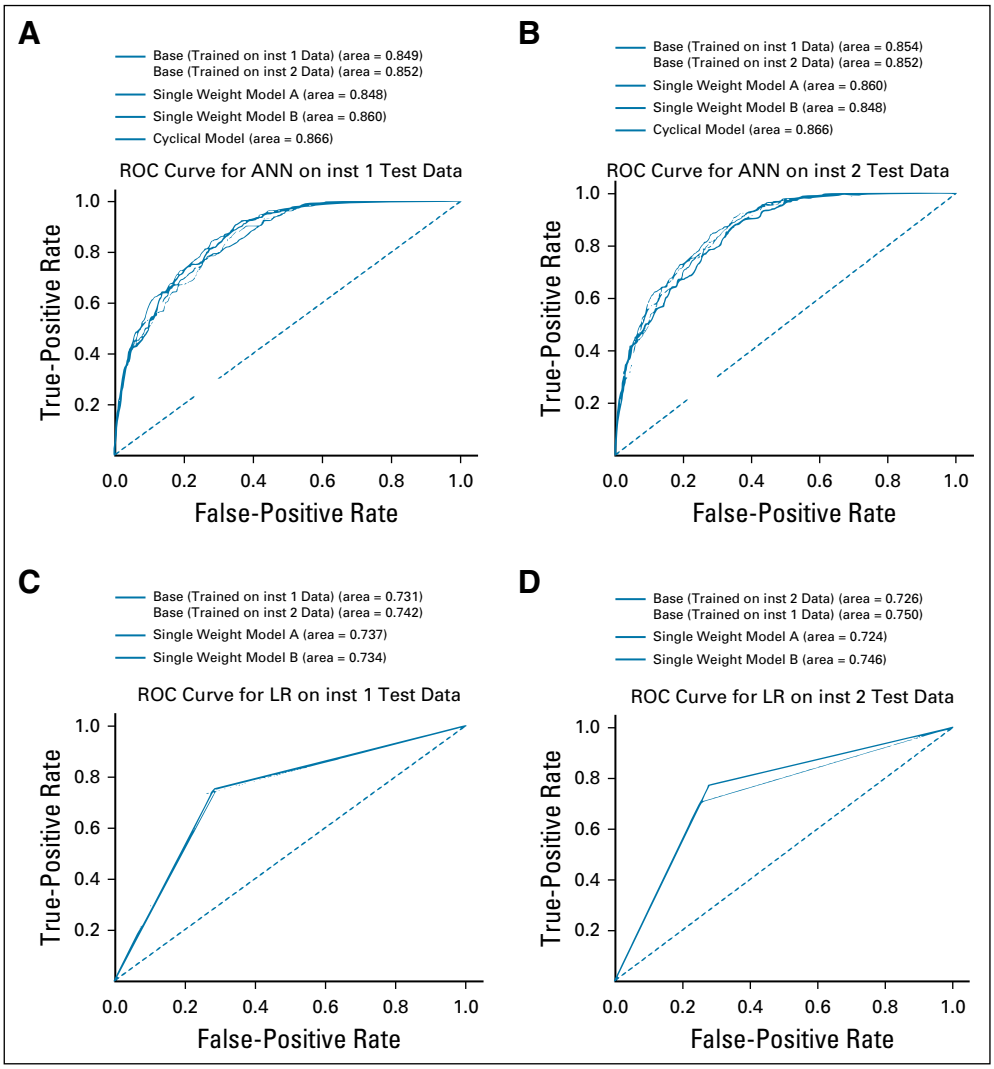

FIG 1. ROC curves corresponding to performance metrics in tables. (A) ROC curve based on ANN models' performances against institution 1 test data. (B) ROC curve based on ANN models' performances against institution 2 test data. (C) ROC curve based on LR models' performances against institution 1 test data. (D) ROC curve based on LR models' performances against institution 2 test data. ANN, artificial neural network; LR, logistic regression; ROC, receiver operating characteristic.

Table 2 represents the performances of LR models across each institution's test data. There was no statistically significant improvement of the accuracy or F1 score provided by the federated learning methods in the case of LR models. Figure 1 shows the receiver operating characteristic (ROC) curves that also show the earlier stated statistical significance.

Actual Implementation on Azure

Table 3 represents the performances of the ANN models across each institution's test data on the Databricks environment. When applied to WFBMC test data, the single weight model B and cyclical weight model had a significantly higher accuracy than both base models ($P < .01$). Against the MUSC's test data, the cyclical weight model showed significant improvement in both F1 score and accuracy ($P < .05$).

Similar to the simulation, the LR federated learning methods did not show much improvement over the base model. Table 4 represents the performances of LR models across each institution's test data, and Figure 2 shows the ROC curves that also show the earlier stated statistical significance.

DISCUSSION

The purpose of this study was to determine whether federated learning methods would improve the performance of the ML models in health care while preserving the security and privacy of the patient data. The results, from both the simulation environment and the actual implementation on Databricks, suggest that federated learning methods do have the potential to improve model performances with a few caveats.

One observation made during this study was that the federated learning methods generally did not improve the performance of LR. This might be due to the lower complexity of LR and lack of an iterative training process (ie, epochs) when compared with ANN models.

For the ANN models, there was generally one federated model that performed better than baseline models for

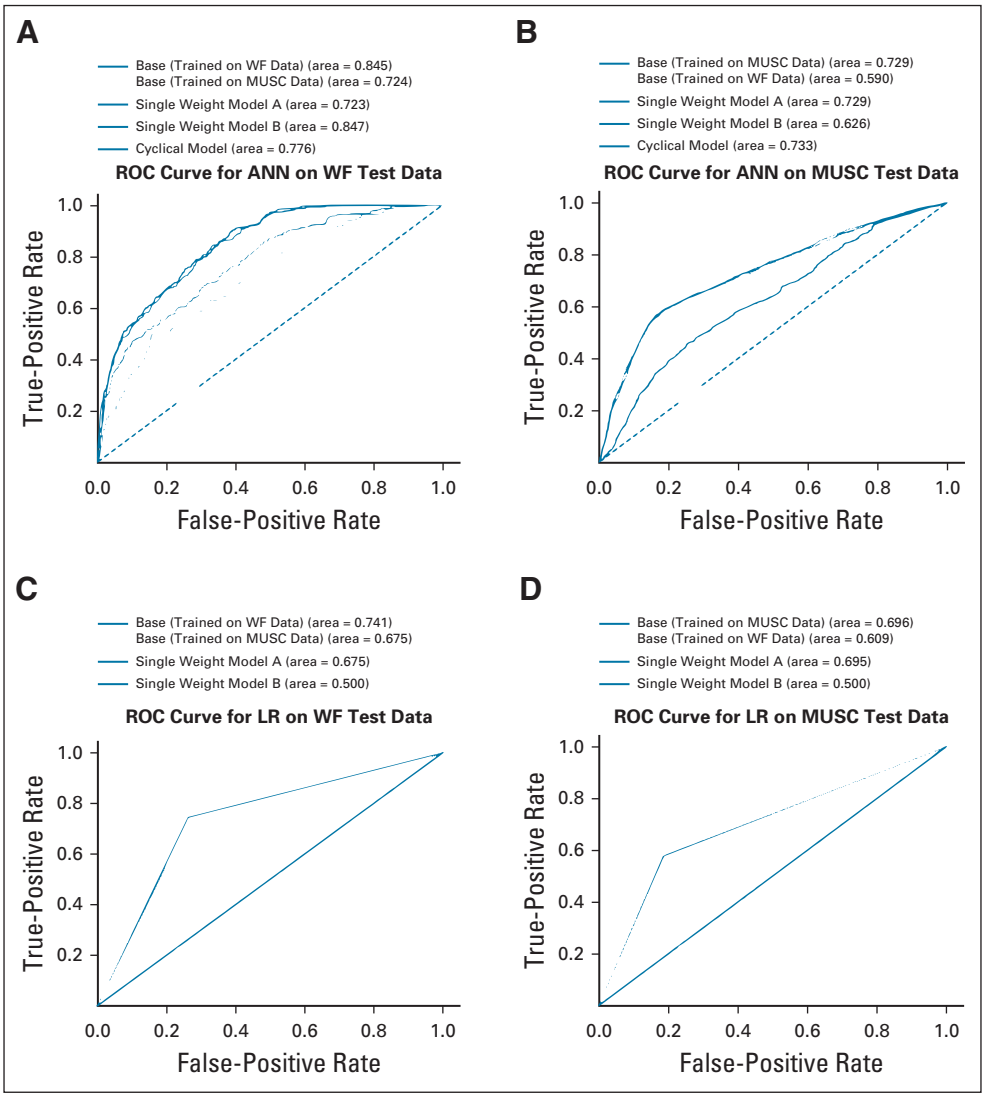

FIG 2. ROC curves corresponding to performance metrics in tables. (A) ROC curve based on ANN models' performances against WF's test data. (B) ROC curve based on ANN models' performances against the MUSC's test data. (C) ROC curve based on LR models' performances against WF's test data. (D) ROC curve based on LR models' performances against the MUSC's test data. ANN, artificial neural network; LR, logistic regression; MUSC, Medical University of South Carolina; ROC, receiver operating characteristic; WF, Wake Forest.

each of the four attempts shown in the results. However, the type of federated model that performed best in each of these situations varied. One such reason may be the training order of institutions in these federated learning methods. As previous studies have shown, the training order of institutions impacts the final performance of the model, especially for single weight training, which was shown a similar effect in our results. Whereas single weight model A, when tested with Wake Forest data, provided an F1 score of 0.3675, the same federated process in single weight model B provided an F1 score of 0.4716. We hypothesize that this is due to the persistent role that the order of institutions plays in single weight training.

One concern with the federated ML methods is their susceptibility to an adversarial attack in the collaborative environment. Several studies have shown that there are multiple attacks such as membership interference or attribute inference that could affect the safety of the models.[23-27] Privacy-preserving AI has been coined and implemented by some; however, it does not yet provide the necessary protection in a federated learning environment and may even lead to the exposure of potentially sensitive data.[28-33] Further evaluation of model security and alternative approaches are still needed and will be reserved for future studies.

One limitation of this work is the use of a relatively simplistic model with few features for testing our federated learning approach. Future work should include the implementation of more complex ML, including deep learning models using this infrastructure.

In this project we demonstrated a federated learning process for ML models in a collaborative academic health center setting going beyond simulation. While this is still an emerging field, our work establishes the potential for federated learning to significantly improve model performance. Previous studies have focused mainly on simulated data; however, we have taken the additional step of

implementing them across institutions. In doing so, we have demonstrated an efficient way of sharing and accessing models across institutions. While our investigation focused on binary classification, the same protocol for nonbinary outcome analysis can be easily implemented with new ML models.

AFFILIATIONS

[1]Department of Cancer Biology, Wake Forest University School of Medicine, Winston Salem, NC
[2]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA
[3]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC
[4]Center for Biomedical Informatics, Wake Forest University School of Medicine, Winston Salem, NC
[5]Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Winston Salem, NC
[6]Department of Implementation Science, Wake Forest University School of Medicine, Winston Salem, NC
[7]Databricks, Boca Raton, FL
[8]Microsoft, Atlanta, GA

CORRESPONDING AUTHOR

Umit Topaloglu, PhD, FAMIA, Wake Forest Baptist Comprehensive Cancer Center, Center for Biomedical Informatics, Cancer Biology & Biostatistics and Data Sciences, Wake Forest University School of Medicine, 1 Medical Center Boulevard, Winston-Salem, NC 27157; e-mail: umit.topaloglu@wakehealth.edu.

SUPPORT

Supported in part by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197) and Betsy Sykes Research Funds (19-02673).

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Ralph D`Agostino
Consulting or Advisory Role: Exelixis, Target Health

Wei Zhang
Consulting or Advisory Role: Astellas Pharma

Metin N. Gurcan
Leadership: Otologic Technologies, Inc.
Consulting or Advisory Role: Otologic Technologies, Inc.

No other potential conflicts of interest were reported.

AUTHOR CONTRIBUTIONS

Conception and design: Suraj Rajendran, Jihad S. Obeid, Hamidullah Binol, Kristie Foley, Philip Austin, Joey Brakefield, Metin N. Gurcan, Umit Topaloglu
Financial support: Jihad S. Obeid
Provision of study materials or patients: Jihad S. Obeid
Collection and assembly of data: Suraj Rajendran, Jihad S. Obeid, Hamidullah Binol, Kristie Foley, Philip Austin, Umit Topaloglu
Data analysis and interpretation: Suraj Rajendran, Jihad S. Obeid, Hamidullah Binol, Ralph D`Agostino, Kristie Foley, Wei Zhang, Metin N. Gurcan, Umit Topaloglu
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

ACKNOWLEDGMENT

The authors acknowledge the use of the services and facilities funded by the National Center for Advancing Translational Sciences (NCATS) and National Institutes of Health (UL1TR001420 and UL1TR001450). They also thank Paul Weese, John Clark, Erin Quigley, and Bill Campman for their help during cloud implementation and Wendell Futrell and Tony Gwyn for their help gathering and organizing radon data sets. Finally, they thank Kathy Walker for the edits on the manuscript.

REFERENCES

1. Fralick M, Colak E, Mamdani M: Machine learning in medicine. N Engl J Med 2019;380:2588-2589
2. Lee J, Yoon W, Kim S, et al: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234-1240
3. Obermeyer Z, Emanuel EJ: Predicting the future-big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-1219
4. Oleynik M, Kugic A, Kasáč Z, et al: Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 2019;26:1247-1254
5. Buda M, Maki A, Mazurowski MA: A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018;106:249-259
6. Sessions V, Valtorta M: The effects of data quality on machine learning algorithms. IQ Concepts, Tools, Metrics, Measures, Models, and Methodologies. 2006
7. Bonawitz K, Eichner H, Grieskamp W, et al: Towards federated learning at scale: System design. SysML 2019
8. Liu D, Miller T, Sayeed R, et al: FADL: Federated-Autonomous Deep Learning for Distributed Electronic Health Record. Machine Learning for Health (ML4H) Workshop at NeurIPS, 2018
9. Chang K, Balachandar N, Lam C, et al: Distributed deep learning networks among institutions for medical imaging. J Am Med Informa Assoc 2018;25:945-954
10. Barlow S: White Paper Comparing the Three Major Approaches to Healthcare Data Warehousing: A Deep Dive Review, HealthCatalyst, 2017
11. Map of Cigarette Use Among AdultslSTATE SystemlCDC: https://www.cdc.gov/statesystem/cigaretteuseadult.html
12. Find Information About Local Radon Zones and State Contact InformationlRadonlUS EPA: https://www.epa.gov/radon/find-information-about-local-radon-zones-and-state-contact-information

13. US EPA O: Find Information About Local Radon Zones and State Contact Information. https://www.epa.gov/radon/find-information-about-local-radon-zones-and-state-contact-information
14. Lubin JH, Boice JD: Lung cancer risk from residential radon: Meta-analysis of eight epidemiologic studies. J Natl Cancer Inst 1997;89:49-57
15. Lubin JH, Boice JD, Edling C, et al: Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. J Natl Cancer Inst 1995;87:817-827
16. HomepagelSCDHEC. https://www.scdhec.gov/
17. Salvaris M, Dean D, Tok WH, et al: Microsoft AI platform, in Deep Learning With Azure. Berkeley, CA, Apress 2018, pp 79-98
18. Blythe M: Tutorial - Access Blob Storage Using Key Vault Using Azure DatabricksIMicrosoft Docs. 2019. https://docs.microsoft.com/en-us/azure/azure-databricks/store-secrets-azure-key-vault
19. Chawla NV, Bowyer KW, Hall LO, et al: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 2002, 321-357, 2002
20. Bock S, Goppold J, Weiß M: An Improvement of the Convergence Proof of the Adam-Optimizer. http://arxiv.org/abs/1804.10587
21. McCready M: GitHub Version Control - Azure DatabricksIMicrosoft Docs. 2020. https://docs.microsoft.com/en-us/azure/databricks/notebooks/github-version-control
22. Introduction—PyGithub 1.45 Documentation. https://pygithub.readthedocs.io/en/latest/introduction.html
23. Yeom S, Giacomelli I, Fredrikson M, et al: Privacy risk in machine learning: Analyzing the connection to overfitting. Proceedings—IEEE Computer Security Foundations Symposium, IEEE Computer Society, 2018, pp 268-282.
24. Ateniese G, Felici G, Mancini LV, et al: Hacking Smart Machines With Smarter Ones: How to Extract Meaningful Data From Machine Learning Classifiers. https://doi.org/10.1504/IJSN.2015.071829
25. Papernot N, Mcdaniel P, Sinha A, et al: SoK: Towards the Science of Security and Privacy in Machine Learning. IEEE European Symposium on Security and Privacy. 2018. https://doi.org/10.1109/EuroSP.2018.00035
26. Hitaj B, Ateniese G, Perez-Cruz F: Deep Models Under the GAN: Information Leakage From Collaborative Deep Learning. http://arxiv.org/abs/1702.07464
27. Shokri R, Stronati M, Song C, et al: Membership inference attacks against machine learning models. Proceedings—IEEE Symposium on Security and Privacy. Institute of Electrical and Electronics Engineers Inc, 2017, pp 3-18.
28. Zhang T, He Z, Lee RB: Privacy-Preserving Machine Learning Through Data Obfuscation. arXiv preprint, 2018. arXiv:1807.01860
29. Xu K, Yue H, Guo L, et al: Privacy-preserving machine learning algorithms for big data systems. Proceedings—International Conference on Distributed Computing Systems. Institute of Electrical and Electronics Engineers· 2015, pp 318-327,
30. Shokri R, Shmatikov V: Privacy-preserving deep learning. CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 2015, 1310-1321.
31. Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al: Differential Privacy-Enabled Federated Learning for Sensitive Health Data. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
32. Choudhury O, Gkoulalas-Divanis A, Salonidis T, et al: Anonymizing Data for Privacy-Preserving Federated Learning. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L
33. Truex S, Steinke T, Baracaldo N, et al: A hybrid approach to privacy-preserving federated learning. Proceedings of the ACM Conference on Computer and Communications Security. Association for Computing Machinery, 2019, pp 1-11.

APPENDIX

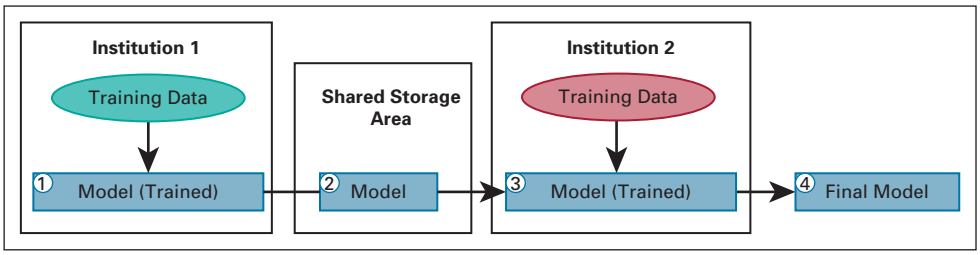

FIG A1. Single weight training mechanism.

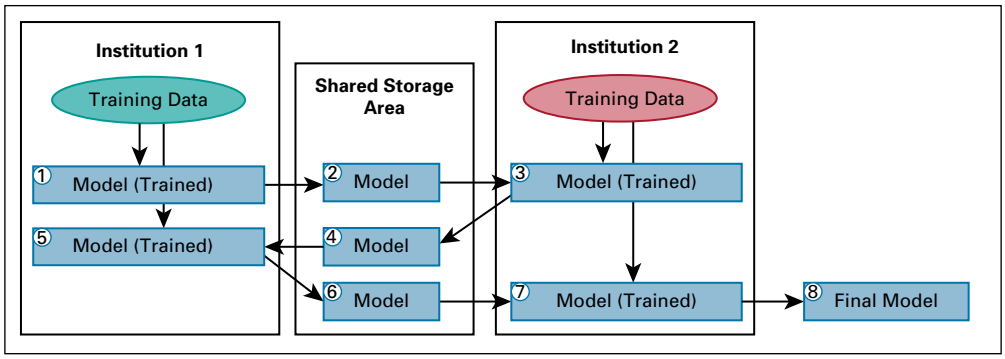

FIG A2. Cyclical weight training mechanism.

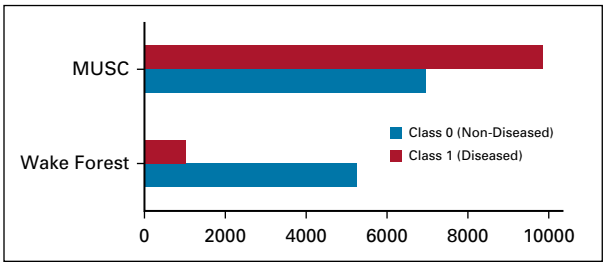

FIG A3. Distributions of each dataset across classes.

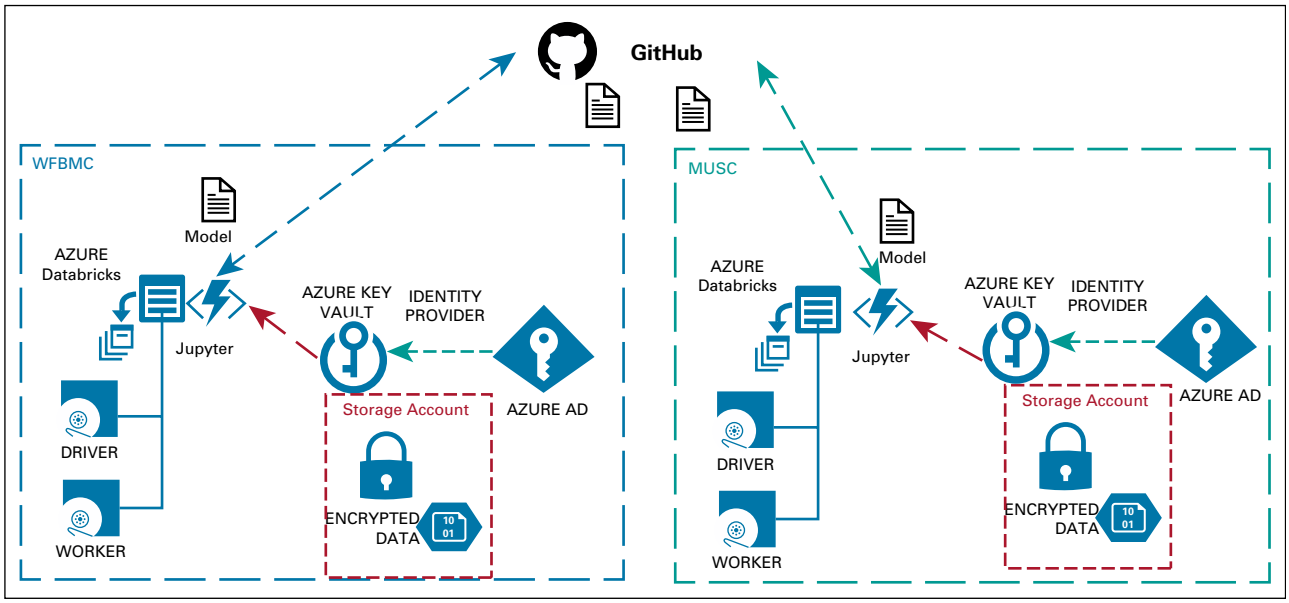

FIG A4. The workflow of the federated learning environment in databricks.

TABLE A1. A Comprehensive Distribution of the Two Institutions' Datasets Across Each Feature

Parameter	Wake Forest	MUSC
Male	2,788	8,580
Female	3,513	8,256
< 29 years old	6	74
30-39 years old	26	260
40-49 years old	216	1,055
50-59 years old	949	3,699
60-69 years old	1,534	5,742
70-79 years old	2,153	4,374
> 80 years old	1,418	1,632
African American	1,007	3,967
Asian	5	46
Native American	11	21
Caucasian	5,254	12,547
Pacific Islander	0	6
Other race	23	187
Refused to state race	1	3
Unknown race	1	59
Nonsmoker	1,044	6,633
Former smoker	3,171	6,905
Current smoker	2,087	3,298
Radon exposure < 2 pCi/L	769	16,157
Radon exposure 2 pCi/L—4pCi/L	3,276	578
Radon exposure > 4pCi/L	2,257	101

Abbreviation: MUSC, Medical University of South Carolina.