Journal of Biomedical Informatics 106 (2020) 103424

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling

Anup Tuladhar[a,b,*], Sascha Gill[b,c], Zahinoor Ismail[b,c,d,e,f], Nils D. Forkert[a,b,c,g], for the Alzheimer's Disease Neuroimaging Initiative

[a] Department of Radiology, University of Calgary, Calgary, Alberta, Canada
[b] Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada
[c] Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada
[d] Department of Community Health Science, University of Calgary, Calgary, Alberta, Canada
[e] Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada
[f] O'Brien Institute for Public Health, University of Calgary, Calgary, Alberta, Canada
[g] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada

ARTICLE INFO

Keywords:
Artificial neural networks
Distributed learning
Machine learning
Medical information systems
Random forest
Support vector machines

ABSTRACT

The development of machine learning solutions in medicine is often hindered by difficulties associated with sharing patient data. Distributed learning aims to train machine learning models locally without requiring data sharing. However, the utility of distributed learning for rare diseases, with only a few training examples at each contributing local center, has not been investigated. The aim of this work was to simulate distributed learning models by ensembling with artificial neural networks (ANN), support vector machines (SVM), and random forests (RF) and evaluate them using four medical datasets. Distributed learning by ensembling locally trained agents improved performance compared to models trained using the data from a single institution, even in cases where only a very few training examples are available per local center. Distributed learning improved when more locally trained models were added to the ensemble. Local class imbalance reduced distributed SVM performance but did not impact distributed RF and ANN classification. Our results suggest that distributed learning by ensembling can be used to train machine learning models without sharing patient data and is suitable to use with small datasets.

1. Introduction

Distributed learning is a comparably novel approach for training machine learning models that circumvents the need for data sharing beyond local health care centers. The approach is particularly appealing for machine learning applications in medicine because the amount of data at a single institution may be inadequate for training accurate and generalizable models [1–3]. This is especially relevant for rare diseases, where very few patients are seen at any single institution [4]. Thus, collaboration between multiple institutions is necessary to train machine learning models on sufficient amounts of data. However, sharing patient data to create a central database for model training (Fig. 1A) is challenging because of legal [5] and ethical concerns [6] about medical data privacy [7–9]. Thus, an alternative to central learning is necessary for healthcare data.

Distributed learning forgoes the need for collecting data in a centralized fashion by performing model training in situ at each local institution (Fig. 1B). These locally trained "agents" only consist of abstracted mathematical parameters and do not contain data from individual patients. After local training, the agents are sent to a central server that combines the local models generated at individual healthcare institutions into a single global model. This global model can leverage insights derived from a greater amount and variety of patient data than any individual model trained at a single institution.

Two main approaches to distributed learning of artificial neural networks (ANN) have been explored in recent years. In federated averaging, multiple ANNs are trained in parallel and the trained parameters are averaged to form a single model [10–13]. In cyclical weight transfer, a single ANN is trained at one institution at a time, moving sequentially between institutions [14]. Although ANNs have become

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; ANN, artificial neural network; ANOVA, analysis of variance; RF, random forest; SVM, support vector machine; SD, standard deviation

[*] Corresponding author at: 3330 Hospital Drive NW, HSC Building, Room 2913, Calgary, AB T2N 4N1, Canada.
E-mail address: anup.tuladhar@ucalgary.ca (A. Tuladhar).

https://doi.org/10.1016/j.jbi.2020.103424
Received 29 October 2019; Received in revised form 16 March 2020; Accepted 17 April 2020
Available online 23 April 2020

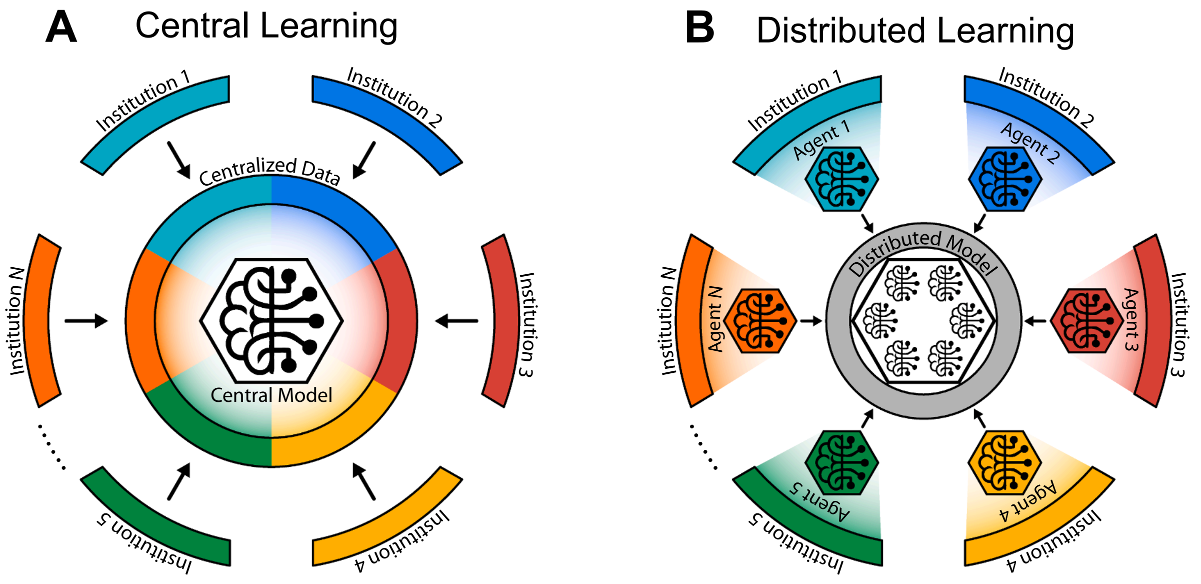

Fig. 1. Distributed learning circumvents the need to share data to train a machine learning model. (A) In central learning, data is collected from a number of institutions into a centralized database. The central model is trained on this centralized database. (B) In distributed learning, model training occurs locally at each institution. Rather than sharing data, the institutions share their locally trained models (agents). The distributed model is constructed from the combination of locally trained agents.

popular due to the improved performance solving various classification problems compared to traditional machine learning techniques [15], it remains an open question if they are the best machine learning classifier for distributed learning. Thus, a detailed comparison with other machine learning classifiers, such as support vector machines (SVM) and random forest (RF), is necessary.

One method to combine local machine learning models (agents) that is independent of the classifier type is to create a global ensemble classifier. Here, all agents are kept independent and each agent makes a prediction on new data, which are combined to one global prediction. However, previous studies evaluating distributed learning with ensembling have used large local datasets [14,16], which may already be sufficient to train accurate models at a single institution. The suitability of this approach in cases of very few training examples per agent has not been explored.

2. Objective

In this study, we investigate distributed learning with ensembling using three types of machine learning classifiers: ANN, SVM, and RF. We compare the classifiers on binary classification tasks in four medical datasets (Table 1). We aimed to investigate the effects of: (1) local dataset size on distributed model performance, particularly for the case of rare diseases with very few examples per institution, (2) the number of collaborating institutions on distributed model performance, and (3) class imbalance at local institutions on distributed model performance. These comparisons are made across all three classifier types and four datasets.

3. Materials

Four medical datasets were used for this work: three collected from the publicly accessible UC Irvine machine learning database (breast cancer, diabetes, and heart disease) and one curated by the Alzheimer's Disease Neuroimaging Initiative (mild cognitive impairment), which were obtained from the ADNI database (http://adni.loni.usc.edu/). The goal of the Alzheimer's Disease Neuroimaging Initiative (ADNI) is to use clinical, neuropsychological, behavioral, genetic, and neuroimaging data to track the progression of Alzheimer's disease.

The classification task for all datasets was to distinguish between patients and healthy subjects. All features in the datasets were normalized using either z-score normalization for continuous values (mean: 0, standard deviation: 1) or min-max normalization (min: 0, max: 1) for discrete values. A summary of each dataset used for simulations is found in Table 1.

4. Methods

All model training and testing was done on Compute Canada and Calcul Quebec computing clusters using Python 3.6.3.

4.1. Model performance evaluation

Model performance was evaluated and reported as the macro-average F1 score. Although the datasets were not balanced overall, the simulated global training data was selected to achieve a balanced class distribution. As a result, the test data was imbalanced (Table 1). To mitigate the effect of imbalanced test data on model performance scores, the macro-average F1 score was chosen to measure performance. Therefore, the F1 score is independently computed for each class in the test set and the unweighted mean of the scores between classes is calculated, *i.e.* the mean of the F1 score for the healthy class and patient class. Thus, unlike accuracy, the macro-average F1 score is not biased by class prevalence in the test set and can be used more intuitively in the case of imbalanced data.

4.2. Classifier design

For each classifier type (ANN, SVM, RF), a number of parameters are available. As we were especially interested in distributed learning performance on small datasets, we determined the classifier parameters that maximized local agent performance when trained on very few (≤ 10) examples across the four datasets. For each dataset, classifier type, and parameter setting (described below), single institution models were evaluated at cases of 2, 4, 8, or 10 training examples. Monte-Carlo cross-validation with 25 iterations was performed for each case, and the global average across iterations and institution sizes was calculated. The parameters that resulted in the best performance, *i.e.* the highest

Table 1
Summary of datasets used.

Disease	Mild cognitive impairment	Breast cancer	Diabetes	Heart disease
Database	Alzheimer's Disease Neuroimaging Initiative	Wisconsin Breast Cancer Diagnostic Dataset	Pima Indians Diabetes Dataset	Cleveland Heart Disease Dataset
Features	Age, sex, years of education, regional brain volumes and surface areas, and regional cortical thicknesses Total: 230	Characteristics of breast mass cell nuclei derived from image (e.g. cell radius, texture, area, smoothness, symmetry) Total: 30	Age, BMI, family history, number of pregnancies, plasma glucose, blood pressure, serum insulin, tricep skin thickness Total: 8	Age, sex, chest pain type, angina, blood pressure, serum cholestoral, blood sugar, heart rate, electrocardiogram results Total: 13
Dataset size	Total: 348 Healthy: 102 (29.3%) Patients: 246 (70.7%)	Total: 569 Healthy: 357 (62.7%) Patients: 212 (37.3%)	Total: 768 Healthy: 500 (65.1%) Patients: 268 (34.9%)	Total: 303 Healthy: 138 (45.5%) Patients: 165 (54.5%)
Global Training Set	Total: 160 Healthy: 80 (50%) Patients: 80 (50%)	Total: 360 Healthy: 180 (50%) Patients: 180 (50%)	Total: 400 Healthy: 200 (50%) Patients: 200 (50%)	Total: 240 Healthy: 120 (50%) Patients: 120 (50%)
Test Set	Total: 188 Healthy: 22 (11.7%) Patients: 166 (88.3%)	Total: 209 Healthy: 177 (84.7%) Patients: 32 (15.3%)	Total: 368 Healthy: 300 (81.5%) Patients: 68 (18.5%)	Total: 63 Healthy: 18 (28.6%) Patients: 45 (71.4%)

macro-average F1 score on the test data, across the four datasets were used for distributed learning (Supplemental Figure S1).

4.2.1. *Artificial neural network models*

Fully connected ANN models were implemented in Keras 2.2.4 with Tensorflow 1.12. The ANNs had three layers: two hidden layers with ReLu units and one sigmoid classification layer. L1 regularization (0.0001) was used. ANN models were trained for 100 epochs with 64 example batch sizes using the "binary cross-entropy" optimizer. We tested ANNs with 16, 128, or 1024 neurons per hidden layer in single institution experiments. The final model used for distributed learning experiments had 128 neurons per hidden layer.

4.2.2. *Support vector machine models*

SVMs were implemented in Scikit-learn 0.21. The C parameter was set to 1.0 and gamma was defined as the reciprocal of the number of features in the dataset. We tested linear, quadratic, cubic, Gaussian, and sigmoid kernels in single institution experiments. The final model used for distributed learning experiments used the sigmoid kernel.

4.2.3. *Random forest models*

RFs were implemented in Scikit-learn 0.21. A balanced RF algorithm was used. We tested tree numbers of 10, 100, 1000, and 10,000 in single institution experiments. The final model used for distributed learning experiments used 1000 trees.

4.3. *Distributed learning*

4.3.1. *Simulating distributed data*

To prevent selection-biased conclusions being drawn from a single split of the data into simulated training and test sets, we performed Monte-Carlo cross-validation, a form of bootstrapping by random sampling with replacement. In each simulation experiment and database, we used 25 iterations of random sub-sampling, with new training and test sets being randomly assigned in each iteration.

To simulate distributed data that is not shared between institutions, we created subsets of training data for each local model by sampling without replacement from the global training set. Thus, each simulated subset was independent. This simulation scheme allowed us to modify the size of local datasets from large (greater than 25% of the total training set) to very small (≤ 10 examples per institution). Additionally, it allowed us to introduce controlled heterogeneity into local datasets, such as non-uniform training set sizes or local class imbalance. The amount of training examples used for simulations is reported in Table 1.

4.3.2. *Distributed model training and testing*

For each classifier type (ANN, SVM, RF), local models ("agents") were trained on these independent subsets, *i.e.* local agents exclusively learned from their own local datasets. The distributed learning model was created by ensembling the predictions of local agents on the test data. For each test example, the local agents made independent predictions on the class probabilities. The distributed model's class probability on the test case is the arithmetic mean of the predicted class probabilities from the local agents. This mean class probability was used to classify the test datasets. For example, in distributed ANNs, a series of ANN agents were trained on the independent local training datasets. Then, for a given test case, each locally trained ANN agent produces a class probability prediction. The arithmetic mean of these class probabilities from the agents is the predicted class probability from the distributed ANN, and the test case is classified using this probability.

The distributed model was compared to central learning and single institution models. Central learning was performed using the entire training dataset, by amalgamating the training data across simulated or real institutions into a single large training set. A single institution model was created by training a model on a single subset of the training data. This effectively simulates the model created by an institution that is not participating in distributed learning, but instead relying solely on its own data. Multiple single institution models were created, and their performance metrics were averaged to estimate single institution performance.

4.4. *Statistics*

Results from simulation experiments are reported as mean + standard deviation (SD), calculated from the results of Monte-Carlo cross-validation iterations. Statistical comparisons were done with Graphpad Prism 8.2 using two-way ANOVAs and Holm-Sidak's post-hoc multiple comparisons test. Statistical significance was set at P-value < 0.01.

5. Results

5.1. *Effect of institution size*

First, we assessed the impact of local dataset size on distributed model performance. Therefore, the local dataset size was iteratively reduced, up to the limit of two examples per local dataset. A central model was trained on the entire training set, while single institution models and distributed learning agents were trained on these progressively smaller local datasets (Fig. 2).

Distributed learning by ensembling statistically improved

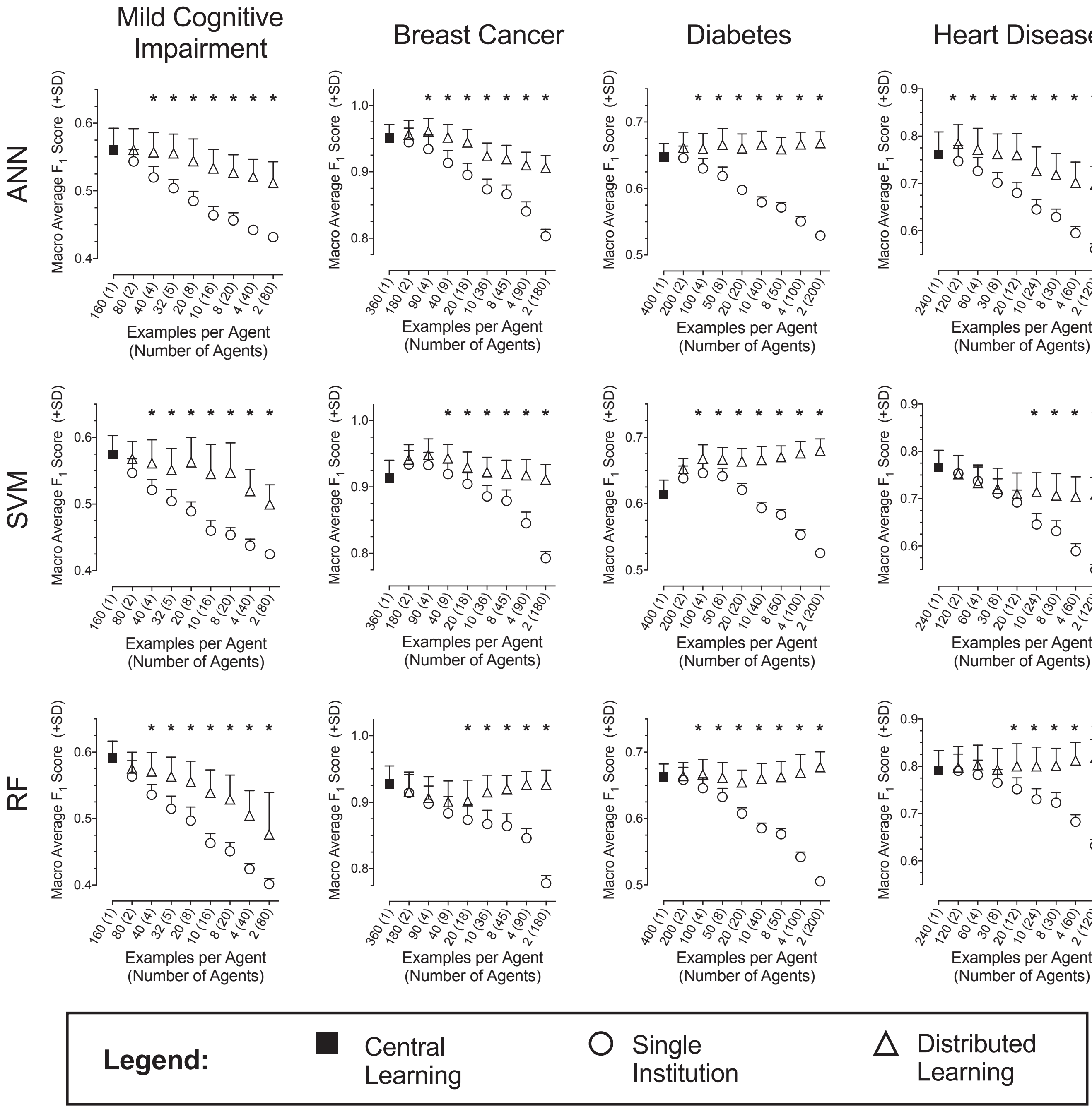

Fig. 2. Distributed learning by ensembling individual agents improves performance compared to individual agents. The performance of artificial neural networks (2 layers with 128 neurons per layer), support vector machines (sigmoid kernel), and random forest (1000 trees) were evaluated on four datasets. Single institution model performance degraded as the size of local training datasets decreased. A distributed learning model, using an ensemble of independently trained agents, significantly increased performance compared to single institution models. Data are expressed as the mean + standard deviation of Monte-Carlo cross-validation runs. Significance was evaluated by a two-way ANOVA, classifier type × examples per agent; $^*P < 0.01$.

performance relative to the average single institution model, for all classifier types (ANN, SVM, RF) and all datasets (two-way ANOVA, classifier type × examples per agent, $P < 0.01$). This effect was most pronounced when local datasets were very small. Even as the average single institution model performance dropped, ensembling agents trained on these very small local datasets resulted in improved performance with distributed learning. While this effect was not as pronounced when local datasets were larger, distributed learning always improved performance over the average single institution model.

In general, distributed model performance decreased relative to the central model as local datasets grew smaller. In the case of small local datasets (2, 4, 8, or 10 examples per institution), distributed RF outperformed distributed ANN and SVM on average across the four datasets (two-way ANOVA, classifier type × database, $P < 0.001$).

As expected, the computational time for training agents at local institutions was reduced as local dataset sizes decreased (Supplemental Figure S2). Generally, SVM models required the shortest training time and RF models required the longest. Inference time for the distributed ensemble models increased proportionally with the number of agents in the ensemble, for all three classifier types, and was generally longer than central learning models (Supplemental Figure S3). Inference times were shortest for distributed SVM and ANN ensembles and longest for

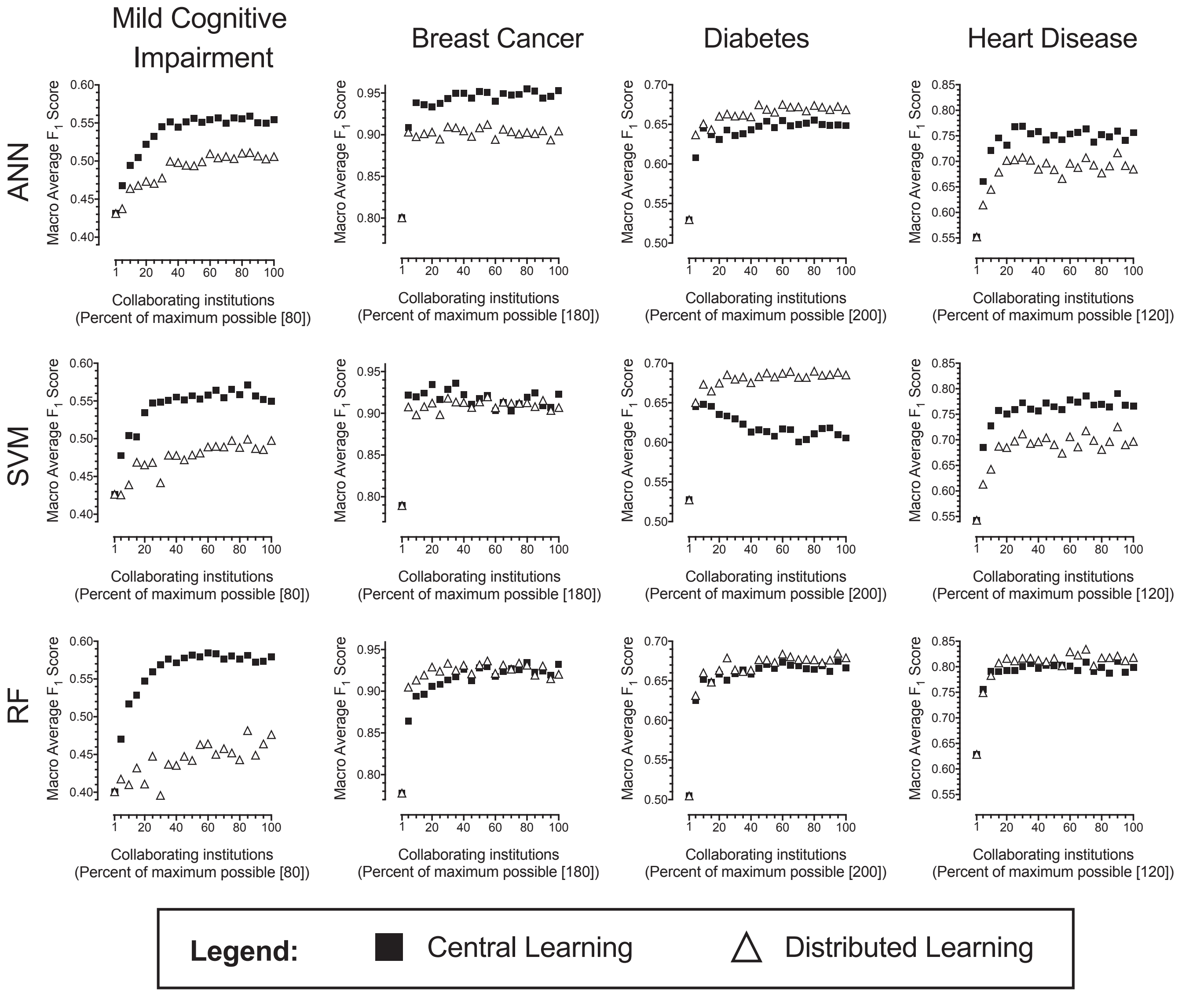

Fig. 3. Distributed learning performance improves with increased collaboration between institutions sharing trained models. Collaboration in central learning is simulated through institutions sharing data into an ever-increasing central database. Distributed learning performance improves for all three classifier types as more institutions contribute locally trained models. Data are expressed as the mean of Monte-Carlo cross-validation runs.

distributed RF ensembles. However, parallelizing and optimizing computations could improve inference computation times.

5.2. *Effect of number of institutions*

Next, we evaluated the effect of the number of collaborating institutions on distributed model performance. In distributed learning, increased collaboration results in more locally trained agents added to the global ensemble. Thus, while local agents see the same limited number of datasets, the distributed model effectively samples from a larger dataset.

As the benefits of distributed learning were most salient in the case of small local datasets, the number of examples per agent were set to two and remained balanced. Starting with a single institution, more collaborating institutions were simulated by progressively adding locally trained agents to the ensemble. As a point of comparison, for each level of collaborating institutions, a central model was simulated by training a single model on the same number of datasets collected into a hypothetical central database.

Across datasets, distributed model performance increased with the number of collaborating institutions (Fig. 3). However, the improvement plateaued, as performance with 30–70% of collaborating institutions was comparable to 100% collaboration. This plateau mirrored the learning curve in central learning, and in general occurred at the same number of collaborating institutions. Conversely, removing a substantial portion of models from the ensemble (30–50%) does not impact performance, demonstrating the stability of the ensembling approach.

5.3. *Effect of unevenly sized training sets at institutions*

We then simulated the scenario of non-uniform training set sizes across institutions on distributed model performance. Given that classifier performance generally improves with more data, we explored whether weighting local agents based on their training set size would improve the classification results of the distributed learning model when local training sets sizes are non-uniform. Again, local training sets remained class-balanced. Keeping the total number of training samples constant, as per Table 1, the size of each agent's local dataset was randomly set at 25% or less of the total training size; thus, local training set sizes were non-uniform.

For non-uniform training set sizes, weighting local agent's input to

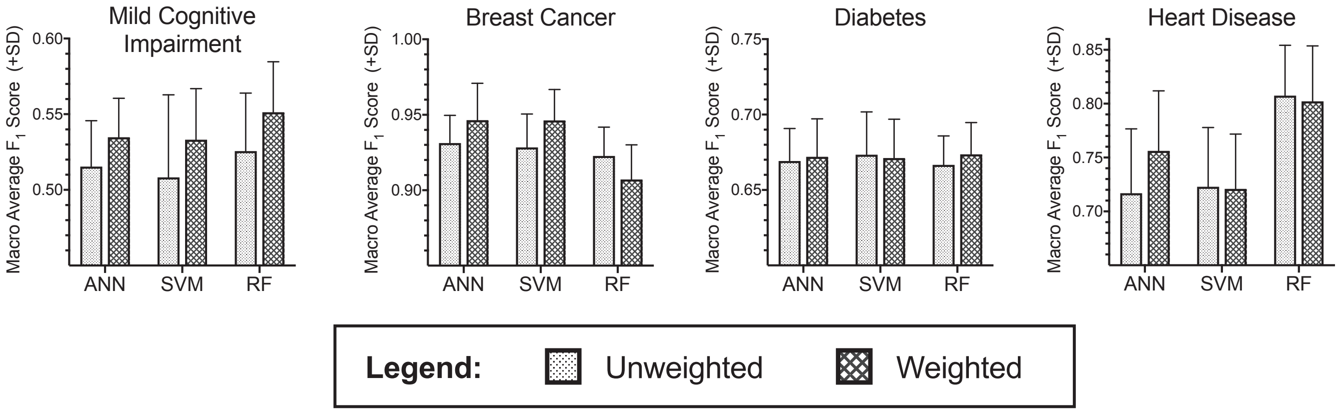

Fig. 4. When simulated institutions used different training set sizes, weighting predictions from local agents based on their training set size improved distributed learning performance compared to unweighted ensembling. Data are expressed as the mean + standard deviation of Monte-Carlo cross-validation runs.

the ensemble based on their training set size generally improved performance compared to the same model without weighting (Fig. 4).

5.4. Effect of imbalanced training sets at institutions

We evaluated the effect of imbalanced training data at local institutions, a more plausible scenario, on distributed model performance. Therefore, the class balance in the global training set was retained while creating class imbalances in the local training subsets, biasing local datasets either toward a majority of healthy examples or a majority of patient examples. In this experiment, the agent dataset size was set to ten while the ratio of majority to minority class examples available to each agent was varied from balanced (5:5) to very imbalanced (9:1).

In the presence of local class imbalance, the average single institution model performed worse compared to the balanced case. Generally, performance of the single institution models worsened with increasing class imbalance (Fig. 5).

Distributed ANN and RF models were not affected by class imbalance and achieved classification results similar to the balanced case in all experiments. However, distributed SVM was negatively impacted by local class imbalance in the mild cognitive impairment, diabetes, and heart disease datasets (Fig. 6).

5.5. Distributed learning on a multi-institutional dataset

Finally, we evaluated distributed learning by ensembling on the actual data distribution of the multi-institutional ADNI dataset. The data used in this study was collected across 72 institutions. However, 29 institutions had to be excluded for this secondary experiment because their local datasets only contained a single class, which prevents proper training of the machine learning models. Of the remaining 43 institutions, 38 institutions were randomly selected for training and the remaining five institutions were used for testing (Table 2). Local agents were trained according to the actual data collected at each institution, and their predictions on the test data were ensembled. This was done for all three classifier types. Central learning was performed on the amalgamation of the data from the 38 training sites.

The multi-institutional ADNI data was non-uniform in local training set sizes (median: 6, interquartile range: 5–8) and imbalanced (median: 66.7% patient examples, interquartile range: 50–77.8%). Unlike the simulated experiments where the global training set was class-balanced, in the multi-institutional ADNI data, the global training set was skewed towards patient examples (66.8%), reflecting the actual data distribution. Though this skew did not negatively impact central learning performance, it did result in poorer distributed learning performance in SVM and RF classifiers, which performed worse than single institution models. Contrary to this finding, distributed ANNs did perform better than single institution models when agents were unweighted, which was further improved by weighting agents by their local training set size (Table 3).

6. Discussion

We demonstrated that popular ANN, RF, and SVM machine learning classifiers can be used for distributed learning, a framework for machine learning without sharing data. Using this framework, institutions

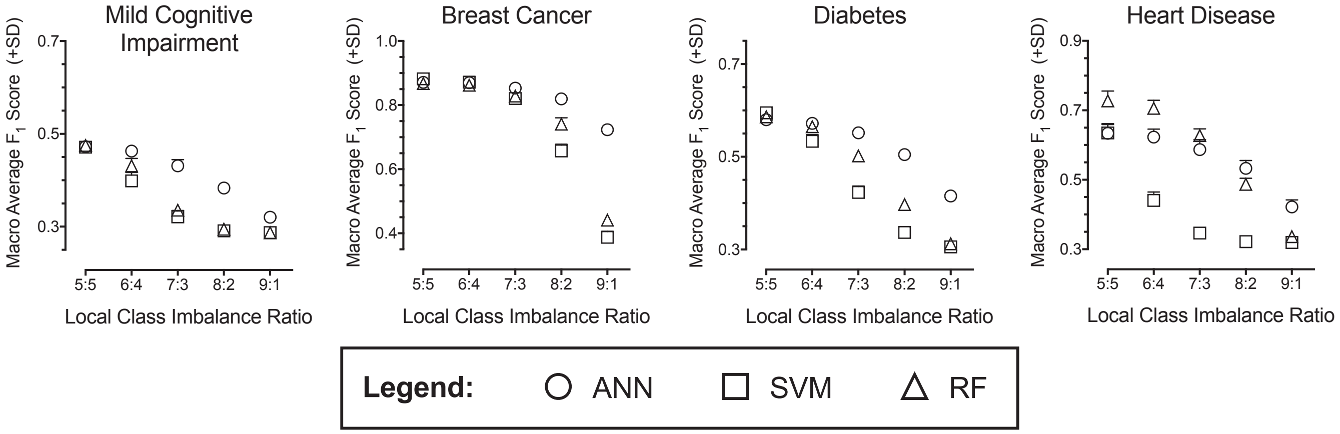

Fig. 5. Single institution models are negatively impacted by local class balance. As the ratio of the majority to minority class in local datasets increased, models trained solely at a single institution perform poorer. Data are expressed as the mean + standard deviation of Monte-Carlo cross-validation runs.

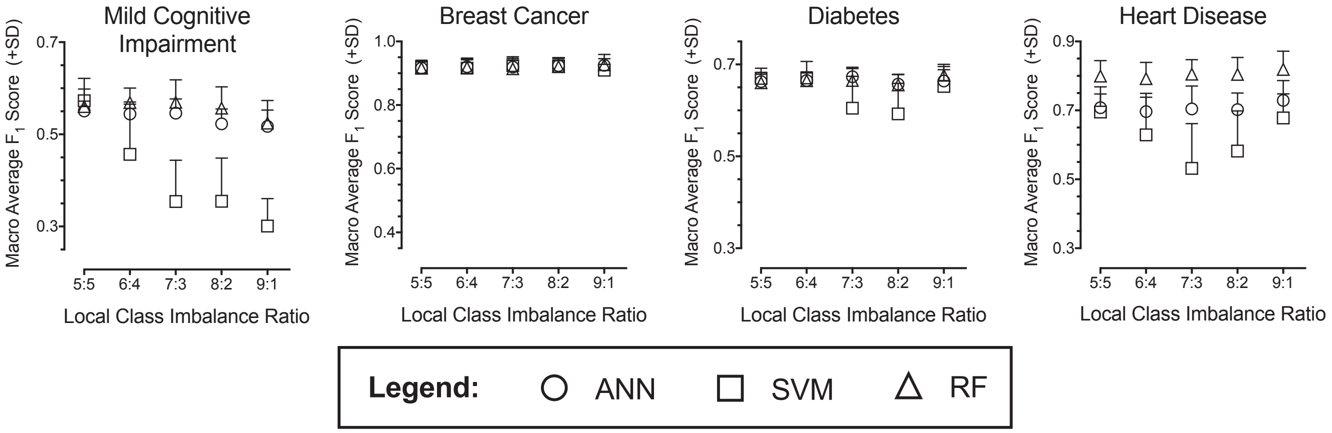

Fig. 6. Distributed learning ANN and RF models are not affected by class imbalance in agent's local datasets. However, class imbalance in agent's local datasets affects distributed SVM performance. Data are expressed as mean + standard deviation of Monte-Carlo cross-validation runs.

Table 2
Statistics of the multi-institutional ADNI dataset.

	Original dataset	Multi-institutional dataset	Training data	Test data
Number of Institutions	72	43	38	5
N across all institutions				
Total	348	283	244	39
Healthy	102	94	81	13
Patients	246	189	163	26
Total examples per institution				
Mean (± SD)	4.83 ± 3.19	6.58 ± 2.81	6.42 ± 2.83	7.80 ± 2.59
Median	4.5	6	6	8
Range (min, max)	1, 14	2, 14	2, 14	4, 11
Interquartile range	2–7	5–8	4.75–8	5.5–10
Healthy examples per institution				
Mean (± SD)	1.42 ± 1.29	2.19 ± 1.03	2.13 ± 1.02	2.60 ± 1.14
Median	1	2	2	3
Range (min, max)	0, 4	1, 4	1, 4	1, 4
Interquartile range	0–2	1–3	1–3	1.5–3.5
Patient examples per institution				
Mean (± SD)	3.42 ± 2.47	4.40 ± 2.40	4.29 ± 2.46	5.20 ± 1.92
Median	3	4	4	5
Range (min, max)	0, 12	1, 12	1, 12	3, 8
Interquartile range	1–5	3–6	2–6	3.5–7
Class balance per institution (as % patient examples)				
Mean (± SD)	70.2 ± 29.5	64.1 ± 15.5	63.7 ± 16.1	67.1 ± 9.85
Median	75	66.7	66.7	72.7
Range (min, max)	0, 100	33.3, 85.7	33.3, 85.7	55.6, 75.0
Interquartile range	56.0–100	50.0–77.8	50–77.8	56.4–75.0

Table 3
Performance of distributed learning on multi-institutional ADNI data subset. Data are the macro-average F1 score on the test set.

Classifier	ANN	SVM	RF
Central learning	0.71	0.69	0.71
Single institution	0.52	0.45	0.46
Distributed learning (unweighted)	0.53	0.40	0.40
Distributed learning (weighted)	0.55	0.40	0.40

can collaborate by sharing models instead of data, while ensembling these locally trained agents generally improves the overall performance.

We have shown that distributed learning is not restricted to a single type of machine learning classifier. Previous works have only evaluated ensembling for training machine learning models using large local datasets containing medical data (120–1500 examples) [14,16]. The limiting case of very small local datasets is particularly interesting for medicine because it simulates the case where institutions may not have previously considered using their data for machine learning because of too small datasets, or the case of rare diseases, where patient populations even at large institutions are very small and may have been understudied [4]. It should be noted that the disease examples used in this work are not considered rare diseases. Due to the problems described above, there are no large enough datasets available for rare diseases that could have been used in this work. Thus, rare diseases had to be simulated by creating small local datasets. However, we believe that the results of this study generally hold true for true rare diseases.

The distributed learning case of training independent models on independent subsets of data used here is conceptually similar to bootstrap aggregation without replacement, an ensemble learning meta-algorithm [17]. This may explain why distributed RF performed best in most test cases. RF models use an ensemble of decision trees trained on subsets of training data. Ensembling RF models trained on distributed data may be similar to training a single large RF model on the hypothetical global training set. However, distributed RF performed much worse compared to centrally trained RF on the challenging mild cognitive impairment task; this was particularly the case with the actual distribution of the multi-institutional ADNI data. Thus, with more challenging datasets, models other than RF may be needed. Additionally, in tasks such as image segmentation, a RF model might not be optimal and alternatives, such as convolutional neural networks [18], will need to be explored for distributed learning.

Using multi-institutional datasets for model training is important because it increases the amount and variety of training data. In central learning, adding more training data generally improves performance, known as a learning curve. We and others [14] show that adding more collaborating institutions mimics this phenomena, as distributed model performance increases as more locally trained agents are added to the global model. This suggests that the distributed model is able to effectively take advantage of an increased amount of training data without actually requiring data sharing. The distributed model approach is also more feasible for healthcare data, as ethical and legal concerns of data privacy in healthcare make the simulated central model approach unlikely to be implemented for many diseases [5–9].

Training models on multi-institution datasets increases the model's generalizability, as it reduces model overfitting to single institution's idiosyncrasies or overcomes limitations, such as local class imbalance. Poor performance of single institution models may be partly attributable to over- or under-fitting of the agents to local datasets, resulting in highly biased models with poor generalizability. However, when sufficient numbers of agents are ensembled, the biases of individual agents are averaged out, which may partly explain the improved classification performance with distributed learning, even in case of very weak classifiers trained using very few samples. While previous studies on ensembling models evaluated local class imbalance, this was limited to imbalance in a single institution [14]. We extended this to multi-institutional data imbalance, where class imbalance is prevalent across all institutions, which may better mimic a real-world scenario for distributed data. We show that while single institution models perform poorly with local class imbalance, distributed ANN and RF models are less affected by local class imbalance and can generalize to unseen data better than distributed SVM models. As local datasets were simulated by sampling without replacement from a single dataset, the variability between institutions may be limited. Therefore, we used a subset of the multi-institutional ADNI database to evaluate distributed learning on a real data distribution, where local institutions vary in size and disease prevalence. Here, distributed ANN outperformed distributed RF and SVM. Correcting for differing sample sizes in local institutions by weighting local agents improved performance. Future work may investigate methods for combining agents that improve ensemble performance, such as accounting for differences in disease prevalence in local agent training datasets or error on the training set. Additionally, the effect of unique patient populations at local institutions, non-identical data distributions, and measurement differences (e.g. different magnetic resonance imaging sequences or hardware) between institutions remains to be investigated.

Ensembling locally trained agents has the advantage of being easy to implement. Existing web platforms for multi-site collaboration and model training could be used to validate and implement this approach [19,20]. As ensembling is theoretically capable of combining multiple types of machine learning classifiers, medical institutions are not required to implement the same classifier and retain the freedom to choose their own model. While cyclical weight transfer [14] and federated averaging [10–13] may offer theoretical performance improvements in the case of few training examples per agent or imbalanced local datasets, as they have the advantage of combining information from all local datasets into a single model, this has yet to be evaluated. Additionally, if a local institution needs to add or remove data from the cyclical or federated distributed model, then the entire model would need to be retrained at all institutions, incurring large computational costs. Here, ensembling offers an advantage over these alternate approaches, as only a single institution would need to retrain its agent to add or remove data from the global distributed model.

For distributed learning by ensembling, inference computational times increased proportionately with the number of agent models from local institutes. However, it bears noting that the computation times depend on various factors, including the hardware used and the exact algorithm implementation. Future work could aim to improve computational efficiency in distributed learning, for example through parallelizing or optimizing computations. Alternatively, locally trained classifiers could be combined into a single global model using model distillation [21], which would reduce computation time at inference.

Future work may extend these studies to other supervised learning problems, such as multi-class classification, regression problems, and image segmentation. Additionally, extensions for distributed learning in unsupervised learning tasks, such as anomaly detection or generative models, may be explored.

7. Conclusions

We demonstrate that ensembling of independently trained models is a promising approach to distributed learning using confidential datasets. Even in situations where local datasets are very small, such as may be the case of rare diseases or widely distributed data, ensembling locally trained agents significantly improves predictive performance compared to a model trained at a single institution with a small number of datasets. Distributed learning performance improves as more locally trained agents are added to the ensemble. When local datasets are class-balanced, ANN, RF, or SVM classifiers may be used to train local agents. However, distributed SVM performs poorer in cases of class-imbalanced local datasets. Distributed RF and ANN tend to be robust in case of imbalanced data at local institutions. Thus, RF or ANN classifiers may be better suited to distributed learning by ensembling. The simplicity of the approach combined with the circumvention of the need to share data may make it easier to develop and deploy machine learning powered solutions in healthcare.

CRediT authorship contribution statement

Anup Tuladhar: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Sascha Gill:** Formal analysis, Data curation, Writing - review & editing. **Zahinoor Ismail:** Formal analysis, Data curation, Writing - review & editing. **Nils D. Forkert:** Conceptualization, Resources, Writing - review & editing, Supervision, Funding acquisition. **:** .

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank members of the Forkert lab and Dr. M. Cecilia Alvarez-Veronesi for thoughtful discussions and feedback on the manuscript.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; MesoScale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work is supported by the Heart and Stroke Foundation of Canada Grant in aid (G-17-0018368), the River Fund at Calgary

Foundation, and the T. Chen Fong Fellowship in Medical Imaging Science to AT.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103424.

References

[1] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, et al., Opportunities and obstacles for deep learning in biology and medicine, J. R. Soc. Interface 15 (2018) 20170387, https://doi.org/10.1098/rsif.2017.0387.

[2] G. Hinton, Deep learning-a technology with the potential to transform health care, JAMA 320 (2018) 1101–1102, https://doi.org/10.1001/jama.2018.11100.

[3] C.D. Naylor, On the prospects for a (Deep) learning health care system, JAMA – J. Am. Med. Assoc. 320 (2018) 1099–1100, https://doi.org/10.1001/jama.2018.11103.

[4] A. Denis, L. Mergaert, C. Fostier, I. Cleemput, S. Simoens, A comparative study of European rare disease and orphan drug markets, Health Policy 97 (2010) 173–179, https://doi.org/10.1016/j.healthpol.2010.05.017.

[5] G.J. Annas, HIPAA regulations - a new era of medical-record privacy? N Engl. J. Med. 348 (2003) 1486–1490, https://doi.org/10.1056/NEJMlim035027.

[6] E. Vayena, A. Blasimme, I.G. Cohen, Machine learning in medicine: Addressing ethical challenges, PLoS Med. 15 (2018) e1002689, , https://doi.org/10.1371/journal.pmed.1002689.

[7] G. Loukides, J.C. Denny, B. Malin, The disclosure of diagnosis codes can breach research participants' privacy, J. Am. Med. Inform. Assoc. 17 (2010) 322–327, https://doi.org/10.1136/jamia.2009.002725.

[8] K. Caine, R. Hanania, Patients want granular privacy control over health information in electronic medical records, J. Am. Med. Inform. Assoc. 20 (2013) 7–15, https://doi.org/10.1136/amiajnl-2012-001023.

[9] W.N. Price, I.G. Cohen, Privacy in the age of medical big data, Nat. Med. 25 (2019) 37–43, https://doi.org/10.1038/s41591-018-0272-7.

[10] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: 2017.

[11] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2019) 1–19, https://doi.org/10.1145/3298981.

[12] T.S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I.C. Paschalidis, W. Shi, Federated learning of predictive models from federated Electronic Health Records, Int. J. Med. Inf. 112 (2018) 59–67, https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[13] L. Huang, A.L. Shea, H. Qian, A. Masurkar, H. Deng, D. Liu, Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records, J. Biomed. Inform. 103291 (2019), https://doi.org/10.1016/j.jbi.2019.103291.

[14] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, et al., Distributed deep learning networks among institutions for medical imaging, J. Am. Med. Inform. Assoc. 25 (2018) 945–954, https://doi.org/10.1093/jamia/ocy017.

[15] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Z. Med. Phys. 29 (2019) 102–127, https://doi.org/10.1016/j.zemedi.2018.11.002.

[16] P. Dluhoš, D. Schwarz, W. Cahn, N. van Haren, R. Kahn, F. Španiel, et al., Multi-center machine learning in imaging psychiatry: A meta-model approach, NeuroImage. 155 (2017) 10–24, https://doi.org/10.1016/j.neuroimage.2017.03.027.

[17] C. Zhang, Y. Ma, Ensemble machine learning: Methods and applications (2012), https://doi.org/10.1007/9781441993267.

[18] K. Kamnitsas, C. Ledig, V.F.J. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, et al., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61–78, https://doi.org/10.1016/j.media.2016.10.004.

[19] J. Que, X. Jiang, L. Ohno-Machado, A collaborative framework for Distributed Privacy-Preserving Support Vector Machine learning, AMIA Annu. Symp. Proc. 2012 (2012) 1350–1359.

[20] D. Meeker, X. Jiang, M.E. Matheny, C. Farcas, M. D'Arcy, L. Pearlman, et al., A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research, J. Am. Med. Inform. Assoc. 22 (2015) 1187–1195, https://doi.org/10.1093/jamia/ocv017.

[21] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, arXiv. stat.ML (2015).