Bioinformatics, 38(8), 2022, 2278–2286
https://doi.org/10.1093/bioinformatics/btac065
Advance Access Publication Date: 9 February 2022
Original Paper

Data and text mining

Federated Random Forests can improve local performance of predictive models for various healthcare applications

Anne-Christin Hauschild [1,*], Marta Lemanczyk[1,†], Julian Matschinske[2,3], Tobias Frisch[4], Olga Zolotareva[2], Andreas Holzinger [5], Jan Baumbach[3] and Dominik Heider [1,*]

[1]Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, [2]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising-Weihenstephan, Germany, [3]Computational Systems Biology, University of Hamburg, Hamburg, Germany, [4]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark and [5]Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität Graz, Graz, Austria

*To whom correspondence should be addressed.

[†]Present address: Hasso-Plattner-Institut for Digital Engineering, University of Potsdam, Potsdam, Germany

Associate Editor: Jonathan Wren

Received on November 4, 2021; revised on January 8, 2022; editorial decision on January 28, 2022; accepted on February 1, 2022

Abstract

Motivation: Limited data access has hindered the field of precision medicine from exploring its full potential, e.g. concerning machine learning and privacy and data protection rules.
Our study evaluates the efficacy of federated Random Forests (FRF) models, focusing particularly on the heterogeneity within and between datasets. We addressed three common challenges: (i) number of parties, (ii) sizes of datasets and (iii) imbalanced phenotypes, evaluated on five biomedical datasets.
Results: The FRF outperformed the average local models and performed comparably to the data-centralized models trained on the entire data. With an increasing number of models and decreasing dataset size, the performance of local models decreases drastically. The FRF, however, do not decrease significantly. When combining datasets of different sizes, the FRF vastly improve compared to the average local models. We demonstrate that the FRF remain more robust and outperform the local models by analyzing different class-imbalances.
Our results support that FRF overcome boundaries of clinical research and enables collaborations across institutes without violating privacy or legal regulations. Clinicians benefit from a vast collection of unbiased data aggregated from different geographic locations, demographics and other varying factors. They can build more generalizable models to make better clinical decisions, which will have relevance, especially for patients in rural areas and rare or geographically uncommon diseases, enabling personalized treatment. In combination with secure multi-party computation, federated learning has the power to revolutionize clinical practice by increasing the accuracy and robustness of healthcare AI and thus paving the way for precision medicine.
Availability and implementation: The implementation of the federated random forests can be found at https://featurecloud.ai/.
Contact: dominik.heider@uni-marburg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The digital revolution in healthcare, fostered by novel high-throughput technologies and electronic health records (EHRs), transitions the field toward a big data era (Constable *et al.*, 2015). Many studies have proven machine learning (ML) to be advantageous for disease diagnosis, prognosis and monitoring of diseases (Fatima and Pasha, 2017). In cancer research, for instance, ML is used to gain deeper insights and understanding of the genetic alterations that are required for cells to develop various stages and severity of cancers (Batra *et al.*, 2017; Jeanquartier *et al.*, 2016; Park *et al.*, 2021a; Wiwie *et al.*, 2019) and thereby enable tailored prognoses and

monitoring of diseases. Moreover, computational models on clinical variables and EHRs are used to assess individualized health risks, for instance, to identify high-risk patients for sepsis in intensive care units (Calvert *et al.*, 2019) or the analysis of longitudinal data for the early detection of heart failure (Zhao *et al.*, 2019). In particular, the combination of big data and artificial intelligence (AI) offers new opportunities to transform healthcare toward precision medicine. Given data for large patient cohorts, we can learn computational models that can predict medical phenotypes such as disease or treatment outcome and extract relevant features (biomarkers), e.g. from expression data. The PAM50 and MammaPrint gene signature panels are examples that aim to include the most pertinent breast cancer marker (Laenkholm *et al.*, 2018; Slodkowska and Ross, 2009). They are currently used as medical diagnostics tools for breast cancer subtyping, guiding individualized breast cancer treatment worldwide. However, both panels are based on small sample sizes (<5k) and a large number of genes (>20k). Thus, studies raised concerns regarding the predictive clinical value of such gene panels (Bösl *et al.*, 2017). The main bottleneck in many studies is the small number of samples compared to many features. This so-called small-n-large-p problem results in a computational issue termed the 'curse of dimensionality'. The Cancer Genome Atlas (TCGA) is the by far most comprehensive repository for clinical cancer omics data worldwide (Weinstein *et al.*, 2013). It contains whole-genome gene expression data for almost five thousand breast cancer patient samples linked to clinical outcomes. However, these few thousand samples stand against more than 19 thousand features that an AI may pick and combine to predict the outcome. Moreover, this small number of patients is unexpected since, in the European Union (EU) alone, there are about 350 thousand new breast cancer cases per year [International Agency for Research on Cancer (IARC)]. Due to the small sample size, there is a risk for model overfitting and a significantly reduced robustness for medical diagnostics.

Moreover, systematic biases within clinical trials, in particular toward white Western participants, have led to medical treatments that were not generally suitable for all ethnic groups (Schork, 2015). As the above breast cancer examples illustrate, modern omics technologies generate massive amounts of data. In addition, a variety of regularization-based methods, e.g. ridge regression or lasso, aim to address small-sample size issues. However, few studies were robustly replicated under heterogeneous clinical conditions, and thus, solely fractions of their results are utilized as prognostic and predictive markers in clinical practice. Big data in healthcare is clearly in its infancy, even in fields such as oncology that are most advanced in omics and one of the best-researched areas of precision medicine.

The aggregation of clinical data including omics and EHRs across institutes, nation- and global-wide, could address these previous limitations of sample size and systematic biases and subsequently move the field toward more accurate precision medicine. However, a global exchange harbors risks to data safety of sensitive patient information and EHRss stored in critical healthcare infrastructure. Essentially data exchange amongst institutions over the internet is posing a roadblock hampering big-data-based medical innovations.

In 2016, the EU passed the General Data Protection Regulation (GDPR) which sets rules for storing and sharing data in and outside the EU (The Council of the European Union, 2016). One main statement of the GDPR is the protection of a person's identity such that it cannot be traced back by third parties directly or even indirectly. Furthermore, anonymization is often not sufficient since a person's identity could be revealed by a singling out (a process of elimination) or through unique combinations of attribute values (Sweeney *et al.*, 2013). Moreover, keeping data centralized, on a shared server or a cloud, for instance, increases the risk of cyberattacks. Thus, existing biomedical information is aggregated in separate data silos restricting access to data and hindering it to exploit its full potential by ML (Rieke *et al.*, 2020). Nevertheless, according to the GDPR, a 'protection by design' technology can be used if it ensures personal privacy at all times. A system that meets these requirements needs to combine advanced distributed architectures such as federated learning (FL) and methods for secure multi-party computation (SMPC) such as differential privacy or homomorphic encryption. The development of the latter has been thoroughly studied (Fang and Qian, 2021; Zapechnikov, 2020). Consequently, we focus on the performance evaluation of FL methods compared to the local and centralized model and address different challenges of biomedical datasets in various clinically relevant applications.

1.1 Federated learning

FL techniques are categorized into horizontal FL (overlapping feature space), vertical FL (overlapping sample space), as well as federated transfer learning (neither feature nor sample space are overlapping) (Park *et al.*, 2021b; Yang *et al.*, 2019b). This study focuses on horizontal FL and aims to overcome the barrier of exchanging raw patient data and move toward large-scale medical data mining. Which, in combination with SMPC methods, can minimize cyber risk. These techniques seek to build a generalized global model without access to a shared dataset (Gan *et al.*, 2017) and therefore require a fundamentally different architecture (see Fig. 1). In the past decade, a variety of federated algorithms have been developed for a multitude of applications (Yang *et al.*, 2019a). Several linear methods, such as distributed regression, were developed. Here an encrypted posterior distribution of coefficients updates a global model (Sundhar Ram *et al.*, 2012; Wang *et al.*, 2013). Moreover, Nasirigerdeh *et al.* (2020) developed sPLINK, a federated GWAS tool. Other approaches implement distributed ensemble learning methods, such as federated decision trees (Strecht *et al.*, 2014), distributed boosting such as SecureBoost framework (Cheng *et al.*, 2019; Lazarevic and Obradovic, 2001) or federated Random Forests (FRF) like FederatedForest (Liu *et al.*, 2019) to mention just a few. Subsequently, federated architectures are rapidly integrated into research and commercial areas such as on mobile applications to minimize data traffic (Konečný *et al.*, 2016a,b; McMahan *et al.*, 2016).

1.2 Related work in clinical research

Until now, very few studies applied FL to medical and health scenarios. The first analyses focused on federated linear approaches. For instance, Lorenzi *et al.* (2017) implemented a multi-centric, sequential and meta-partial least squares approach to model associations between genetic markers and anatomical surface features in Alzheimer's Disease.

Other studies and initiatives use federated approaches, for example, to find clinically similar patients (Lee *et al.*, 2018), predicting hospitalizations due to cardiac events (Brisimi *et al.*, 2018), ICU stay time or mortality (Roy *et al.*, 2019). More recently, the sPLINK tool enables federated genome-wide association studies as a robust alternative to meta-analysis (Nasirigerdeh *et al.*, 2020). The FedHealth, a federated transfer learning framework, has been developed for wearable healthcare devices (Chen *et al.*, 2020). Most recently, a FRF was applied in a collaborative clinical research network to model effective prognosis prediction (Li *et al.*, 2020). Moreover, newly established consortia, such as the FeatureCloud initiative (https://featurecloud.eu), the German Cancer Consortium's Joint Imaging Platform (https://dktk.dkfz.de/en) or the Medical Institutions Collaborate to Improve Mammogram Assessment AI, aim to enable decentralized research across medical and research institutions by using FL architectures.

While different FL algorithms have been designed and applied to clinical data in pilot studies, these rarely examine the various challenges frequently emerging in biomedical dataset analyses. Medical data is somewhat different in many aspects from other branches in data mining. In particular, the heterogeneity within and between medical datasets regarding ethical, legal or social confounders, but also imbalances with phenotype prevalence or cohort sizes (Cios and Moore, 2002).

1.3 Approach

Here we use one of the most typical representatives of federated ensemble learning (FEL) algorithms, the FRF, to public data, treated as

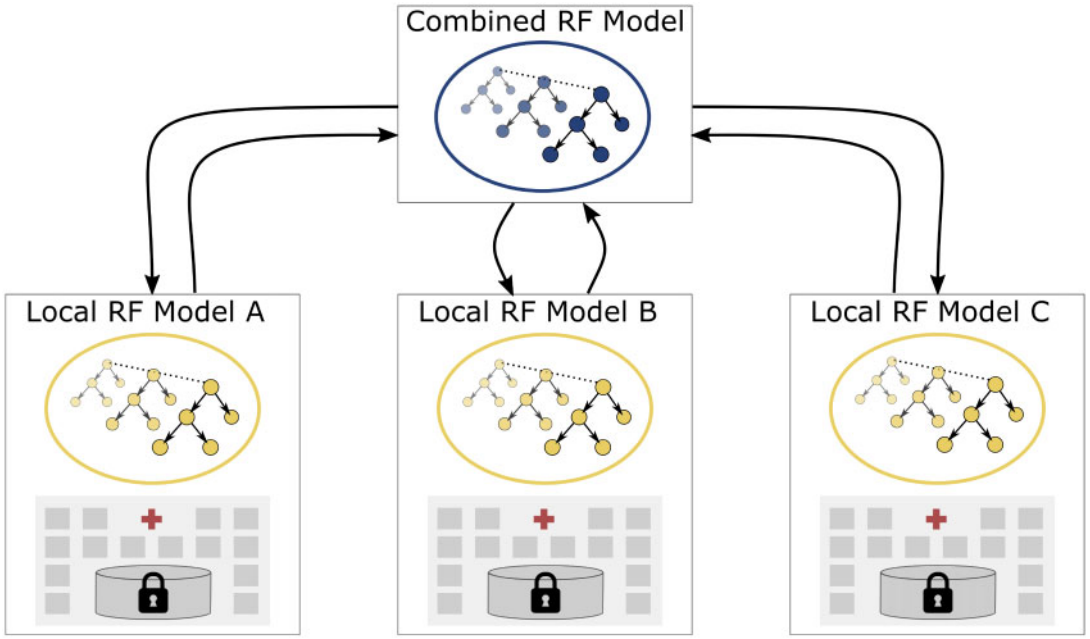

Fig. 1. Principle of FRF deployed in clinical research and practice

confidential and distributed, to evaluate whether horizontal FL can outperform the local models and has the same predictive and prognostic power as classical, data-centralized approaches. In particular, we focus on challenges such as the heterogeneity within and among datasets. Moreover, we emulated three common challenging scenarios among clinical research, (i) a different number of participating parties and their datasets; (ii) different sizes of datasets; and (iii) different class imbalances within datasets, and evaluated the corresponding model performances.

1. **Different number of participants or sites contributing to the global model:** Depending on the available patients and resources, a single study can sometimes only obtain a small number of samples. Sharing data among multiple collaboration partners can obtain a larger dataset and potentially a better model. Thus, it is critical to evaluate whether we can improve the local performance when using a combined federated model for prediction, particularly when the data is distributed amongst an increasing number of sites.
2. **Different number of samples per site:** Depending on the available patients and resources, silos may have a varying number of samples among the contributed datasets and subsequently differences in the predictive power of trained submodels. Therefore, it is crucial to investigate whether and to what extend the combined federated model can compensate for the dataset imbalance and if a size-dependent weighting is beneficial.
3. **Different balances in the predicted phenotype:** An additional challenge in medical data science are biases, i.e. an unevenly distributed target phenotype within the datasets. Thus, it is essential to analyze how such biases can affect the performance of the local and combined model.

2 Materials and methods

2.1 Datasets

For the performance analysis of FRF we used five different clinical [Indian liver patient data (ILPD), hepatocellular carcinoma (HCC)] and biomedical [breast cancer diagnosis (BCD), lung tumor diagnosis (LTD)] datasets, as well as a cross-silo [breast cancer dataset from TCGA (BCTCGA)] dataset. To cover a wide variety of clinically relevant applications, we included different input and output variables, such as disease classification, subtyping, as well as survival as target or output and a variety of predictive or input variables, such as clinical data, laboratory, imaging, as well as genetics data was used. (See Table 1 for number of samples, features and class imbalance.) The following paragraphs give a short description of these public datasets.

2.1.1 ILPD

The ILPD set includes data from 583 liver patients (Ramana *et al.*, 2012). Experts classified the positive instances as patients with a liver disease. The features are clinical measurements as well as age and sex.

2.1.2 HCC

All patients suffer from chronic liver diseases. Positive instances are patients who were diagnosed with HCC (Best *et al.*, 2016). The dataset consists of clinical and biometric features.

2.1.3 BCD

The BCD dataset was retrieved from the hospital of the University of Wisconsin (Wolberg and Mangasarian, 1990). The predictive features were collected from a digital image of a fine needle aspirate of a breast mass. Characteristics for each cell nucleus in the images are measured. The dependent variable is the categorization of breast cancer in benign and malignant (positive class) tumors.

2.1.4 LTD

The LTD dataset GSE30219 consists of gene expression data of lung tumor patients (Rousseaux *et al.*, 2013). A patient is classified as positive if the survival time was higher than 30 months.

2.1.5 BCTCGA

As a cross-silo example for the application of FL algorithms we selected a Breast Cancer dataset from TCGA (BCTCGA). The dataset contains expression profiles of human breast tumors of patients from the TCGA-BRCA cohort (Liu *et al.*, 2018) originating from 19 different institutes. While the breast cancer samples classify into 4–6 subtypes, for evaluation purposes, we focus on a binary

Table 1. Datasets used to evaluate the performance of FL methods

Name	No. of samples	Disease	Classes	No. of variables
ILPD	583	Liver disease	Positive: 416/negative: 167	10 numeric
HCC	685	Hepatocellular carcinoma	Positive: 282/negative: 403	7 numeric
BCD	569	Breast cancer	Malignant: 212/benign: 357	30 numeric
LTD	293	Lung tumor	Survival time > 30 month: 178/<30 month: 115	22 600 numeric
BCTCGA	1069 (19 sites)	Breast cancer	Luminal A: 494/Other subtypes: 575	20 500 numeric

classification of the most frequent subtype luminal A versus other subtypes. Since the data originate from different institutes, we will use this dataset to evaluate the performance of FRF in a cross-silo setting.

2.2 Federated Random Forests

In this study, we use FRF, one of the most common representatives of FEL algorithms. RF models are particularly suited for these scenarios. On the one hand, they have been widely used and proven to be very efficient in accurately modeling biomedical data for various tasks (Boulesteix *et al.*, 2012). On the other hand, these models have the advantage that they are easily parallelizable and executable on computing clusters or graphics card servers (Riemenschneider *et al.*, 2017). Subsequently, this parallelization is easily extendable to distributed modeling on distantly located datasets and recombination of resulting local models. In contrast to a centralized approach, only the model, not the data, is transferred amongst partnering sites, leading to a substantial decrease in communication overhead. However, a significant challenge of many federated approaches is communication efficiency, since many FL algorithms, such as deep neural networks, require frequent communication for the exchange of model parameters throughout the training phase (McMahan *et al.*, 2016). In contrast, FRF is trained separately for each data silo and solely requires two communication steps per model update. Let N be the number of distinctly stored datasets, also called silos, D_i with $i \in \{1, \ldots, N\}$. Traditional ML would merge all datasets $D = D_1 \cup \ldots \cup D_N$ and build a classical joint data-centralized RF model $M_{\text{Centralized}}$ on D to benefit from insights of the entire dataset. In a federated scenario, the datasets D_i cannot be shared amongst entities, and neither D nor $M_{\text{Centralized}}$ can be generated. Therefore, the goal is to build a combined RF model M_{Combined} integrating knowledge from all datasets D_i without sharing the actual data (Yang *et al.*, 2019a) (see Fig. 1). At first, each entity locally performs a separate ML on its private data to fit a local RF model M_{Loc_i}. Subsequently, these local models $M_{Loc_1}, \ldots, M_{Loc_N}$ are aggregated at a central node and integrated to a combined model M_{Combined}. Thereby, solely the abstract models are exchanged, and the private data remains locally. The overarching goal of this study is to evaluate in detail the competitiveness of combined RF models compared to the locally trained and the classically trained data-centralized RF model on different biomedical datasets and with respect to various data realities.

2.3 Evaluation of challenges in healthcare data

To benchmark the performance of FRF in comparison to a data-centralized approach, we will utilize different datasets and aspects of data heterogeneity (ILPD, HCC, BCD, LTD). Therefore, we emulate three common challenging scenarios among clinical research: (i) a different number of participants or sites contributing to the global model, (ii) different number of samples within the shared datasets and (iii) different balances in the phenotype to be predicted, and evaluated the corresponding model performances. Finally, we will evaluate the performance of the federated RF model in a cross-silo example (BCTCGA).

1. **Different number of participants or sites contributing to the global model:** To analyze the effects of the distribution of resources amongst an increasing number of sites, we evaluate all models on the same sized data and split these into an increasing number of participants. To cover a broad spectrum of scenarios, we split the data into 2–100 separate silos. Finally, we compare the performance of the local models, a federated combined model and a data-centralized model trained on the entire training dataset.
2. **Different number of samples per site:** To investigate the effects of data-size imbalance between contributing participants, we split the data into two silos of varying complementary sizes, e.g. 5% and 95% or 25% and 75%. Subsequently, we compare the performance of the local (small and big) model with a non-weighted and weighting combined model.
3. **Different balances in the predicted phenotype:** To evaluate how the federated ML models perform compared to the local and the data-centralized approach when the class balance of the data differs, we select silos from each dataset that corresponds to a specific class imbalance.

2.4 Evaluation procedure

Figure 2 gives an overview of the evaluation procedure described in the following.

1. **Split dataset into training and test data**

   For the Monte Carlo cross-validation, we separated each dataset into training (90%) and test (10%). For a robust evaluation of the performance, this split procedure and all following steps are repeated at least 100 times. Subsequently, the data-centralized RF model is trained on the entire training set, while steps two and three generate the local and combined models, respectively. Finally, the performances of the classical models, the local models and the combined models are evaluated on the corresponding test datasets within each cycle. The following section describes the preparation for different heterogeneity assessed scenarios.

2. **Generate distributed datasets**

   To simulate the different scenarios with distributed heterogeneous datasets, we performed the following three procedures.

   a. Different number of sites:

      The original training dataset D is randomly split in same-sized silos $D_i \in \{D_1, \ldots, D_N\}$ with $N \in \{2, 5, \ldots, 100\}$. Each silo D_i represents one participant. For each silo, one local model M_i is trained. Then, the algorithm generates 100 decision trees for each local RF model to create the combined model.

   b. Different number of samples per site:

      The training dataset D is split in two silos D_1 and D_2 with $|D_1| < |D_2|$. The splitting thresholds were set in 5% steps so that the different splits: D_1 contains 5,10,...,40, 45% of the training data (small dataset) while D_2 respectively contains 95, 80,..., 60, 55% (big dataset). In addition, we investigate if weighting the models based on their sample size influences the combined models. For the non-weighted scenario, each local model contributes the same number of decision trees to the combined RF. In the weighted model, each

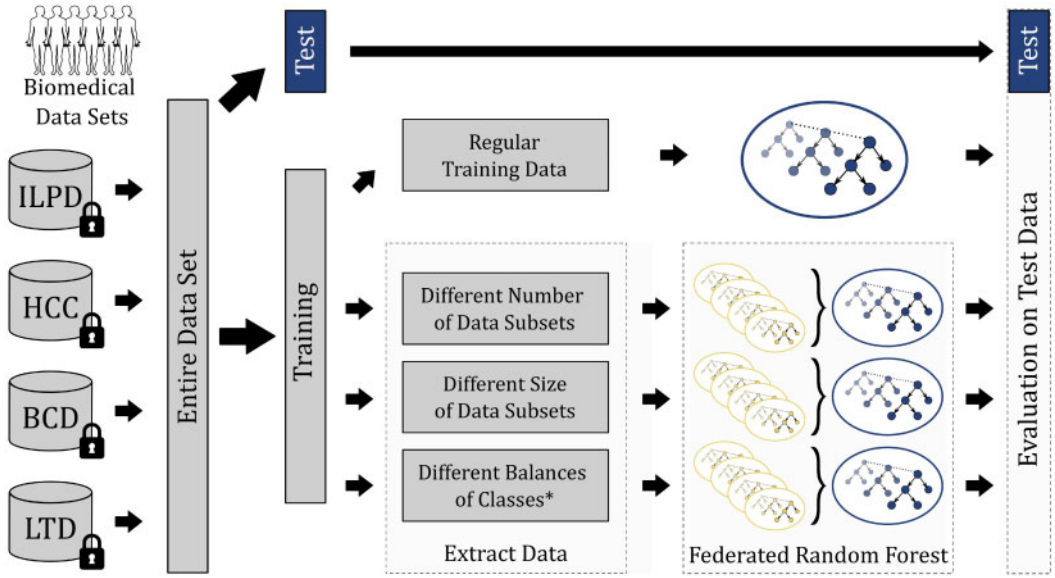

Fig. 2. Evaluation procedure comparing data-centralized ML models with the federated approaches. Therefore, for each clinical dataset, three different heterogeneity scenarios are evaluated, (i) a different number of silos, (ii) different sizes of silos and (iii) different balances of classes, i.e. prevalences. Subsequently, both the local, combined and data-centralized models are trained on these datasets. Finally, their performance is evaluated and compared. *Note that the data-centralized model is trained on the imbalanced data for the class imbalance evaluation to achieve a fair comparison

decision tree gets a weight assigned for the consensus decision relative to the number of the respective training samples

c. Different balances in the predicted phenotype:

We sample the datasets such that the percentage of the positive instances equals 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%. At first, a test dataset is sampled with the respective percentage. Then, depending on the balance, the remaining dataset was down-sampled such that the ratio and equal training sizes are guaranteed. Subsequently, the training set is split into two silos maintaining the ratio of positive and negative samples. Finally, we sampled the same number of decision trees for the combined model.
For all scenarios, the number of sampled decision trees was determined in such a way that $|M_{\text{Centralized}}| = |M_{\text{Combined}}|$.

3. **Combination of models**

As ensemble classifiers, RFs are easy to combine and therefore well suited for the federated setting. Let D_i with $i \in \{1, \ldots, N\}$ be distinctly stored datasets. The data-centralized model $M_{\text{Centralized}}$ is a RF that was trained by the entire dataset $D = D_1 \cup \ldots \cup D_N$. In the federated approach, for each silo D_i, a local model M_{Loc_i} is built. Each model equals a set of k decision trees $M_{Loc_i} = \{m_{Loc_i,1}, \ldots, m_{loc_i,k}\}$. To create the federated combined model, a subset of decision trees is randomly sampled from each local model. The number of sampled decision trees depends on the specific scenario. To create the federated combined model, all subsets of decision trees are merged into a single combined model $M_{\text{Combined}} = \{m_{Loc_1,1}, \ldots, m_{loc_1,k}, \ldots, m_{Loc_N,1}, \ldots, m_{loc_N,k}\}$. However, each analysis ensures that the number of decision tree models in the data-centralized and combined model is equal $|M_{\text{Centralized}}| = |M_{\text{Combined}}| = 100$. Finally, all chosen decision trees are merged into a combined model M_{Combined}.

4. **Evaluation of cross-silo example**

Equally to the previous analysis, we used Monte Carlo cross-validation to evaluate the cross-silo FRF. Thus, the BCTCGA dataset is separated into training (90%) and test (10%). For a robust evaluation of the performance, this split procedure and all following steps are repeated more than 100 times. Subsequently, the data-centralized RF model is trained on the entire training set, while a local model is trained on the data corresponding to a specific institute in the dataset. Finally, the performances of the data-centralized models, the local models and the combined models are evaluated on the corresponding test datasets within each cycle.

5. **Performance evaluation on test sets**

We evaluate the performance of the $|M_{\text{Centralized}}|$ model, all $M_{Loc_1}, \ldots, M_{Loc_N}$ models and the $|M_{\text{Combined}}|$ model for all repetitions of the generated and cross-silo scenarios. In addition, the performance was evaluated by the receiver operating curve (ROC). Therefore, the sensitivity and specificity are calculated and plotted for each positive-class-probability-cut-offs of the model. Subsequently, the area under the ROC (ROC-AUC) is integrated and serves as a quality measure of the model. However, the ROC-AUC does not adequately evaluate datasets that show strong imbalance in the phenotype. Thus we additionally analyze the performance using the precision-recall curve (PR) and the area under the PR curve (PR AUC) accordingly in these circumstances

The Python source code that allows for reproducing the results, the datasets and plots can be found in the following public GitHub repository of the project: https://github.com/jm9e/FL_Pipeline. Moreover, a detailed description of all required libraries can be found in the requirements.txt file inside the public Git repository.

3 Results

3.1 Different number of sites

Our analysis of the effects of increasing data distribution amongst a growing number of data silos (2–100) showed a similar pattern for all datasets. Exemplarily, Figure 3 depicts the performance evaluation of the ILPD and HCC datasets. As expected, for an increasing number of sites and subsequently decreasing number of samples, the performance (AUC) of the local models decreases drastically while the variance amongst runs increases. The performances of the combined model remain fairly stable and hardly vary from the data-centralized baseline model. However, certain limits apply when the

local models do not have a sufficient number of samples (no. of sample < 10), seen in particular for smaller datasets like LTD but also ILPD and BCD for a large number of sites. Solely, the HCC dataset does not show any differences, most likely due to its large size and the small number of variables. However, amongst all datasets, the test performance (AUC on the separate test data) of the combined model is substantially better than the performance of the local models. The performance analysis of BCD and LTD and detailed information about variance and significance can be found in Supplementary Material.

3.2 Different number of samples per sites (unbalancedness)

We evaluate the effects of an unevenly balanced number of samples amongst the different sites. Thus, the training data was separated into two data silos (D_1 and D_2) containing a specific percentage of the training data. For instance, the 'Small Dataset Model' is trained on a small dataset comprising $D_1 = 5\%$ while the 'Big Dataset Model' is trained on a large dataset comprising the remaining $D_2 = 95\%$ of the data; accordingly all other splits are evaluated: 'Small Dataset Models' $D_1 = \{5, 10, \ldots, 45, \text{Balanced}\}$ and 'Big Dataset Models' $D_2 = \{95, 90, \ldots, 65, \text{Balanced}\}$. Subsequently, the algorithm sampled and combined the local models in a weighted or unweighted fashion. Finally, the performances of the local (D_1 versus D_2) and combined models are compared to the data-centralized model.

Figure 4 exemplarily shows the performance of the analysis on the ILPD and LTD datasets. In general, the local models based on smaller datasets with fewer samples perform worse than those trained on larger datasets, resulting in an increasing and decreasing curve. Thus, the performance differs strongest where silo sizes are most different (see left part of the Fig. 4). Moreover, the combined models tend to successfully compensate for the lower performance of the local models based on the smaller datasets. The comparison of the weighted combined and unweighted combined models shows different results based on the overall sizes of the example datasets. For instance, for smaller datasets, such as LTD and BCD, the weighted combined models show a clear advantage on two local models that are based on large differences in sample sizes. However, for larger datasets, such as ILPD and HCC, the size-dependent weighting of the different local models does not have a strong effect on combined model performance. This effect is most likely due to the fact that even the smaller local models are based on sufficiently large data.

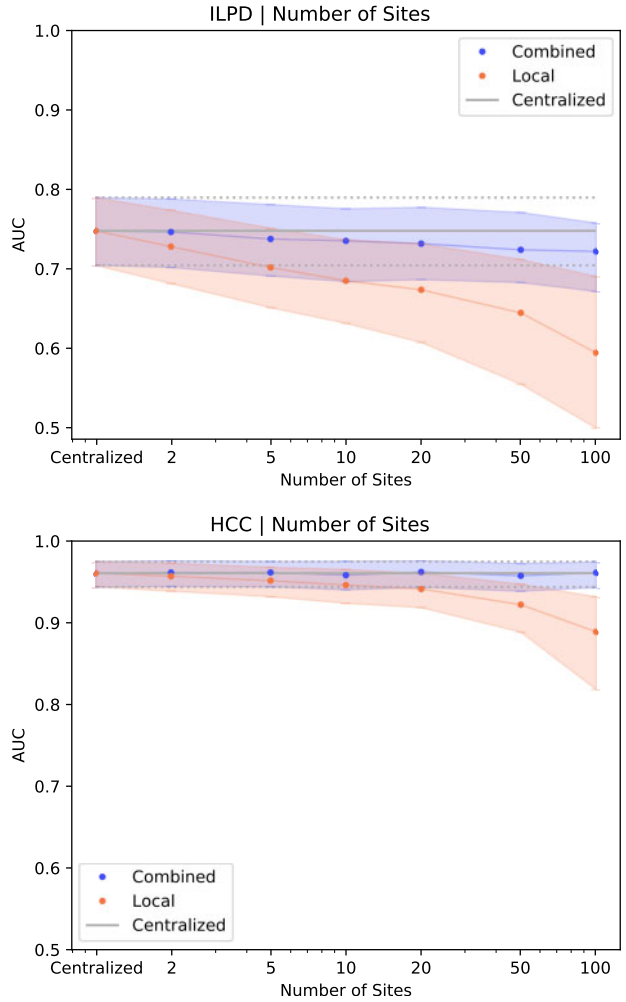

Fig. 3. Influence of increasing distribution of samples amongst a growing number of data silos (2–100) on the performance of a combined model. A local model is trained on each silo of the training data, respectively. Subsequently, the local models are merged into a combined model. This Figures depicts the performance on the ILPD and HCC datasets for both the local and combined models in red and blue, respectively. Both show different overall performances but demonstrate the same effects. Moreover, there is no significant difference between data-centralized and combined models. For comparison, the performance of the data-centralized model and the corresponding quantiles are indicated by horizontal black lines

3.3 Different imbalances in the predicted phenotype

To evaluate the influence of a bias in phenotype, we simulated differently imbalanced participant data. For instance, a dataset

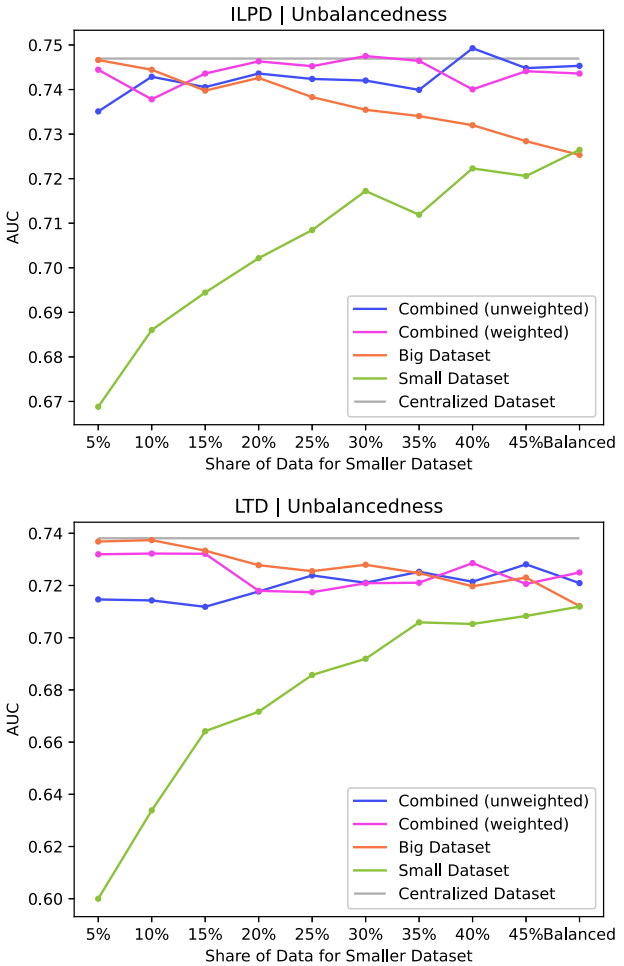

Fig. 4. Influence of unevenly sized data per site. On the one hand, the plots show the performance of the local models build on the small (green) and large (orange) silos of the data. On the other hand, the performance of the weighted and unweighted models are shown

containing 10% positive and 90% negative cases. Respectively, the percentage of positive samples corresponded to 10, 20,. . ., 80, 90%. To analyze the effect of the phenotype imbalance on the model's goodness of fit, we calculated both AUC and area under the PR AUC. For comparison, we also included the phenotype-balanced data (50% versus 50%). Figure 5 exemplarily depicts (i) the AUC and (ii) PR AUC performances for the ILPD dataset. In general, both the local and combined AUC performance variance increases with increasing imbalances for both negative and positive phenotypes, particularly for local models trained on a small number of positive samples. Consequently, the local and combined models are more stable (minor variance) and show better performance when the dataset almost balanced. Since the PR AUC focuses on the positive cases, the area under the PR AUC performance for all models improves with an increasing number of positive instances in the imbalanced dataset. Overall, the performance of the combined models consistently increases compared to the corresponding local models.

See Supplementary Section S3 for dataset HCC, BCD and LTC.

3.4 Cross-silo example—breast cancer TCGA data

To access the applicability of our analysis in a cross-silo scenario, we applied our FRF approach on a publicly available breast cancer TCGA dataset that originated from 19 different hospitals and institutions. The results depicted in Figure 6 show that, as previously observed, the combined model succeeds the performance of the models trained on the local institution-specific datasets. However, the results do not show significant improvement based on weighting the local models based on their sample size. Moreover, while the difference between the combined and data-centralized models is rather small, the combined model shows better performance than the local models.

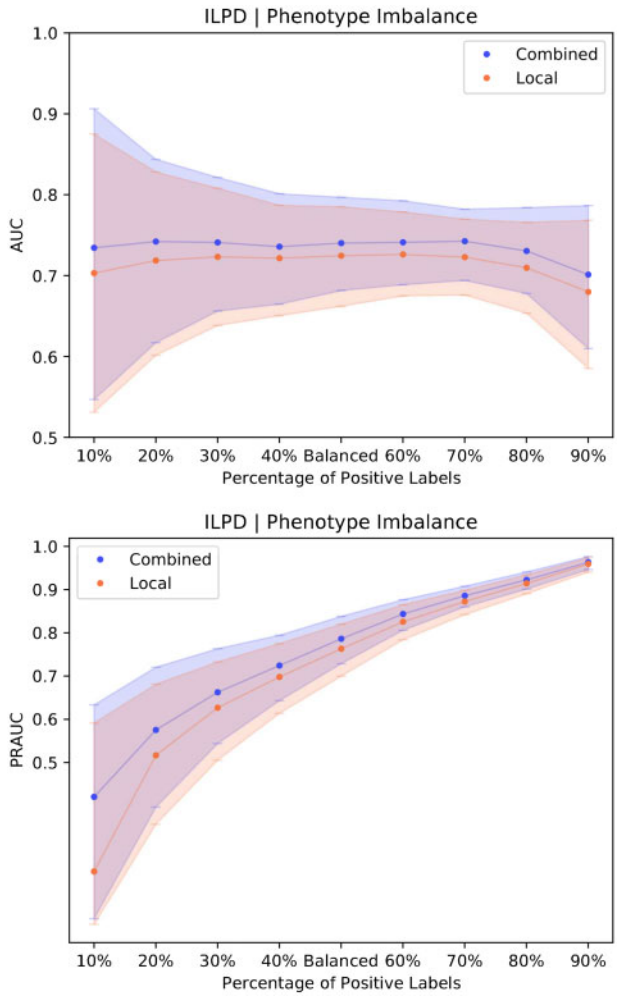

Fig. 5. Comparison of the performance of the federated local and combined models with the data-centralized model on differently imbalanced datasets. The first box plot visualizes the AUC of models trained 10–90% positive samples respectively. The second box plot depicts the corresponding area under the PR curve

4 Discussion

In clinical practice, studies are often limited to small local datasets due to a small number of patients or limited resources per hospital. While clinical studies using ML are usually successfully done on sufficiently large public data or larger cohorts, a lack of access to sufficient data due to privacy regulations hinders the transition to clinical applications. The distributed architecture of FL can address such challenges and comprises a paradigm shift from centralized data approaches. To clarify whether FL methods such as FRF can compete with classical approaches, we performed a thorough benchmark of the efficacy of FRF on five standard biomedical datasets, including clinical information, laboratory results, image-based parameters and gene expression data; and compared the results of the combined FRF model with the local and data-centralized ML approach. Data generated by biomedical research is different in many aspects compared to data from other domains. The heterogeneity within and between datasets, in particular concerning ethical, legal or social confounders, as well as imbalances with phenotype prevalence or cohort sizes, pose challenges for ML and AI in general (Cios and Moore, 2002). In this study, we evaluated three common challenging scenarios among clinical research, (i) a different number of participants or sites (here called silos), (ii) a different number of samples per site (unbalancedness) and (iii) different imbalances of phenotypes, and evaluated the corresponding model performances (AUC and precision-recall AUC). In addition, we validated our approach on a real-world cross-silo example based on a public distributed breast cancer dataset comprising expression profiles of patients from 19 different healthcare institutions.

Our results on the **number of sites** scenario (I) consistently show on all datasets that FRF enables to build a more powerful combined model based on the combination of local models trained on preexisting distributed data silos. The combined FRF models outperform the local models in all distribution scenarios on all example datasets. Strictly speaking, the performance advantage increases the smaller the size of local datasets. Thus, we conclude that for most clinical applications, the use of FRF architectures can aid in overcoming the obstacles of privacy and data governance challenges amongst participating institutions and improve overall phenotype prediction such as disease classification or treatment recommendations.

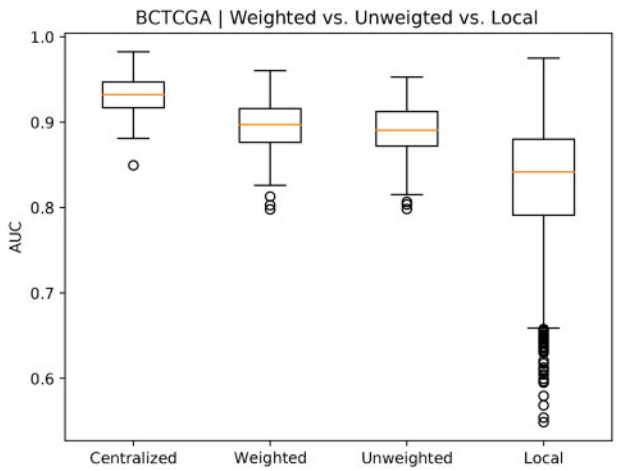

Fig. 6. Comparison of the performance of the cross-silo federated local and combined models with the data-centralized model on a cross-silo distributed breast cancer dataset

Our analysis of **unbalancedness** scenario (II) reveals that combined models tend to successfully compensate, in particular, the lower performance of the local models based on the smaller datasets. Moreover, in scenarios of small sample sizes and significant size differences between the data silos, weighting functions (here proportional number of trees per dataset size) on the combined model represents a clear advantage in terms of prediction quality. Solely examples of extreme unbalances showed a small but significantly decreased accuracy on the combined model compared to the model trained on the larger dataset. However, the performance was comparable when using size-dependent weighting for the combined model. In conclusion, FRF are further recommended to improve prediction performance for circumstances of small datasets and extreme unbalancedness. In addition, the results show that weighing functions can aid in compensating for the extreme unbalancedness.

The investigation of FRF for the varying **imbalances of phenotypes** scenario (III) reveals that combined models are consistently more robust toward particular biases within the parameter distributions. Thus in clinical applications that frequently show challenges of unbalanced parameters, the employment of FRF can additionally result in more robust predictions.

Moreover, in most scenarios (I, II and III), the combined models have shown to be comparable to the classic data-centralized model. However, exceptions apply mainly to models of extremely small sample sizes (mostly scenarios I), leading to small but significant differences (no. of sample < 10). Thus, while we can generally expect a performance improvement of the combined models in comparison to local models, it might in extreme cases not compare to a global data-centralized model, which, however, is often not feasible under privacy and legal perspectives.

Finally, when reviewing the results of the cross-silo dataset, both the unweighted and the weighted combined models (both AUC $<$ 90%) show slightly worse performance than the data-centralized model (AUC $>$ 92.5%). In this cross-silo scenario, the weighting function shows a slight improvement. However, for all simulated-distributed and cross-silo datasets, the combined models show a significant improvement compared to the performance of the corresponding local models. Thus, we can conclude that both medical research and studies and clinical practice would greatly benefit from the application of FRF.

5 Conclusion

With recent developments in AI and ML, tremendous opportunities for medical research are at our fingertips. However, up to now, limited sample sizes, biased datasets and limited data access, for instance, due to privacy and legal regulations, have hindered the field from exploiting the full potential of computational methodology. This limitation is particularly the case for ML-based on patient information, covered by data protection rules such as the GDPR. Therefore, we require a paradigm shift from centralized data lakes toward tailored ML architectures such as distributed or FL approaches integrated with SMPC, and encryption techniques to adhere to these privacy requirements. Such systems ensure that the actual data never leaves the owner and thus can be integrated with existing infrastructure and data silos. Thus, they can overcome mentioned boundaries, ease collaborations across institutes without tedious paperwork and lengthy processes since no data exchange is needed, and therefore initiate the transition from research to clinical practice.

While some FL methods exist and are already used in business health applications and mobile apps (Konečný *et al.*, 2016b; McMahan *et al.*, 2016), the deployment of the current state-of-the-art methodology to clinical settings is still in its infancy.

In the current study, we focus on the evaluation of horizontal FRF. This federated ML method seeks to build a generalized ensemble model without access to a shared data basis (Gan *et al.*, 2017). Therefore, we assume that the comparison results between federated and data-centralized models will be consistent with datasets that underwent secure data conversions such as homomorphic encryption.

The results consistently confirm for all scenarios, simulated (number of data silos, unbalancedness, imbalanced phenotypes) and cross-silo datasets, that the combined models show a significant improvement compared to the performance of the corresponding local models. When local models are built on datasets of decreasing sample size, the performance of the combined models cannot compete with a data-centralized model. The quality improvement compared to the local model is particularly strong. Thus, we conclude that in circumstances where privacy and legal regulations prohibit a data-centralized approach, the employment of an FRF will, with high probability, lead to more robust and generalizable prediction models.

The application of FL methods such as FRF can potentially benefit various stakeholders. Data remains under the governance of healthcare and research institutions and is stored complying with data protection rules such as the GDPR. Such laws comprise the right of a patient to revoke a use-consent. Saving the data locally guarantees the possibility of deleting data (including all trained models), thereby lowering patients' reluctance to become data donors. ML and clinical researchers can benefit from potentially vast collections of relevant biomedical data. Data collected by institutions from different geographic locations, of demographics and other varying factors, allow less biased (compensating, e.g. socio-economic and ethnic confounders) and more generalizable models. Subsequently, such models trained on a national or global scale have a strong potential to perform well on clinical decisions regardless of the treatment location, which will have relevance, especially for patients in rural areas. Moreover, a broad application of FL methods will be relevant, particularly for rare or geographically uncommon, diseases that are likely to be diagnosed faster and more accurate (Rieke *et al.*, 2020).

In addition to the general advantages of FL, due to their tree-based structure, FRF models are well equipped to evaluate variable importance and therefore are particularly suited to enhance the interpretability of subsequent combined models. Studies have shown that explainability and causality of such models, alongside model accuracy and robustness, is the most crucial factor for acknowledgment and acceptance of ML technologies in clinical practice (Holzinger, 2021; Janzing and Schölkopf, 2017; Schwarz and Heider, 2019). Thus, in the future, we will focus on explainability and the optimization of the FRF models to account for the heterogeneity and noisiness of biomedical data. Furthermore, these methods routinely have to account for inhomogeneous data sites (Kargupta *et al.*, 2000). These challenges are frequently found in medical data records.

In summary, we believe that federated ML-based architectures like FRF have the potential to increase the accuracy and robustness of healthcare AI, revolutionize both clinical research and practice, and pave the way for the precision medicine of the 21st century by building more accurate predictive models enabling more focused personalized treatment (Hamburg and Collins, 2010; Rieke *et al.*, 2020).

Data availability

All data is publicly available as described in material and methods.

Funding

This work was supported by the European Union's Horizon2020 research and innovation programme under Grant Agreement No 826078. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information in contains.

Conflict of Interest: none declared.

References

Batra,R. *et al.* (2017) On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.*, **3**, 6.

Best,J. *et al.* (2016) Der GALAD-Score, ein AFP-, AFP-L3- und DCP-basierter Diagnosealgorithmus verbessert die Detektionsrate des hepatozellulären

Karzinoms im BCLC-Frühstadium signifikant. *Z. Gastroenterol.*, **54**, 1296–1305.
Bösl,A. *et al.* (2017) MammaPrint versus EndoPredict: poor correlation in disease recurrence risk classification of hormone receptor positive breast cancer. *PLoS One*, **12**, e0183458.
Boulesteix,A.L. *et al.* (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisc. Rev. Data Min. Knowl. Discov.*, **2**, 493–507.
Brisimi,T.S. *et al.* (2018) Federated learning of predictive models from federated Electronic Health Records. *Int. J. Med. Inf.*, **112**, 59–67.
Calvert,J. *et al.* (2019) Machine-learning-based laboratory developed test for the diagnosis of sepsis in high-risk patients. *Diagnostics*, **9**, 20.
Chen,Y. *et al.* (2020) FedHealth: a federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.*, **35**, 83–93.
Cheng,K. *et al.* (2021) SecureBoost: a lossless federated learning framework. *IEEE Intell. Syst.*, **36**, 87–98.
Cios,K.J. and Moore,G.W. (2002) Uniqueness of medical data mining. *Artif. Intell. Med.*, **26**, 1–24.
Constable,S.D. *et al.* (2015) Privacy-preserving GWAS analysis on federated genomic datasets. *BMC Med. Inf. Dec. Mak.*, **15**, S2.
Fang,H. and Qian,Q. (2021) Privacy preserving machine learning with homomorphic encryption and federated learning. *Fut. Internet*, **13**, 94.
Fatima,M. and Pasha,M. (2017) Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.*, **09**, 1–16.
Gan,W. *et al.* (2017). Data mining in distributed environment: a survey. *WIREs Data Mining and Knowledge Discovery*, 7(6), e1216.
Hamburg,M.A. and Collins,F.S. (2010) The path to personalized medicine. *N. Engl. J. Med.*, **363**, 301–304.
Holzinger,A. (2021) Explainable AI and multi-modal causability in medicine. *i-com*, **19**, 171–179.
Janzing,D. and Schölkopf,B. (2017) *Elements of Causal Inference Foundations and Learning Algorithms*. The MIT Press, Cambridge.
Jeanquartier,F. *et al.* (2016) Machine learning for in silico modeling of tumor growth. *Machine Learning for Health Informatics*, 415–434.
Kargupta,H. *et al.* (2000) Collective data mining: a new perspective toward distributed data mining. *Adv. Distrib. Parallel Knowl. Discov.*, 133–184.
Konečný,J. *et al.* (2016a) Federated learning: strategies for improving communication efficiency. *arXiv*, *preprint arXiv:1610.05492*.
Konečný,J. *et al.* (2016b) Federated optimization: distributed machine learning for on-device intelligence. *arXiv*, *preprint arXiv:161002527*.
Lænkholm,A.-V. *et al.* (2018) JOURNAL OF CLINICAL ONCOLOGY PAM50 risk of recurrence score predicts 10-year distant recurrence in a comprehensive danish cohort of postmenopausal women allocated to 5 years of endocrine therapy for hormone receptor-positive early breast cancer. *J. Clin. Oncol.*, **36**, 735–740.
Lazarevic,A. and Obradovic,Z. (2001) The distributed boosting algorithm. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, pp. 311–316.
Lee,J. *et al.* (2018) Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR Med. Inf.*, **6**, e7744.
Li,J. *et al.* (2020) A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. *Artif. Intell. Med.*, **103**, 101814.
Liu,J. *et al.*; Cancer Genome Atlas Research Network. (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, **173**, 400–416.
Liu,Y. *et al.* (2020) Federated forest. *IEEE Trans. Big Data*, **1**.
Lorenzi,M. *et al.* (2017) Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study. In: *12th International Symposium on Medical Information Processing and Analysis*, Tandil, Argentina, Vol. 10160, pp. 1016016.
McMahan,H.B. *et al.* (2016) Communication-efficient learning of deep networks from decentralized data. *Artif. Intell. Stat.*, 54, 1273–1282.
Nasirigerdeh,R. *et al.* (2022) sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. Genome Biology, *23*, 32.
Park,Y. *et al.* (2021a) Integrative analysis of next-generation sequencing for next-generation cancer research toward artificial intelligence. *Cancers*, **13**, 3148.
Park,Y. *et al.* (2021b) Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. *NAR Genomics Bioinf.*, **3**, lqab104.
Ramana,B.V. *et al.* (2012) A critical comparative study of liver patients from USA and INDIA: an exploratory analysis. *Int. J. Comput. Sci. Issues*, **9**, 506.
Rieke,N. *et al.* (2020) The future of digital health with federated learning. *NPJ Digit. Med.*, **3**, 1–7.
Riemenschneider,M. *et al.* (2017) eccCL: parallelized GPU implementation of ensemble classifier chains. *BMC Bioinformatics*, **18**, 371.
Rousseaux,S. *et al.* (2013) Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.*, **5**, 186ra66.
Roy,A.G. *et al.* (2019) Braintorrent: a peer-to-peer environment for decentralized federated learning. *arXiv*, *preprint arXiv:1905.06731*.
Schork,N.J. (2015) Personalized medicine: time for one-person trials. *Nature*, **520**, 609–611.
Schwarz,J. and Heider,D. (2019) GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making. *Bioinformatics*, **35**, 2458–2465.
Slodkowska,E.A. and Ross,J.S. (2009) MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Exp. Rev. Mol. Diagn.*, **9**, 417–422.
Strecht,P. *et al.* (2014) Merging Decision Trees: a case study in predicting student performance. In: *International Conference on Advanced Data Mining and Applications*, Guilin, China, pp. 535–548.
Sundhar Ram,S. *et al.* (2012) A new class of distributed optimization algorithms: application to regression of distributed data. *Optim. Methods Softw.*, **27**, 71–88.
Sweeney,L. *et al.* (2013) Identifying participants in the personal genome project by name (A Re-identification Experiment). arXiv:1304.7605.
The Council of the European Union (2016) General data protection regulation. *Technical report*, The European Parliament.
Wang,S. *et al.* (2013) EXpectation Propagation LOgistic REgRession (EXPLORER): distributed privacy-preserving online model learning. *J. Biomed. Inf.*, **46**, 480–496.
Weinstein,J.N. *et al.* (2013) The cancer genome atlas pan-cancer analysis report. *Nature Genetics*, 45, 1113–1120.
Wiwie,C. *et al.* (2019) Time-resolved systems medicine reveals viral infection-modulating host targets. *Syst. Med.*, **2**, 1–9.
Wolberg,W.H. and Mangasarian,O.L. (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci. USA*, **87**, 9193–9196.
Yang,Q. *et al.* (2019a) Federated machine learning. *ACM Trans. Intell. Syst. Technol.*, **10**, 1–19.
Yang,Q. *et al.* (2019b) Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.*, **10**, 1–19.
Zapechnikov,S. (2020) Privacy-preserving machine learning as a tool for secure personalized information services. *Proc. Comput. Sci.*, **169**, 393–399.
Zhao,J. *et al.* (2019) Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci. Rep.*, **9**,