Computers in Biology and Medicine 136 (2021) 104716

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Privacy preserving distributed learning classifiers – Sequential learning with small sets of data

Fadila Zerka [a,b,*], Visara Urovi [c], Fabio Bottari [b], Ralph T.H. Leijenaar [b], Sean Walsh [b], Hanif Gabrani-Juma [b], Martin Gueuning [b], Akshayaa Vaidyanathan [a,b], Wim Vos [b], Mariaelena Occhipinti [b], Henry C. Woodruff [a,d], Michel Dumontier [c], Philippe Lambin [a,d]

[a] The D-Lab, Department of Precision Medicine, GROW – School for Oncology, Maastricht University, Maastricht, the Netherlands
[b] Radiomics (Oncoradiomics SA), Liège, Belgium
[c] Institute of Data Science (IDS), Maastricht University, the Netherlands
[d] Department of Radiology and Nuclear Medicine, Maastricht University Medical Centre+, Maastricht, the Netherlands

ARTICLE INFO

Keywords:
Distributed learning
Sequential learning
Rare disease
Medical data privacy

ABSTRACT

Background: Artificial intelligence (AI) typically requires a significant amount of high-quality data to build reliable models, where gathering enough data within a single institution can be particularly challenging. In this study we investigated the impact of using sequential learning to exploit very small, siloed sets of clinical and imaging data to train AI models. Furthermore, we evaluated the capacity of such models to achieve equivalent performance when compared to models trained with the same data over a single centralized database.
Methods: We propose a privacy preserving distributed learning framework, learning sequentially from each dataset. The framework is applied to three machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), and Perceptron. The models were evaluated using four open-source datasets (Breast cancer, Indian liver, NSCLC-Radiomics dataset, and Stage III NSCLC).
Findings: The proposed framework ensured a comparable predictive performance against a centralized learning approach. Pairwise DeLong tests showed no significant difference between the compared pairs for each dataset.
Interpretation: Distributed learning contributes to preserve medical data privacy. We foresee this technology will increase the number of collaborative opportunities to develop robust AI, becoming the default solution in scenarios where collecting enough data from a single reliable source is logistically impossible. Distributed sequential learning provides privacy persevering means for institutions with small but clinically valuable datasets to collaboratively train predictive AI while preserving the privacy of their patients. Such models perform similarly to models that are built on a larger central dataset.

1. Introduction

The application of artificial intelligence (AI) (i.e., machine/deep learning models) within the clinical decision making process, also referred to as precision medicine, has become a research topic of increasing interest [1,2]. The rising number of published AI models in the literature that support diagnosis/prognosis is a testament to this.

The most common way to train AI models, often referred to as "centralized training", is when the data is sourced from a single centralized database and the training of the classification AI model is local to a single machine. This approach however is not ideal during collaborative efforts where data sharing and centralization is strictly regulated by legal and ethical considerations. For instance, the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) act as safeguards to protect the privacy of patient data. Distributed learning (i.e., federated learning, ensemble learning, or sequential learning) offers a promising solution to this centralization barrier, allowing development and validation of predictive models while preserving the privacy the patient data. Federated learning, the most conventional form of distributed learning, involves a master server that coordinates the initialization and aggregation of learning within a consortium of partners [3,4]. Ensemble learning consists of training independent models on local data, and each model's predictions on new data are grouped to a single global prediction [5].

* Corresponding author. Clos Chanmurly 13, 4000, Liège, Belgium.
E-mail address: fadila.zerka@radiomics.bio (F. Zerka).

https://doi.org/10.1016/j.compbiomed.2021.104716
Received 18 May 2021; Received in revised form 16 July 2021; Accepted 28 July 2021
Available online 31 July 2021

Sequential distributed learning is an extension of distributed learning enabling the partners of a consortium to iteratively update a model with their respective local datasets. The last model in the queue is the final model [6,7]. These approaches are particularly appealing in the cases of small datasets (e.g., low clinical volume or rare diseases) in which the amount of data available to a single center is below the threshold to develop robust and generalizable AI. Since the performance and the robustness of an AI model is directly related to the number of samples on which it was trained and validated [2], the scarcity of data coupled with lengthy procedures required to centralize data can derail initiatives to develop clinical decision support tools.

While distributed learning has been well established with applications in multicentric studies [3,6,8–10], and previous work on ensemble distributed learning on small local datasets has indicated promising performance [11,12], the impact of the network data-scape (e.g., small batch sizes) has yet to be systematically investigated for sequential distributed learning. In this work, we investigate the performance of Stochastic Gradient Descent (SGD) based classifiers trained using a sequential distributed learning approach. We evaluate the influence on model performance when using micro batch sizes (as small as $n = 1$) to replicate cases where a participating institution (or partner) may only provide a single case record to the consortium to support training. To this extent we examine the influence of micro batch sizes on sequential learning model performance compared to the equivalent (i.e., same data) centralized model using a variety of Radiomics and clinical open-source datasets.

2. Background and significance

2.1. Model optimizers - stochastic gradient descent

This work also explores stochastic gradient descent (SGD) which is an iterative optimization method. It is a commonly used optimization technique applied to various machine and deep learning algorithms [13]. Upon each training iteration, the SGD optimizer fine-tunes the algorithm, minimizing the error of the model. As opposed to standard gradient descent optimizers, where the error is reduced over the entirety of the training dataset, SGD randomly selects small training batches and approximates the gradient for the random batch. The iterative process of batch selection is performed by randomly shuffling the dataset and minimizing over all batches, offering the advantage of avoiding local minima and reducing model optimization time.

2.2. Challenges in medical image analysis

Multicentric studies are needed to develop robust AI and to demonstrate the clinical relevance of imaging AI. This kind of studies face many challenges such as:

1) Data collection (described in section “Medical data sharing”);
2) Data heterogeneity, caused by the difference in acquisition and reconstruction settings amongst the different medical centers [14]. To ensure better model building in a heterogeneous domain, the raw data and/or the features derived from it must be harmonized [15,16];
3) And Inter-reader variability, the automation of manual tasks, such as organ and lesion delineation, requires to learn from ground truth masks delineated manually by experienced radiologists [17].The difference in experience and trainings of the clinicians leads to a variation on the ground truth delineations, which in turn represents a challenge in segmentation model training and validation [17].

2.3. Medical data sharing

Despite the efforts made to publicly share medical data in public repositories, including, the cancer imaging archive (TCIA; https://www.cancerimagingarchive.net/), and the NIH BioLINCC (https://biolincc.nhlbi.nih.gov/home/), among others [18], data sharing remains very difficult, especially in low prevalence rare diseases. Within the context of rare-diseases, data sharing limitations can hinder rare disease research and development, as well documented cases may be limited in number. This proves especially difficult in situations where a single institution may want to extract hidden insights using machine learning approaches, such as a diagnostic or prognostic biomarker. Initiatives, such as the European Joint Program on Rare Diseases (EJP RD; https://www.ejprarediseases.org/index.php/about/), began to address this issue and has illustrated the potential of data in driving precision medicine and accelerating rare disease diagnosis/prognosis.

The importance of datatype (e.g., genotype, phenotype and endotype among others) in modeling patients with rare diseases, is well demonstrated within the literature [19,20]. However, de-identification of patient data prior to sharing, does not necessarily guarantee preservation of privacy [21] as patient personal information can potentially be re-identified from the de-identified features (e.g., up to 99.98% of the American population in any dataset can be identified using only 15 demographic features) [22]. This risk increases as the dimensionality of data increases. In order to protect patient sensitive information, data acquisition and sharing is therefore tightly regulated by ethical and legal constraints [23]. In this context, distributed sequential learning is an important approach to facilitate data analysis across institutions while preserving data privacy.

2.4. Distributed learning

Distributed learning was first applied to clinical decision support systems in 2013 [2]. Distributed learning infrastructures enable the efficient training of machine/deep learning models by isolating training data in respective local databases of each collaborative center. Distributed learning can be applied in various forms. In federated learning, each of the collaborators connects to a master server that initializes and updates learning. After initialization, each collaboration center trains a portion of the model on their local data then provides the resulting model weights to the master server. The master server in turn aggregates the weights, updates the model, and shares the updated model weights with the collaborators within the network. Each collaborator then retrains the local models based on the updated weights and sends them back to the master server to close the loop, which operates until a convergence threshold is reached [3,4]. Another form of distributed learning is sequential learning, differing in learning management architecture: 1) learning orchestrated by a cloud server such as the Personal Health Train (PHT; https://www.dtls.nl/fair-data/personal-health-train/) [9,24], or 2) decentralized learning as applied in Chained-Distributed Machines Learning (C-DistriM) [6]. Each iteration in a sequential learning process corresponds to an update of the model from one collaborator. This type of learning is slower when compared to federated learning, where the learning is parallel, but is not subject to the logistical concerns (mainly related to the variation of the internet connection speed across the partners) related to federated learning [25].

In distributed learning data is not visible to the researchers. For this reason, researchers have to rely on the statistical information derived from the local data to build a global model. To reach an optimal performance some modeling steps such as feature selection and inference have to be adapted [4,26,27]. In this regard, the literature has demonstrated both federated learning and distributed learning achieve a comparative performance to traditional centralized learning approaches [4,6,10,28,29].

2.5. Machine learning classifiers

In this work we consider three machine learning classifiers, depicted in Fig. 1, in distributed sequential settings:

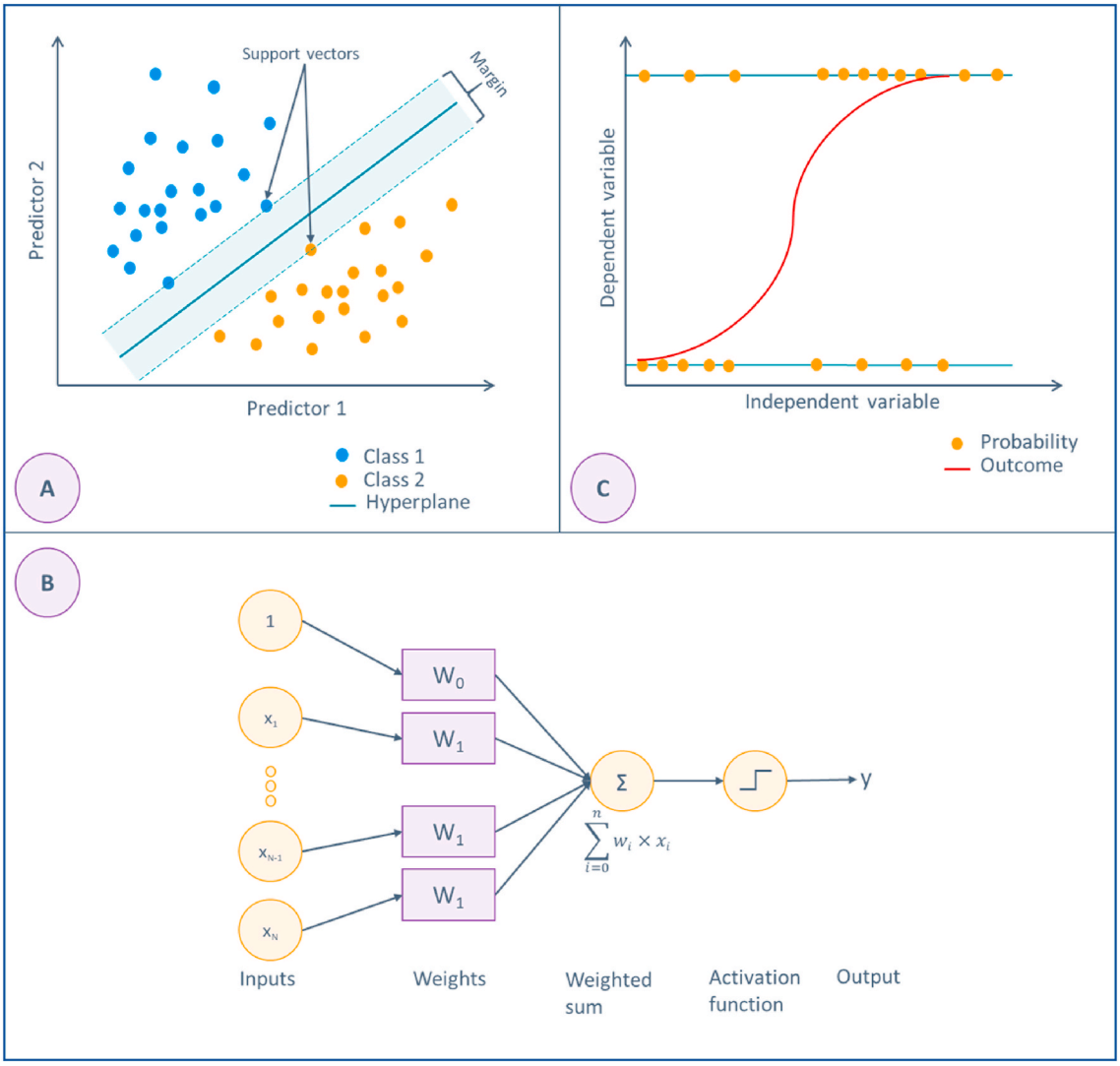

Fig. 1. A) support vector machine, B) logistic regression, C) perceptron.

1. Support vector machines (SVM), a supervised learning algorithm, applied mostly towards classification, but also for regression and the detection of outliers. SVMs work by establishing two parallel hyperplanes, separating the different classes of the feature space. The best fit is established as the one that maximizes the distance between both hyperplanes [30]. To accommodate data variability (i.e., linearly separable or not) various kernels such as linear, radial basis function have been established to optimize the distance between hyperplanes [31]. In this work we applied linear SVM techniques.
2. Logistic regression is a statistical method used for analyzing a feature space in which there are one or more independent variables that identify a predefined outcome. The assumption is that multiple linear regressions of the independent variables are transformed using a logit function to form a conditional probability of the outcome variable. Logistic regression assumes that the feature space possesses a linear relationship with the outcome, making it a linear algorithm with a nonlinear transform [30].
3. Perceptron, a single-layer neural network used for linear classification. The hidden layer mimics the design of a network of neurons within the human brain. Similarly, a perceptron network predicts classifications based on patterns within a series of input features correlating to a specific outcome [30]. The process is as follow: 1) the input features are multiplied by their corresponding weights, that are randomly selected at the first iteration, 2) sum the results of step (1) to generate a weighted sum, 3) calculate the outcome (output) by applying the weighted sum to the activation function, that maps the outcome into values ranging between two predefined values (labels) such as [0,1].

3. Methods

3.1. Data

In this study, four open-source datasets were collected from two different public repositories: the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php) and cancerdata.org (https://www.cancerdata.org/). The characteristics of these datasets are illustrated in Table 1. The datasets include:

1. breast cancer Wisconsin dataset [32].
2. Indian liver dataset [33],
3. NSCLC-Radiomics dataset [34,35],
4. Stage III NSCLC dataset [36].

These datasets were used to train and test the selected machine learning classifiers. Each of these datasets consists of a feature space corresponding to a binary outcome, as illustrated in Table 1. In addition to these four data sets, we extended our analysis to test sequential

Table 1
Dataset characteristics.

Dataset	Disease	Outcome	Patients	Training samples	Validation samples	Test samples	Number of Features (after selection)	Feature type
Breast cancer Wisconsin [32]	Breast cancer	Breast mass malignancy	569	341	114	114	15	clinical
Indian liver [33]	Liver disease	Liver disease	583	349	117	117	7	demographics, clinical
NSCLC-Radiomics dataset [34,35]	Non-small-cell lung cancer	2-year survival	421	224	75	75	4	radiomics
Stage III NSCLC [36]	Stage 3 non-small-cell lung cancer	Overall survival (binary)	548	328	110	110	8	clinical, dosimetric

distributed learning on deep neural networks applied to smaller sets of (MNIST) dataset [37].

The breast cancer Wisconsin dataset consists of features calculated from a digitized image of a fine needle aspirate (FNA) of a breast mass [38], and an outcome defined as "malignant" or "benign". ANOVA test was used to perform select the robust features. 50% of the features were discarded based on ANOVA's F-ratio, reducing the total number of features from 30 to 15.

The Indian liver dataset [33] consists of a set of demographics and clinical features (all patient records were collected from North East of Andhra Pradesh, India) for patients with liver disease. A Pearson pairwise feature correlation was performed. Highly correlated features (i.e., with Pearson correlation coefficient > 0.7) were discarded, reducing the total number of features from 11 to 8. Four patients had missing values corresponding to one feature, the missing data were imputed based on the mean value of the corresponding feature vector.

The NSCLC-Radiomics dataset [34,35] consists of radiomics features extracted, using RadiomiX (Radiomics/Oncoradiomics SA, Liège, Belgium) based on quantitative image analysis technology, from gross tumor volumes (GTV) of standard CT images corresponding to 422 patients. Gross tumor volume segmentations were performed by trained oncologists. Of the 421 records, 44 subjects were discarded during the radiomic features calculation phase. The discarded subjects had GTV segmentations with multiple unconnected volumes. In these cases, signature feature "compactness" cannot be calculated since it is defined for a single volumetric object. The outcome (survival) in NSCLC-Radiomics was converted into two-year survival (binary). New feature selection was not performed, the four predictive features reported in the original study [39] were used.

Data from Ref. [36] is referred to as The Stage III NSCLC dataset. The dataset consists of a combination of clinical, dosimetric features and clinical outcome (survival) for lung cancer patients. Missing data were imputed, using the scikit-learn (version 0.22) imputation transformer. The imputation was based on the mean values of each feature. No feature selection was performed for this dataset, instead predictive features reported in the original study [36] were used to train the models.

As a means to mitigate classifier scaling bias, features in all training sets were independently normalized to the interval [0,1] and the same normalization factor was applied to their respective validation and test sets. The primary objective of this work was to assess model performance variability across unique training scenarios in centralized vs. distributed SGD training approaches. Improving the prediction performance of for models trained with these datasets was out of the scope of this work.

3.2. Experiment design

Three commonly used machine learning (ML) classifiers were selected to conduct this study (Support Vector Machine (SVM), Logistic Regression, and Perceptron). Each classifier satisfies the inclusion criteria:

1. The classifier can be trained in a sequential manner,
2. The classifier has previously been applied and accepted in medical image analysis scientific community [8,40].

The open-source SGDClassifier package (scikit-learn v0.22, Google Summer of Code) in Python (v3.6) was used to implement the selected classifiers [41].

Each dataset was split into training, validation, and test sets (60% training, 20% validation, and 20% testing). Training, validation, and test sets were stratified based on the outcome label to guarantee equal percentage of positive and negative samples on each subset. The validation data was used for hyperparameter tuning and the test set was used evaluate the model performance.

For each dataset, we simulated four training cases:

Centralized: a centralized learning approach where the entirety of the training set was used by a single partner to fit the model – used as the reference for distributed learning approaches.

Case 1. a distributed learning approach composed of 2 partners (2 subsets), randomly distributed between each partner (67% and 33% of the dataset).

Case 2. a distributed learning approach representing an extreme case where each partner contributes with a single datapoint (i.e., from a single patient). In this case the model was updated at each iteration incorporating one additional datapoint.

Case 3. a repeat of Case 2, with the exception of randomly shuffling the dataset to observe the effect the order of the training data (of medical centers) has on the resulting model.

3.3. Optimization of training parameters

3.3.1. Centralized model

For each classifier and dataset pair, we trained a central model and used it as the reference to compare performance of each corresponding distributed model. Hyperparameters tuning was performed for optimal performance. The primary tuned parameter specified the regularization parameter used to calculate learning rate, herein referred to as alpha (α), and number of iterations (epochs). Default values with respect to the classifier were used for remaining parameters such as tolerance, and penalty. To tune the hyperparameters we defined a set of alpha values ranging between $1e^{-7}$ and 1, as illustrated in Fig. 2. For each classifier:

1. The validation set performance (determined by the area under the curve (AUC) of the Receiver Operating Characteristic curve (ROC)), was estimated for each parameter α.
2. The resulting models were compared based on their performances.
3. The best α parameter was then selected according to the model comparison outcome.
4. Each classifier was subsequently retrained using the best corresponding α parameter.
5. Finally, the performance of the global model was then evaluated against the appropriate test set.

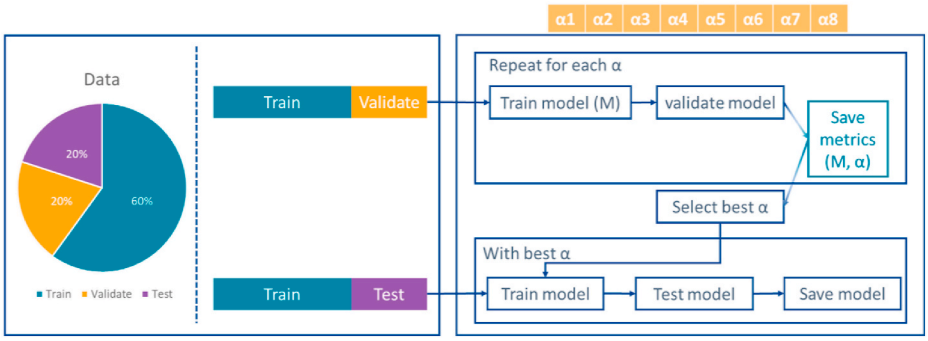

Fig. 2. Model training design: 1) each dataset is divided into training, validation, and testing sets; 2) train a model and validate it for each α; 3) compare the AUCs corresponding to each α; 4) select the value of α that returned the best AUC; 5) train the main model using the selected α, 6) test the model using the test data, 7) save the model.

3.4. Distributed learning models

Hyperparameters tuning was also performed for distributed learning cases, as illustrated in Fig. 3. Model optimization was performed over 5 key steps:

1. For each simulated partner, estimate the validation set performance (AUC) corresponding to each parameter α.
2. Compare the resulting models based on their performances.
3. Select the best α parameter according to the model comparison outcome.
4. Retrain each classifier using the best corresponding α parameter in a sequential manner.
5. Finally, the performance of the global model was then evaluated against the appropriate test set.

Finally a pairwise comparison of the final aggregated models AUC values corresponding to each classifier and dataset was performed using DeLong tests [42].

To consider the impact of shuffling the local training datasets, and their size on model performance in cases where partners have multiple datapoints each, we extended the experiments conducted in this study. To sufficiently realize these experiments, we sourced the Modified National Institute of Standards and Technology (MNIST) dataset [37], a commonly used large dataset suited to test deep neural networks. Data description, the different data splits, model architecture, and results are available in Supplementary Materials (Section A1).

4. Results

4.1. Results based on dataset

The combination of 4 datasets, training cases and 3 model architectures resulted in 48 uniquely trained models. Table 2 summaries model performance for each architecture and training use case reported as the AUC and a 95% confidence interval (CI). Models trained with the breast cancer dataset outperformed models trained with other datasets in all model architectures. The Indian dataset had notably better performance in specific training use cases and model architectures when compared to classification performance for either NSCLC dataset. Logistic regression and perceptron architectures had improved performance over SVM for classification in either NSCLC dataset.

4.2. Results based on training use case

Fig. 4 depicts the AUC values corresponding to each study case for each pair of classifier and dataset. For each classifier, the derived AUC values per use case (centralized, Case 1, Case 2, Case 3) trended with a high degree of similarity but were not identical. Shuffling and local dataset size variability produced observable differences in the ROC curves. However, the Pairwise DeLong tests [42] were used to compare the ROC curves for each of the training scenarios and found no statistically significant differences (p-values > 0.05), as summarized in Table 3 organized by classifier and training dataset. Furthermore, each model trained in a distributed fashion did not differ significantly from

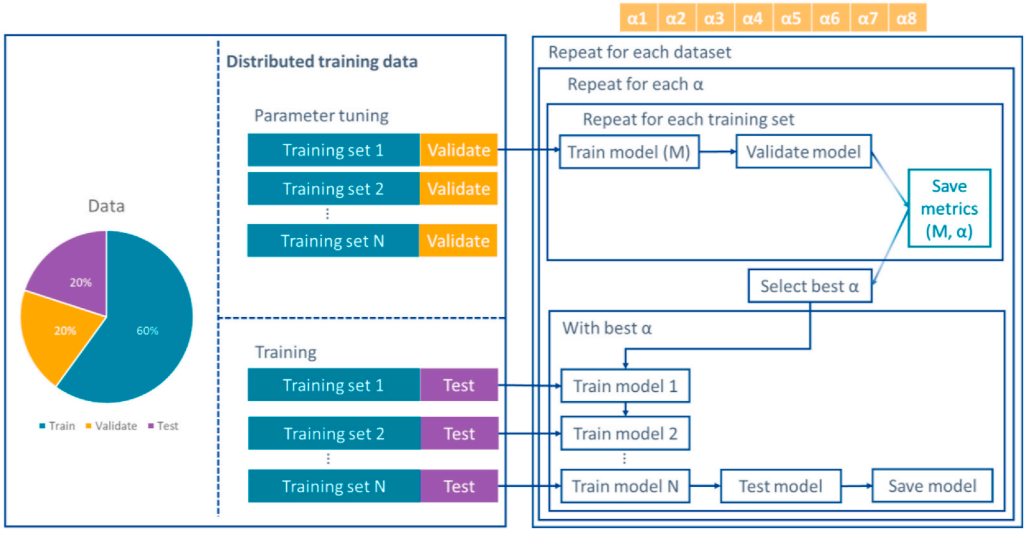

Fig. 3. Distributed model training design: 1) each dataset is divided into training, validation, and testing sets; 2) each training set is split into distributed sets; 3) train a model for each distributed dataset and validate it for each α; 5) compare the AUCs corresponding to each α; 6) select the value of α that returned the best AUC; 7) train the main model using in a sequential manner across the distributed datasets using the selected α, 8) evaluate all the distributed models using the same test, 9) save the final model in the queue as the main model.

Table 2
Discrimination performance (AUC) obtained by training centralized and distributed classifiers (SVM, logistic regression, and Perceptron) using four different datasets (Breast cancer, Indian Liver, NSCLC-Radiomics dataset, and Stage III NSCLC).

Classifier	*Training scenario*	*AUC (95% CI)*			
		Breast cancer	Indian Liver	NSCLC-Radiomics dataset	Stage III NSCLC
Support vector machine	Centralized	0.99 (0.98–1)	0.76 (0.68–0.85)	0.64 (0.51–0.77)	0.64 (0.48–0.79)
	Case 1	0.99 (0.99–1)	0.77 (0.69–0.86)	0.64 (0.51–0.77)	0.61 (0.46–0.75)
	Case 2	0.98 (0.98–1)	0.74 (0.65–0.83)	0.65 (0.52–0.77)	0.60 (0.46–0.76)
	Case 3	0.98 (0.97–1)	0.75 (0.67–0.84)	0.62 (0.49–0.75)	0.61 (0.46–0.76)
Logistic Regression	Centralized	0.98 (0.98–1)	0.76 (0.67–0.84)	0.72 (0.61–0.84)	0.70 (0.57–0.82)
	Case 1	0.97 (0.95–0.99)	0.76 (0.67–0.85)	0.71 (0.59–0.82)	0.69 (0.56–0.82)
	Case 2	0.97 (0.94–0.99)	0.73 (0.64–0.82)	0.72 (0.61–0.84)	0.65 (0.52–0.78)
	Case 3	0.99 (0.98–1)	0.74 (0.65–0.83)	0.70 (0.58–0.82)	0.67 (0.55–0.79)
Perceptron	Centralized	0.99 (0.98–1)	0.78 (0.70–0.86)	0.72 (0.61–0.84)	0.70 (0.55–0.84)
	Case 1	0.98 (0.96–1)	0.76 (0.68–0.85)	0.68 (0.55–0.80)	0.69 (0.56–0.82)
	Case 2	0.99 (0.98–1)	0.74 (0.65–0.83)	0.67 (0.54–0.79)	0.67 (0.53–0.81)
	Case 3	0.99 (0.98–1)	0.78 (0.70–0.86)	0.66 (0.54–0.79)	0.69 (0.56–0.81)

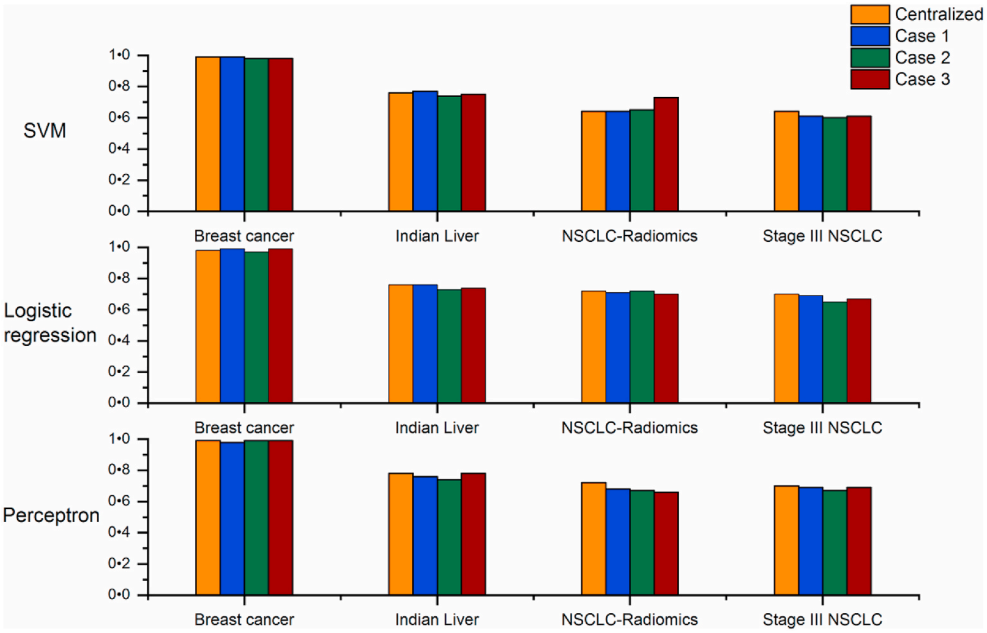

Fig. 4. AUC of each classifier and dataset pair, for each dataset four different models have been trained (Centralized, Case 1: two partners, Case 2: each center holds one patient, Case 3: each center holds one patient with a shuffle in the order of the partners while training).

the reference centralized trained model (p-values > 0.05). These results were validated on a CNN classification model using the MNIST dataset. Detailed results of the MNIST experiments are presented in Supplementary Materials (Section A). ROC curves corresponding to each training scenario and dataset is reported in the Supplementary Materials (Section B).

4.3. Results based on classifier architecture

In most cases the average absolute difference in the AUC values were blow 5%. The average difference in the AUC of the centralized training over the Breast cancer, Indian Liver, NSCLC-Radiomics dataset, Stage III NSCLC datasets was reported as 0.67%, 1.75%, 8.33%, and 6.24%, respectively. Differences in the AUC for each training scenario versus each classifier has been summarized in Table 1, Supplementary Materials (Section C). The maximum average difference of the AUC values for the distributed learning classifiers per dataset increases up to 8.74%, 7.66%, and 8.64% for case, Case 2, Case 3, respectively. It should be noted that in extreme cases certain scenarios had AUC differences above 10%, highlighted in Table 1 of the Supplementary Materials (Section C). These results suggest the optimal classifier chosen is highly dependent on the characteristics of the dataset.

5. Discussion and future work

High quality datasets with sufficient training datasets are required for machine learning models to converge and generalize [2]. When working with patient data, there are important ethical and legal considerations to be managed, when considering sharing patient data between institutions.

The results presented in this work demonstrate that sequential distributed learning on small, isolated datasets (including extreme cases of model updated using a single datapoint at a time) achieves equivalent performance to models trained in conventional centralized learning. Similar conclusions were observed in the case of multiclass classification using the MNIST dataset [37]. We observed, by applying a pairwise DeLong [42] comparison, that the AUC for distributed learning models do not differ with statistical significance from models trained in centralized scenarios.

The results in Tables 2 and 3 and the ROC curves indicate that there is a difference in the performance of different classifiers, and this difference can vary from one dataset to another. We noted that the average

Table 3
p-values corresponding to the pairwise Delong test.

Model	Test	Dataset			
		Breast cancer	Indian Liver	NSCLC-Radiomics dataset	Stage III NSCLC
		p-value	**p-value**	**p-value**	**p-value**
Support vector machine	Central model vs Case 1	0.155	0.594	0.926	0.810
	Central model vs Case 2	0.785	0.475	0.871	0.718
	Central model vs Case 3	0.082	0.861	0.53	0.760
	Case 1 vs Case 2	0.143	0.429	0.969	0.953
	Case 1 vs Case 3	0.071	0.710	0.685	0.980
	Case 2 vs Case 3	0.271	0.580	0.234	0.941
Logistic Regression	Central model vs Case 1	0.061	1	0.422	0.606
	Central model vs Case 2	0.076	0.250	0.196	0.541
	Central model vs Case 3	0.904	0.538	0.373	0.663
	Case 1 vs Case 2	0.425	0.224	0.401	0.652
	Case 1 vs Case 3	0.052	0.652	0.823	0.787
	Case 2 vs Case 3	0.066	0.677	0.285	0.762
Perceptron	Central model vs Case 1	0.147	0.448	0.062	0.600
	Central model vs Case 2	0.427	0.267	0.332	0.809
	Central model vs Case 3	0.370	0.858	0.210	0.662
	Case 1 vs Case 2	0.322	0.274	0.838	0.868
	Case 1 vs Case 3	0.329	0.212	0.719	0.936
	Case 2 vs Case 3	0.946	0.129	0.848	0.856

AUC difference between the classifiers can increase up to 8.33%, 8.74%, 7.66%, and 7.66% with respect to each use case (centralized, Case 1, Case 2, Case 3) and dataset (Breast cancer, Indian Liver, NSCLC-Radiomics dataset, Stage III NSCLC). Even though this margin may be perceived as inconsequential, the clinical risk of decisions based on predictions must be considered as with all changes in model performance. Cases with AUC differences above our 10% threshold (highlighted in red), indicate that this specific classifier is suboptimal for the dataset in question. Thus, with respect to learning [43], we recommend to select the classifier based on comparative performance of the different centralized and distributed classifiers, or base the selection justified criteria related to the data characteristics that will be used to fit the model.

Previous reports on distributed ensemble learning [11,12], showed the potential of application of this particular type of distributed learning to small siloed datasets. For example, Tuladhar et al. [12] reports that grouping models learned locally from either artificial neural network, SVM, or random forest could efficiently exploit small sets of data to build global models. These results suggest that the application of ensemble learning on small dataset is feasible. While other authors have demonstrated that grouping local logistic regression models, is promising in the case of small datasets [11]. In addition to that, they proposed a model update based on the distributed sets of data information to improve the global model performance on small datasets. Results of these studies [11,12] showed an overall improvement in global model performance compared to models trained in a single institution data. These results, however, cannot be extended to the case of distributed sequential learning.

Our results demonstrate that sequential distributed learning can be beneficial for the application of AI for outcome prediction in favor of medical institutes holding very small datasets. Practical examples of small datasets can be 1) pediatric cases that tend to suffer from small sample sizes [44], 2) early phase clinical trials where the sample size tend to have around 20 subjects [45], and 3) rare diseases as they have a very low prevalence (<5/10000 in the European population) [46], making it nearly impossible for a single medical center to collect enough data to train machine learning models. Even with these limitations, and with considerably small datasets (20–100 datapoints), researchers have been using machine learning to build diagnosis and prognosis models for rare diseases [47]. The generalizability of trained models is directly related to the quality and quantity of the training data [48]. In this regard, distributed learning provides opportunity to develop generalizable models with small high-quality datasets in multi-center applications while also mitigating the need to share data and maintaining the privacy of all patient information, such as imaging, genomic, or clinical insight.

Batch size is well known to have an effect on final model performance [49], where evidence suggests that large batch size does not always relate to better model performances [50]. Conversely, in distributed learning applications, 1) a smaller batch size has been linked to the privacy of the training data, as it considerably reduces the ability to reproduce training data from shared model weights in case of weights leakage [51], 2) it has been well documented that the order of training partners in a distributed network influences the performance of the model [28]. Our results suggest that the centralized and distributed models are not statistically different. Therefore, we see distributed sequential learning as a viable tool for multicentric precision medicine studies, particularly in applications with small datasets such as rare diseases and could also be applied in pediatrics and early phase clinical trials.

The tuning of each classifier prior training of the final model is an essential step in achieving robust and generalizable models as this is dependent on the nature of data used in training. The need for tuning hyperparameters stems from the fact that the classifiers investigated in this work are using SGD as an optimizer, and thus cannot avoid this optimization step. The main parameter that needs to be optimized is the learning rate; as it controls the manner in which model is modified according to the estimated error at every iteration/update of the model weights. The process of learning rate selection is challenging as a small value theoretically facilitates better performances but in contrast can increase training phase time significantly. On the other hand, a larger learning rate value can result in an unstable training phase, as the model updates very quickly in each iteration causing it to converge to a flat (i.e., less optimal) minima.

Tuning model hyperparameters is also imperative for distributed classifiers, as we showed in this study. Furthermore, we observed that there is a need to investigate different combinations of hyperparameters and number of iterations. Hyperparameter and training settings such as number of iterations, coupled with a set of model selection criteria (based for example on a comparison of model accuracies or model parameters) [52], can be beneficial to reduce the risk of overfitting. However, this leads in turn to one limitation, related to the longer execution time in comparison to traditional centralized training. This increase in time is accounted by the need to investigate all the training data across all the participating partners to set optimal hyperparameters. In addition to this, it is important to consider

communication costs, as all parameter tuning and model updates occur over the internet. Characterizing the time required for training is challenging as the duration is highly dependent on each partner's internet bandwidth. Future directions of this work will include analysis to characterize the model training duration in a distributed fashion, identify the scalability of the infrastructure to accommodate larger loads by increasing available computational power either though scale up (additional hardware) or scale out (additional nodes) and investigate the elasticity or ability to dynamically handle varying loads of data.

6. Conclusion

This study demonstrates 1) the proof-of-concept of sequential distributed learning applied on small sizes of data, narrowed down to a single datapoint at a time 2) the opportunities associated with this type of distributed infrastructures on the application of AI in low prevalence diseases. We simulated three different distributed learning cases using three classifiers and four different datasets. Our results indicate that sequentially training the models using (extremely) small datasets delivers statistically similar performance (p-values > 0.05) in comparison to the conventional centralized approach. This work provides a validation of the potential of distributed learning in case of small datasets and a new opportunity to data driven outcome modeling in rare disease research. Furthermore, this work can be used to continuously update predictive models as new data is available. Finally, future work is planned to estimate and optimize the scalability of sequential distributed learning infrastructures in real world settings.

Funding

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno). This research is also supported by the Dutch Technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, Aspasia NWO (grant n°91716421) and the Technology Program of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL - n° 673780), EUROSTARS (DART - n° E10116, DECIDE - n° E11541), the European Program PREDICT - ITN - n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 "CLEARLY" - n° UM 2017–8295), Interreg V-A Euregio Meuse-Rhine ("Euradiomics" - n° EMR4), DRAGON (Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement n° 101005122), EuCanImage (European Union Horizon 2020 research and innovation program under grant agreement n° 952103), and DEEP-MAM (Eurostar grant n° E12931).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fadila Zerka, Akshayaa Vaidyanathan, Fabio Bottari, Martin Gueuning, Hanif Gabrani-Juma, Mariaelena Occhipinti are salaried employees/ receive renumeration from Radiomics (Oncoradiomics SA). Dr Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Radiomics (Oncoradiomics SA), ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Radiomics (Oncoradiomics SA), BHV, Merck, Varian, Elekta, ptTheragnostic and Convert pharmaceuticals. Dr Lambin has shares in the company Radiomics (Oncoradiomics SA), Convert pharmaceuticals SA and The Medical Cloud Company SPRL and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics (Oncoradiomics SA) and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics (Oncoradiomics SA) and Health Innovation Ventures and three non-issues, non licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889. Ralph T.H. Leijenaar has shares in the company Radiomics (Oncoradiomics SA) and is co-inventor of an issued patent with royalties on radiomics (PCT/NL2014/050728) licensed to Radiomics (Oncoradiomics SA). Sean Walsh and Wim Vos have shares in the company Radiomics (Oncoradiomics SA). Michel Dumontier has shares in The Medical Cloud Company SPRL. Rest of the co-authors have no known competing financial interests or personal relationships to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104716.

References

[1] P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, E. Roelofs, W. van Elmpt, P.C. Boutros, P. Granone, V. Valentini, A.C. Begg, D. De Ruysscher, A. Dekker, Predicting outcomes in radiation oncology—multifactorial decision support systems, Nat. Rev. Clin. Oncol. 10 (2013) 27–40, https://doi.org/10.1038/nrclinonc.2012.196.

[2] P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – an approach towards decision support systems enabling customised radiotherapy', Radiother. Oncol. 109 (2013) 159–164, https://doi.org/10.1016/j.radonc.2013.07.007.

[3] T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology 4 (2017) 24–31, https://doi.org/10.1016/j.ctro.2016.12.004.

[4] M. Bogowicz, A. Jochems, T.M. Deist, S. Tanadini-Lang, S.H. Huang, B. Chan, J. N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R.H. Brakenhoff, I. Nauta, S.E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebers, F.W.R. Wesseling, S. Keek, S. Sanduleanu, R.T.H. Leijenaar, M. R. Vergeer, C.R. Leemans, C.H.J. Terhaard, M.W.M. van den Brekel, O. Hamming-Vrieze, M.A. van der Heijden, H.M. Elhalawani, C.D. Fuller, M. Guckenberger, P. Lambin, Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer, Sci. Rep. 10 (2020) 4542, https://doi.org/10.1038/s41598-020-61297-4.

[5] M. Kirienko, M. Sollini, G. Ninatti, D. Loiacono, E. Giacomello, N. Gozzi, F. Amigoni, L. Mainardi, P.L. Lanzi, A. Chiti, Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI, Eur. J. Nucl. Med. Mol. Imag. (2021), https://doi.org/10.1007/s00259-021-05339-7.

[6] F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (C-DistriM), IEEE Access 8 (2020) 183939–183951, https://doi.org/10.1109/ACCESS.2020.3029445.

[7] S. Lugan, P. Desbordes, L.X.R. Tormo, A. Legay, B. Macq, Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn), ArXiv:1906.07690 [Cs, Stat], 2019, http://arxiv.org/abs/1906.07690. (Accessed 30 August 2019).

[8] F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T.H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic review of privacy-preserving distributed machine learning from federated databases in health care, JCO Clinical Cancer Informatics (2020) 184–200, https://doi.org/10.1200/CCI.19.00047.

[9] T.M. Deist, F.J.W.M. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, A. Dekker, Distributed learning on 20 000+ lung cancer patients – the Personal Health Train, Radiother. Oncol. 144 (2020) 189–200, https://doi.org/10.1016/j.radonc.2019.11.019.

[10] A. Jochems, T.M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, C. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin, A. Dekker, Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries, Int. J. Radiat. Oncol. Biol. Phys. 99 (2017) 344–352, https://doi.org/10.1016/j.ijrobp.2017.04.021.

[11] T.-T. Kuo, J. Kim, R.A. Gabriel, Privacy-preserving model learning on a blockchain network-of-networks, J. Am. Med. Inf. Assoc. 27 (2020) 343–354, https://doi.org/10.1093/jamia/ocz214.

[12] A. Tuladhar, S. Gill, Z. Ismail, N.D. Forkert, Building machine learning models without sharing patient data: a simulation-based analysis of distributed learning by ensembling, J. Biomed. Inf. 106 (2020) 103424, https://doi.org/10.1016/j.jbi.2020.103424.

[13] N. Ketkar, Stochastic gradient descent, in: N. Ketkar (Ed.), Deep Learning with Python: A Hands-On Introduction, Apress, Berkeley, CA, 2017, pp. 113–132, https://doi.org/10.1007/978-1-4842-2766-4_8.

[14] J. Weese, C. Lorenz, Four challenges in medical image analysis from an industrial perspective, Med. Image Anal. 33 (2016) 44–49, https://doi.org/10.1016/j.media.2016.06.023.

[15] G. Vegas-Sánchez-Ferrero, M.J. Ledesma-Carbayo, G.R. Washko, R. San José Estépar, Harmonization of chest CT scans for different doses and reconstruction methods, Med. Phys. 46 (2019) 3117–3132, https://doi.org/10.1002/mp.13578.

[16] R. Da-ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J.F. Rameé, O. Pradier, U. Schick, D. Visvikis, M. Hatt, Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies, Sci. Rep. 10 (2020) 10248, https://doi.org/10.1038/s41598-020-66110-w.

[17] Medical image computing and computer assisted intervention – MICCAI 2018: 21st international conference, granada, Spain, september 16-20, in: A.F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Proceedings, Part I, Springer International Publishing, Cham, 2018, https://doi.org/10.1007/978-3-030-00928-1.

[18] R. Banzi, S. Canham, W. Kuchinke, K. Krleza-Jeric, J. Demotes-Mainard, C. Ohmann, Evaluation of repositories for sharing individual-participant data from clinical studies, Trials 20 (2019) 169, https://doi.org/10.1186/s13063-019-3253-3.

[19] C. Faviez, X. Chen, N. Garcelon, A. Neuraz, B. Knebelmann, R. Salomon, S. Lyonnet, S. Saunier, A. Burgun, Diagnosis support systems for rare diseases: a scoping review, Orphanet J. Rare Dis. 15 (2020) 94, https://doi.org/10.1186/s13023-020-01374-z.

[20] E. Turro, W.J. Astle, K. Megy, S. Gräf, D. Greene, O. Shamardina, H.L. Allen, A. Sanchis-Juan, M. Frontini, C. Thys, J. Stephens, R. Mapeta, O.S. Burren, K. Downes, M. Haimel, S. Tuna, S.V.V. Deevi, T.J. Aitman, D.L. Bennett, P. Calleja, K. Carss, M.J. Caulfield, P.F. Chinnery, P.H. Dixon, D.P. Gale, R. James, A. Koziell, M.A. Laffan, A.P. Levine, E.R. Maher, H.S. Markus, J. Morales, N.W. Morrell, A. D. Mumford, E. Ormondroyd, S. Rankin, A. Rendon, S. Richardson, I. Roberts, N.B. A. Roy, M.A. Saleem, K.G.C. Smith, H. Stark, R.Y.Y. Tan, A.C. Themistocleous, A. J. Thrasher, H. Watkins, A.R. Webster, M.R. Wilkins, C. Williamson, J. Whitworth, S. Humphray, D.R. Bentley, N. Kingston, N. Walker, J.R. Bradley, S. Ashford, C. J. Penkett, K. Freson, K.E. Stirrups, F.L. Raymond, W.H. Ouwehand, Whole-genome sequencing of patients with rare diseases in a national health system, Nature 583 (2020) 96–102, https://doi.org/10.1038/s41586-020-2434-2.

[21] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference, Science 339 (2013) 321–324, https://doi.org/10.1126/science.1229566.

[22] L. Rocher, J.M. Hendrickx, Y.-A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nat. Commun. 10 (2019) 3069, https://doi.org/10.1038/s41467-019-10933-3.

[23] W. Xia, Z. Wan, Z. Yin, J. Gaupp, Y. Liu, E.W. Clayton, M. Kantarcioglu, Y. Vorobeychik, B.A. Malin, It's all in the timing: calibrating temporal penalties for biomedical data sharing, J. Am. Med. Inf. Assoc. 25 (2018) 25–31, https://doi.org/10.1093/jamia/ocx101.

[24] Personal Health Train, Dutch techcentre for life sciences. (n.d.). https://www.dtls.nl/fair-data/personal-health-train/(accessed January 28, 2021).

[25] S. Park, Y. Suh, J. Lee, FedPSO, Federated learning using particle swarm optimization to reduce communication costs, Sensors 21 (2021), https://doi.org/10.3390/s21020600.

[26] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Distributed feature selection: an application to microarray data classification, Appl. Soft Comput. 30 (2015) 136–150, https://doi.org/10.1016/j.asoc.2015.01.035.

[27] S.M. Ayyad, A.I. Saleh, L.M. Labib, A new distributed feature selection technique for classifying gene expression data, Int. J. Biomath. (IJB) 12 (2019) 1950039, https://doi.org/10.1142/S1793524519500396.

[28] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, J. Kalpathy-Cramer, Distributed deep learning networks among institutions for medical imaging, J. Am. Med. Inf. Assoc. 25 (2018) 945–954, https://doi.org/10.1093/jamia/ocy017.

[29] A. Choudhury, S. Theophanous, P.-I. Lønne, R. Samuel, M.G. Guren, M. Berbee, P. Brown, J. Lilley, J. van Soest, A. Dekker, A. Gilbert, E. Malinen, L. Wee, A. L. Appelt, Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – a proof-of-concept study, Radiother. Oncol. 159 (2021) 183–189, https://doi.org/10.1016/j.radonc.2021.03.013.

[30] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, New York, NY, 2013, https://doi.org/10.1007/978-1-4614-6849-3.

[31] M. Achirul Nanda, K. Boro Seminar, D. Nandika, A. Maddu, A comparison study of kernel functions in the support vector machine and its application for termite detection, Information 9 (2018) 5, https://doi.org/10.3390/info9010005.

[32] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, (n.d.). https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (accessed January 6, 2021).

[33] Ilpd (Indian liver patient dataset) - dataset by uci, Data.World. (n.d.). https://data.world/uci/ilpd-indian-liver-patient-dataset (accessed January 6, 2021).

[34] H.J.W.L. Aerts, E. Rios Velazquez, R.T.H. Leijenaar, P. Chintan, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R. Barkovich, P. Lambin, Data from NSCLC-Radiomics [Data Set], 2019, https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[35] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imag. 26 (2013) 1045–1057, https://doi.org/10.1007/s10278-013-9622-7.

[36] C. Oberije, D. De Ruysscher, R. Houben, M. van de Heuvel, W. Uyterlinde, J. O. Deasy, J. Belderbos, A.-M.C. Dingemans, A. Rimner, S. Din, P. Lambin, A validated prediction model for overall survival from stage III non-small cell lung cancer: toward survival prediction for individual patients, Int. J. Radiat. Oncol. Biol. Phys. 92 (2015) 935–944, https://doi.org/10.1016/j.ijrobp.2015.02.048.

[37] Y. Lecun, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 47.

[38] K.P. Bennett, O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optim. Methods Software 1 (1992) 23–34, https://doi.org/10.1080/10556789208805504.

[39] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006, https://doi.org/10.1038/ncomms5006.

[40] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Methodol. 19 (2019) 64, https://doi.org/10.1186/s12874-019-0681-4.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[42] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (1988) 837, https://doi.org/10.2307/2531595.

[43] T.M. Deist, F.J.W.M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, L. Wee, O. Morin, D. R. Raleigh, W. Bots, J.H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, P. Lambin, Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers, Med. Phys. 45 (2018) 3449–3459, https://doi.org/10.1002/mp.12967.

[44] V. Rahimzadeh, C. Schickhardt, B.M. Knoppers, K. Sénécal, D.F. Vears, C. V. Fernandez, S. Pfister, S. Plon, S. Terry, J. Williams, M.S. Williams, M. Cornel, J. M. Friedman, Key implications of data sharing in pediatric genomics, JAMA Pediatr 172 (2018) 476, https://doi.org/10.1001/jamapediatrics.2017.5500.

[45] Phase I Trials - an Overview | ScienceDirect Topics, (n.d.). https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/phase-i-trials (accessed July 7, 2021).

[46] D. Taruscio, L. Vittozzi, A. Rocchetti, P. Torreri, L. Ferrari, The occurrence of 275 rare diseases and 47 rare disease groups in Italy. Results from the national registry of rare diseases, Int. J. Environ. Res. Publ. Health 15 (2018), https://doi.org/10.3390/ijerph15071470.

[47] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, S. Thun, The use of machine learning in rare diseases: a scoping review, Orphanet J. Rare Dis. 15 (2020) 145, https://doi.org/10.1186/s13023-020-01424-6.

[48] T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in radiation therapy: challenges and opportunities, Br. J. Radiol. 90 (2017) 20160689, https://doi.org/10.1259/bjr.20160689.

[49] X. Qian, D. Klabjan, The Impact of the Mini-Batch Size on the Variance of Gradients in Stochastic Gradient Descent, ArXiv:2004.13146 [Cs, Math], 2020, http://arxiv.org/abs/2004.13146. (Accessed 18 January 2021).

[50] I. Kandel, M. Castelli, The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset, ICT Express 6 (2020) 312–315, https://doi.org/10.1016/j.icte.2020.04.010.

[51] M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16, 2016, pp. 308–318, https://doi.org/10.1145/2976749.2978318.

[52] M.W. Browne, Cross-Validation Methods | Elsevier Enhanced Reader, (n.d.). https://doi.org/10.1006/jmps.1999.1279.