FeatureCloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Deliverable D4.3
"Test report on different classifier ensembles"

WP4
"Supervised Federated Machine Learning"

Disclaimer

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number: 826078			**Acronym: FeatureCloud**	
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Reza RAZAVI (CNECT/H/03)			
Project Coordinator	Jan BAUMBACH, TECHNISCHE UNIVERSITAET MUENCHEN (TUM)			
Deliverable	D4.3 “Test report on different classifier ensembles”			
Work Package	WP4 “Supervised Federated Machine Learning”			
Date of Delivery	**Contractual**	31/12/20	**Actual**	22/12/20
Nature	REPORT	**Dissemination Level**	PUBLIC	
Lead Beneficiary	03 MUG			
Responsible Author(s)	Dominik Heider, Anne-Christin Hauschild, Marta Lemanczyk (UMR)			
	Tobias Frisch (SDU)			
	Andreas Holzinger (MUG)			
	Jan Baumbach (TUM)			
Keywords	Federated Privacy-by-Design Machine Learning, Tree-based federated ensemble learning, Evaluation of challenges in biomedical data			

Table of Content

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Objectives of the deliverable based on the Description of Action (DoA)

The objective of WP4 is amongst others, to contribute to theoretical and experimental research, design, and development of federated interactive learning approaches following “privacy by design and architecture”. Therefore, this deliverable experimented with and evaluated graph-based learning approaches, as representative analysis we focussed the tree-based federated ensemble learning (TFEL) algorithm random forest (Objective 6).
Data science in the health domain is often challenged with problems such as the “small-n-large-p” (a small number of samples n and a large number of variables p) leading to a “curse of dimensionality”. Moreover, datasets in medical research sometimes show systematic or non-systematic biases, such as rare phenotypes, for instance.
The combination of information in distributed datasets using federated learning approaches in the FeatureCloud project addresses the small-n challenge. In this deliverable, we will evaluate the performance of TFEL methods focussing on frequent challenges (WP4 Task 6). In detail, we will investigate the three following scenarios:

1. Build different random forest models for differing number of subgroups
2. Build different random forest models for differing subset sizes
3. Build different random forest models for differently imbalanced phenotype classes

2 Executive Summary

Recent developments in artificial intelligence (AI) and machine learning (ML) hold tremendous opportunities for medical research. A manyfold of studies have proven ML to be advantageous for disease diagnosis, prognosis, and monitoring of diseases (Center for Devices and Radiological Health 2019; 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; Wiwie et al. 2019; Jeanquartier et al. 2016) and thereby enable tailored prognoses and monitoring of diseases. Moreover, computational models on clinical variables and electronic health records are used to assess individualized health risks, for instance to identify high-risk patients for sepsis in intensive care units (Calvert et al. 2019; Desautels et al. 2016) or the analysis of longitudinal data for the early detection of heart failure. However, to move technological advances towards clinical practice a large amount of data is required. Up to now, limited data access due to privacy concerns and security risks hindered the field to tap into the full potential of available computational methodology. This is particularly the case for machine learning based on patient data concerning data protection rules such as the General Data Protection Regulation (GDPR). Therefore, FeatureCloud focuses on novel approaches such as Federated Privacy-by-Design Machine Learning (FPDML), that exchanges models instead of the sensitive patient data.

In this deliverable, we evaluate random forest, a representative for TFEL methods, on biomedical data sets, focussing on challenges such as the heterogeneity within and between data sets.
Moreover, we emulated three common challenging scenarios among clinical research, (i) a different number of participating parties and their data sets; (ii) different sizes of data sets; and (iii) different class imbalances within data sets, and evaluated the corresponding model performances.

i) Different number of participants contributing to the global model: Depending on the available patients and resources, a single study can sometimes only obtain a small number of samples. By sharing data among multiple collaboration partners, a better model can be trained. We will evaluate the performance of the local submodels, a federated combined model and a classical model trained on the whole data set and analyse the effects of the same sized data originating from an increasing number of participants.
ii) Different number of samples within the shared data sets: Depending on the available patients and resources, subsets can have varying sizes among the contributed data sets and

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

subsequently differences in predictive power of trained submodels. Therefore, we investigate if the weighting of the contribution of the submodels depending on the size has an influence on the performance of the combined model.

iii) Different balances in the phenotype to be predicted: An additional challenge in medical data science are imbalanced data sets, i.e., the target phenotype is extremely unevenly distributed within the data set. We evaluate how the federated machine learning models behave compared to the classical-centralized approach when the class balance of the data differs.

Our comprehensive evaluation of federated ensemble learning models on different biomedical data sets and different challenges frequently occurring in clinical studies showed that the distributedly trained combined models performed equally accurately compared to the classical models trained on entire data sets.
Lastly, we will introduce future steps in interpretability of the developed ensemble models, since tree-based models are particularly suited to evaluate variable importance. It has been shown that besides model accuracy and robustness the most important factor for acknowledgement and acceptance of such technologies in clinical practice is Explainability and Causability of such models (Pearl 2009; Peters, Janzing, and Schölkopf 2017).

In summary, federated machine learning architectures can aid to overcome the boundaries of clinical research, enabling collaborations across institutes, and can help to overcome data shortage and biases. These technologies have the power to revolutionize both clinical research and practice and pave the way for precision medicine of the 21st century (Hamburg and Collins 2010).

3 Introduction (Challenge)

3.1 Background

The digital revolution in healthcare, fostered by novel high-throughput technologies and electronic health records (EHR), transitions the field towards a big data era (Constable et al. 2015). In particular, the combination of big data and artificial intelligence (AI) offers new opportunities to transform healthcare towards precision medicine. Given large 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 relevant breast cancer marker (Lænkholm et al. 2018; Slodkowska and Ross 2009). They are currently used as medical diagnostics tools for breast cancer subtyping, guiding the selection of 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).
In general, the main bottleneck in many studies is the small number of samples compared to a large number of features, a scenario which is called the small-n-large-p problem, resulting in a computational issue termed the curse of dimensionality.

The Cancer Genome Atlas (TCGA, (Weinstein et al. 2013)) is the by far most comprehensive repository for clinical cancer omics data worldwide. It contains whole-genome gene expression data for almost five thousand breast cancer patient samples, which are linked to clinical outcomes. These few thousand samples, however, stand against more than 20 thousand features that an AI may pick and combine to predict the outcome. This is particularly surprising since, in the European Union alone, there are about 350 thousand new breast cancer cases per year (International Agency for Research on Cancer[1]). The consequence is the risk for model overfitting and a significantly reduced robustness of this kind of medical diagnostics tool.

[1] http://eco.iarc.fr/EUCAN/CancerOne.aspx?Cancer=46\&Gender=2

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Moreover, systematic biases within clinical trials, in particular towards 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, but yet most studies are limited by small sample sizes and suffer from curse of dimensionality. Hence, solely fractions of these studies can be utilized for mining prognostic and predictive markers. 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 Electronic Health Records (EHR) across institutes, nation- and global-wide could address these previous limitations of sample size and systematic biases and subsequently, move the field towards more accurate precision medicine.
However, a global exchange harbors risks to data safety of sensitive patient information and EHRs stored in critical healthcare infrastructure. Especially data exchange amongst institutions over the internet is posing a roadblock hampering big data-based medical innovations.

In 2016 the EU passed the GDPR which sets rules for storing and sharing data in and outside the EU (Voigt and Von dem Bussche 2017). 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 mostly 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, Abu, and Winn 2013). Moreover, keeping data centralized, for instance on a shared server or a cloud, increases the risk of cyberattacks.
To tackle these problems and to guarantee a privacy-preserving data exchange in a biomedical environment, a novel approach is needed. Article 25 of the GDPR states that if a person's privacy is ensured all the time, a ’protection by design’ technology can be used.

In FeatureCloud, we focus on a novel strategy to overcome the legal barrier of exchanging raw patient data and thus enable true large-scale medical data mining in a cyber-risk-minimizing manner. FPDML enables the combination of globally distributed data and the training of global computational models without the need to send any confidential data over any communication network.

Here we employ one of the most common representatives of TFEL methods, random forest, to public data, treated as confidential and distributed, to evaluate whether federated machine learning has the same predictive and prognostic power as classical, centralized approaches. Such a transformative security-by-design concept will further minimize the cyber-crime potential and enable secure cross-border collaborative data mining endeavors and ultimately, allow the medical field to enter the “big data” era also in practice.

3.2 Related Work

Privacy-preserving data mining and distributed learning have numerous applications in various areas of research. Each poses different constraints concerning privacy, results, data distribution, and collaborative or cooperative computing.

3.2.1 Tree-based Federated Ensemble Learning

The basic principles of tree-based federated ensemble learning (TFEL) and parallelized online learning were coined by research fields such as distributed data mining, meta-learning, or collective learning (Park and Kargupta 2002). In recent years, federated architectures rapidly integrated into research and commercial areas such as on mobile applications to minimize data traffic (Konečný, Brendan McMahan, Yu, et al. 2016; Konečný, Brendan McMahan, Ramage, et al. 2016; McMahan et al. 2017).

In parallelized online learning, multiple computers are used to improve the performance of a machine learning model but share access to the same data or the exchange of intermediate results (Li et al.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

2014) which is heavily used in the area of Deep Learning. In contrast, federated ensemble learning seeks to build a generalized model without access to a shared data basis (Gan, Lin, and Chao 2017) and therefore requires a fundamentally different architecture also known as privacy-by-design.
Let N be the number of distinctly stored data sets D_i with i={1,..., N }.
To benefit from insights of the entire data, traditional machine learning would merge all datasets $D=D_1 \cup ... \cup D_N$ and build a classical joint model $M_{Classical}$ on D.
In a federated scenario, the data sets D_i can not be shared amongst entities and neither D nor $M_{Classical}$ can be generated.
Therefore, the goal is to build a combined model $M_{Combined}$ integrating knowledge from all datasets D_i without sharing the actual data (Yang et al. 2019).
At first, each entity locally performs a separate machine learning on its private data to fit a sub-model $M_{Submodel(i)}$. Subsequently, these submodels are aggregated at a central node and combined to the combined model $M_{Combined}$. Thereby, solely the machine learning models are exchanged and the private data remains locally.

As privacy and data security gained increasing traction over the past decade, multiple federated algorithms have been developed. For instance, distributed regression, where an encrypted posterior distribution of coefficients is sent to a shared server that stores and updates a global model (Sundhar Ram, Nedić, and Veeravalli 2012; Wang et al. 2013).
Other approaches implement distributed ensemble learning methods, such as federated decision trees (Strecht, Mendes-Moreira, and Soares 2014; Liu et al. 2020; Tuladhar et al. 2020), or distributed boosting (Lazarevic and Obradovic 2001), to mention just a few.
Another important advantage of TFEL is their built-in ability for intuitive variable importance evaluation. In particular, in clinical research the most important factor for acknowledgement and acceptance besides model accuracy and robustness in clinical practice is interpretability and explainability of such models (Pearl 2009; Peters, Janzing, and Schölkopf 2017).
Therefore, we will introduce future steps in interpretability of the developed ensemble models in Section 6.

3.2.2 Federated Ensemble Learning in Clinical Research

Until recently, these approaches had not been designed with medical data in mind and only a few studies employ federated ensemble learning.
For the longest time, the focus persisted on privacy-preserving approaches, such as on homomorphic encryption or differential privacy.
For example, the centralized computation of a kernel matrix on encrypted data sets (Yu, Jiang, and Vaidya 2006), a privacy-preserving Genome-Wide Association Study (GWAS) based on frequency counting, and X^2 statistics (Constable et al. 2015).
However, more recently, few pilot studies applied federated learning to biomedical data. Ming et al. and others proposed a virtual data pooling approach to enable large-scale meta-analysis of distributed neuroimaging data (Ming et al. 2017; Choudhury et al. 2019). Therefore, a decentralized iterative gradient descent optimization process is integrated into classical machine learning methods, namely t-distributed nonlinear embedding (tSNE), shallow and deep neural networks, joint independent component analysis (ICA), and independent vector analysis (IVA). However, this approach requires a central component to coordinate the iterations and access the distributed entities.
In contrast, Lorenzi et al. implemented a multi-centric dimensionality reduction approach using eigenvalue decomposition, which does not require iteration over centers (Lorenzi et al. 2017). They focussed on sequential and meta- partial least squares to model associations between genetic markers and anatomical surface features in Alzheimer's Disease.
In another study, Brisimi et al. analyzed electronic health records utilizing a federated l_1 -regularized sparse support-vector machine based on a modified version of the Primal-Dual Splitting algorithm (Brisimi et al. 2018; Davis 2015). Most recently, few studies, for instance Liu et al. in 2020 or Tuladhar

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

et al. in 2020, addressed the application of ensemble learning algorithms on biomedical data (Liu et al. 2020; Tuladhar et al. 2020)

3.2.3 Motivation

Medical data is rather different in many aspects from other branches in data mining: In particular, the heterogeneity within and between medical data sets with respect to ethical, legal, or social confounders, but also imbalances with phenotype prevalence or cohort sizes (Cios and Moore 2002). In general, it is expected that the accuracy $acc_{Combined}$ of the model $M_{Combined}$ is approximating the accuracy $acc_{Classical}$ of the classical model $M_{Classical}$.
The overarching goal of this Deliverable is to evaluate in detail the competitiveness of combined ensemble models in comparison to the classical trained model on different biomedical datasets, taking into account various challenges in biomedical data analysis. In particular, we will thoroughly investigate their robustness to (1) differing number of participants (i.e. differing number of subsets), (2) differing subset sizes, and (3) imbalanced phenotype classes.

4 Methodology

To benchmark the performance of federated machine learning approaches in comparison to a standard centralized approach, different aspects of data heterogeneity were taken into account. Federated ensemble learning approaches 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, Janitza, and Kruppa 2012).
On the other hand, these models have the advantage that they can easily be parallelized and executed on computing clusters or graphics card servers (Riemenschneider et al. 2017). This parallelization can be easily extended to distributed modeling on distantly located data sets and recombination of resulting submodels.

4.1 Data sets

For the performance analyses, we used four different clinical (ILPD, HCC) and biomedical (BCD, LTD) data sets see **Table 1** for the number of samples, features, and class balance.

Table 1. The following data sets are used to evaluate the performance of federated learning methods.

Name	Samples	Positives/Negatives	# Variables
ILPD	583	416/167	10 numeric
HCC	685	282/403	7 numeric
BCD	569	212/357	30 numeric
LTD	293	178/115	22,600 numeric

The following paragraphs give a short description of these public data sets:

- **ILPD:** The Indian Liver Patient Data (ILPD) set consists of liver patients (Ramana, Babu, and Venkateswarlu 2012). The positive instances were classified by experts as patients with liver disease. The features are clinical measurements as well as age and sex.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

- **HCC:** All patients suffer from chronic liver disease. Positive instances are patients who were diagnosed with Hepatocellular Carcinoma (HCC) (Best et al. 2016). The data set consists of clinical and biometric features.
- **BCD:** The breast cancer diagnosis (BCD) data set 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).
- **LTD:** The lung tumor diagnosis (LTD) data set 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.

4.2 Subsets

4.2.1 Different number of participants (subsets):

Depending on the available patients and resources, a single study can sometimes only obtain a small number of samples which can lead to weak predictive modeling and overfitting effects, i.e., the resulting model is not able to generalize to new data. By sharing data among multiple collaboration partners, a better model can be trained. We hypothesize that the federated approach performs competitively to the classical model trained on the whole data set, regardless of the number of participants the data is divided between.

4.2.2 Different sizes of subsets:

Another issue is the varying sizes among the contributed data sets and subsequently differences in the predictive power of trained submodels.
Therefore, we investigate whether the weighting of the contribution of the submodels depending on the size influences the performance of the combined model.

4.2.3 Different balances between classes:

An additional challenge in medical data science are imbalanced data sets, i.e., the target phenotype is extremely unevenly distributed within the data set, as it is the case, for instance, in rare diseases where the prevalence of the disease is low. We evaluate how the federated machine learning models behave compared to the classical-centralized approach when the class balance of the data differs.

4.3 Evaluation Workflow

The following section describes the different steps of the evaluation workflow as depicted in Figure 1.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

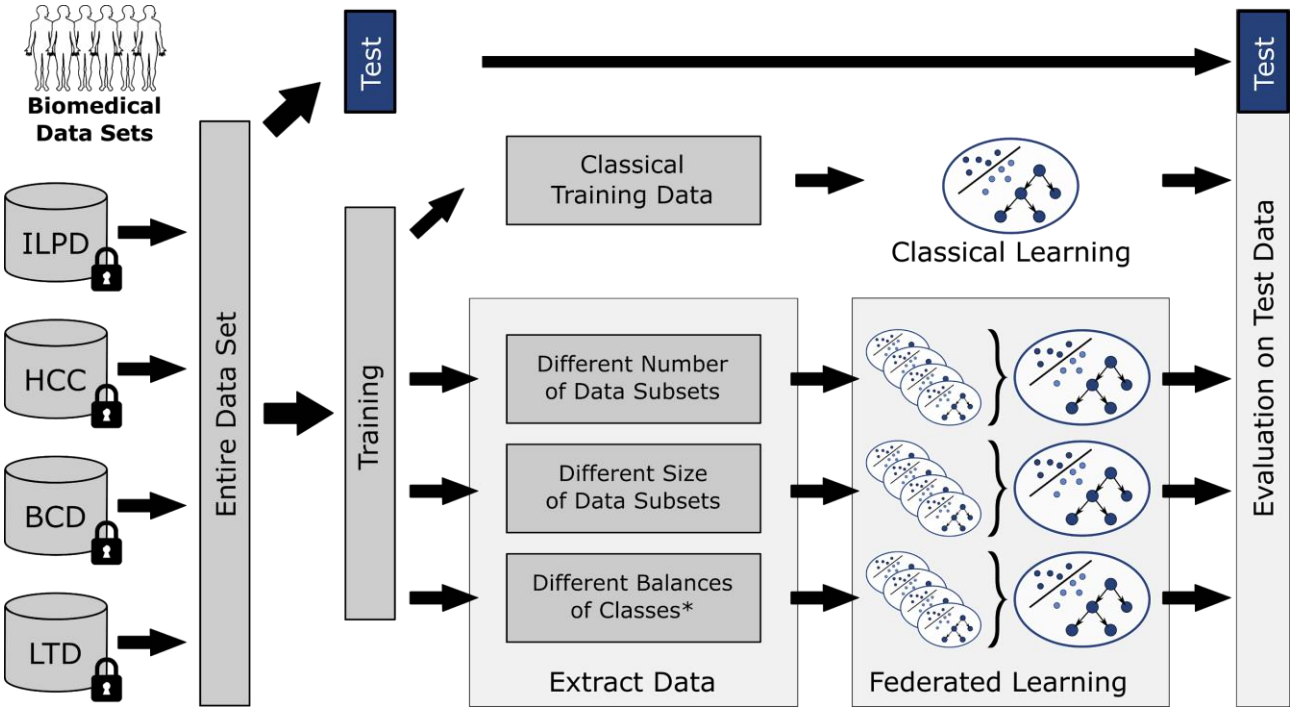

Figure 1. Evaluation Workflow comparing classical model machine learning models with the federated approaches. Therefore, for each clinical data set, three different scenarios are used, (i) a different number of subgroups, (ii) different sizes of subgroups, and (iii) different balances of classes, i.e., prevalences. Subsequently, both the classical model as well as the federated models are built on these data sets. Finally, their performance is evaluated and compared. *Note that for the class imbalance evaluation, the classical model is trained on the emulated data to achieve a fair comparison.

4.3.1 Data set preparation for test data

Performances of the classical models, the submodels, and the combined models are evaluated on the same test data sets within each cycle. For the evaluation of the different number of subsets, subset sizes, and imbalances, each cycle is analyzed by 10-fold cross-validation. The test data set for different balances is described in the following section.

4.3.2 Data set preparation for training

a) **Different number of subsets**
The original training data set D is randomly split in same-sized subsets $D_i = \{D_1,...,D_n\}$ with $N = \{2,...,5\}$ (In our analysis we expect a rather small number of hospitals to contribute, therefore we focus on up to 5 participants). For each subset, one submodel M_i is trained. To create the combined model, the same number of decision trees was sampled from each submodel. The sample number was determined in such a way that $|M_{Classical}| = |M_{Combined}|$.

b) **Different sizes of subsets**
To see the effects on combined models, we analyzed submodels with various size ratios. The training data set D is split into two subsets D_1 and D_2 with $|D_1|<|D_2|$. The splitting thresholds were set in 10% steps so that there are four different splits: D_1 contains 10, 20, 30, and 40% of the training data and D_2 respectively 90, 80, 70, and 60%. Additionally, we investigate if weighting the models based on their sample size influences the combined models. For the non-weighted scenario, each submodel contributes the same number of decision trees for the combined random forest. In the weighted model, for each submodel, the number of decision trees is sampled relative to the size of the training subset.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

c) **Different balances between classes**
We sample the data sets such that the percentage of the positive instances equals 10, 30, 50, 70, and 90%. At first, a test data set is sampled with the respective percentage. Depending on the balance, the remaining data was up- or down-sampled such that the ratio and equal training sizes are guaranteed. Subsequently, the training set is then split into two subsets maintaining the ratio of positive and negative samples. For the combined model, the same number of decision trees was sampled.

4.3.3 Combination of models

Since random forests are already ensemble classifiers, they are well suited for the federated machine learning approach. Let D_i with i={1,..., N } be distinctly stored data sets. The classical model $M_{Classical}$ would be a random forest that was trained by the merged data set $D=D_iU$ ... U D_N. In the federated approach, for each subset D_i, a submodel $M_{Submodel(i)}$ is built. Each model equals a set of k decision trees $M_{Submodel(i)}=\{m_{Submodel(i,1)},...,m_{Submodel(i,k)}\}$. To create the federated combined model, a subset of decision trees is randomly sampled from each submodel. The number of sampled decision trees depends on the specific scenario. However, each analysis ensures that the number of decision tree models in the classical and combined model is equal to $M_{Classical}= M_{Combined}$. Finally, all chosen decision trees are merged into a combined model $M_{Combined}$.

4.3.4 Performance evaluation on test sets

For a thorough evaluation of the robustness of the results, each experiment (a,b,c) was repeated 100 times. The performance was measured by the *Area under the Curve* (AUC). Moreover, for the analysis of different class balances, the *Precision-Recall AUC* (PR AUC) was employed additionally.

4.4 Implementation

We use the following packages from the *scikit-learn* library:

- sklearn.ensemble: Used to build the models. The Gini impurity is applied on the splits for the nodes and each classical model, the random forest consists of 500 decision trees.
- sklearn.model_selection: The stratified 10-fold cross-validation is executed with the StratifiedKFold method.
- sklearn.metrics: Used for the calculation of the performance measurements *AUC* and *PR AUC*.

5 Results

The results of our analysis are depicted in the following boxplot figures. Each boxplot represents the performance of the different models (classical, combined, sub) and their variation among the 100 evaluation runs.
The light blue box plots represent the performances of the classical models while the dark blue lines show the performances of the combined models. For comparison, we included the performances of submodels, which are represented by the yellow boxplots.

5.1 Different number of subsets

The analysis of the influence of the number of subsets on the performance of the combined model yields similar results in all four data sets. Exemplarily, **Figure 2** depicts the performance evaluation of the ILPD and HCC data sets.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

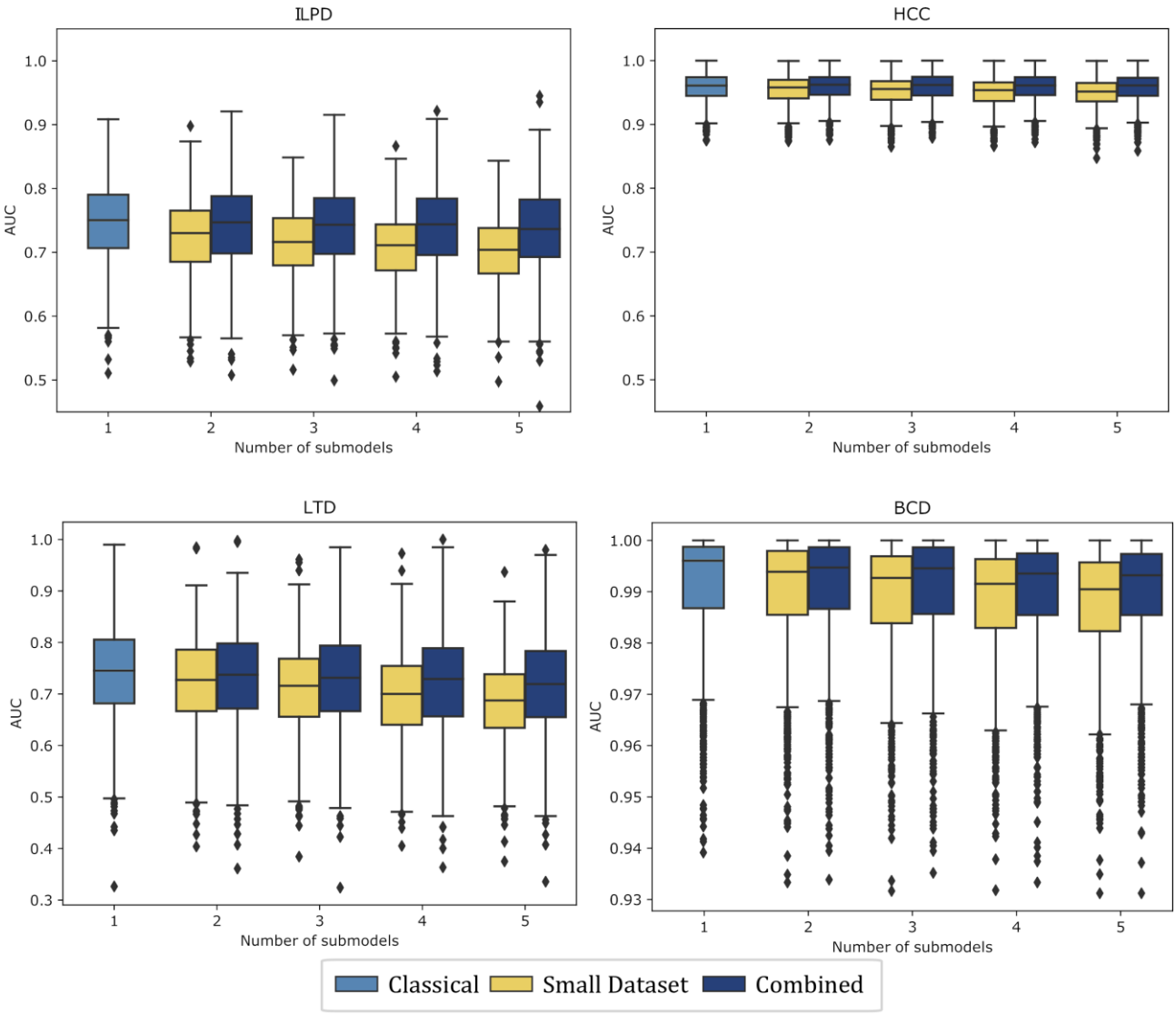

Figure 2. Influence of the number and size of local submodels on the performance of a combined model. The original data sets are divided into two to five subsets. Each subset was used to train one model. Subsequently, these submodels were merged into a combined model. Here we compare the results of the ILPD and HCC data sets. Both show different overall performances but demonstrate the same effects. There is no significant difference between classical and combined models.

It can be observed that the performances of the combined models do not differ strongly from the classical models (less than 1% in all cases). Moreover, there was no significant difference found between the AUCs of the classical and the model combining two subsets. However, for the ILPD, BCD, and LTD data sets, significant differences (even if small effect sizes) occur for combined models of three, four, or five subsets. Solely, the HCC data set does not show any differences, most likely due to its large size and the small number of variables. In contrast, with the increasing number of subsets, the number of samples in each set decreases, resulting in a slightly decreasing performance of the single models. Nevertheless, the differences between classical and combined models are much smaller in comparison to the submodels. In clinical practice, studies are often limited to small local data sets. Employing federated approaches will enable the combination of distributedly trained submodels to a more powerful combined model.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

5.2 Different sizes of subsets

To evaluate the effects of differently sized data on subset models, we separated the submodels into two groups. The original data set was divided into two subsets containing a specific percentage of the data to train each model. For instance, the small data set equals 10% and the large data set = 90% of the data; accordingly other splits were evaluated: 20% vs. 80%, 30% vs. 70% or 40% vs. 60%. Subsequently, these submodels were first either weighted or not weighted, then combined, and finally compared with the classical model.

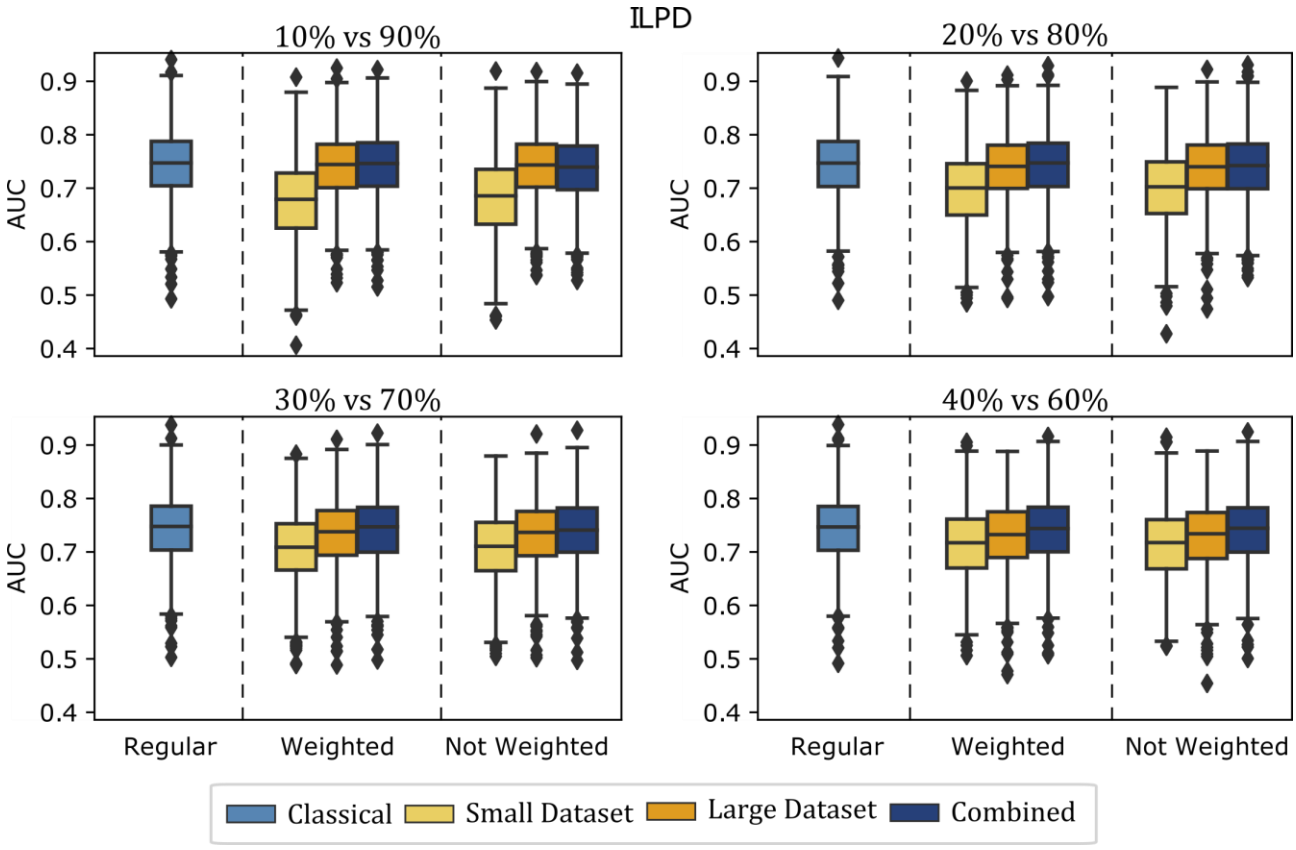

Figure 3. Influence of different sized imbalances of the subsets. Example visualization of the results for the models trained on the ILPD data set split by 10% / 90% and 40% / 60%. The boxplots show the performance of the models based on the smaller subsets (light yellow) and the larger subsets (dark yellow). Finally, we compare both the weighted and not weighted combined models.

Figure 3 shows an example of the results of the ILPD data sets. In general, the submodels based on smaller data sets perform worse compared to those trained on larger data sets. This effect is particularly present if the data sets differ strongest in size as shown in the upper left part of **Figure 3**. However, it can be observed in all data sets that the combined models tend to compensate for the worse performance of the submodels based on the smaller data sets. However, there is no significant difference between the weighted combined and classical models, see "Other supporting documents", Chapter 11.

5.3 Different balances of classes

We use both AUC and PR AUC to measure the performances of the models built on differently imbalanced data. For instance, we evaluated data sets of 10% positive and 90% negative cases accordingly imbalances were evaluated: 20% vs. 80%, 30% vs. 70% , 70% vs. 30% or 90% vs. 10%. Subsequently, We also included balanced data (50% vs. 50%) for comparison. In **Figure 4a**, the AUC performances of all models on imbalanced data sets vary more than for the models trained on

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

the balanced data. Since the precision-recall curve focuses on the positive cases, the area under the precision-recall curve (PR AUC) performance for all models is improving according to the increasing number of positive instances in the imbalanced data set. This can be observed in **Figure 4b**. In summary, the performances of the combined models do not significantly differ from the performances of the classical models, see "Other supporting documents", Chapter 11 for details. However, the combined models' performances slightly increase compared to their submodels.

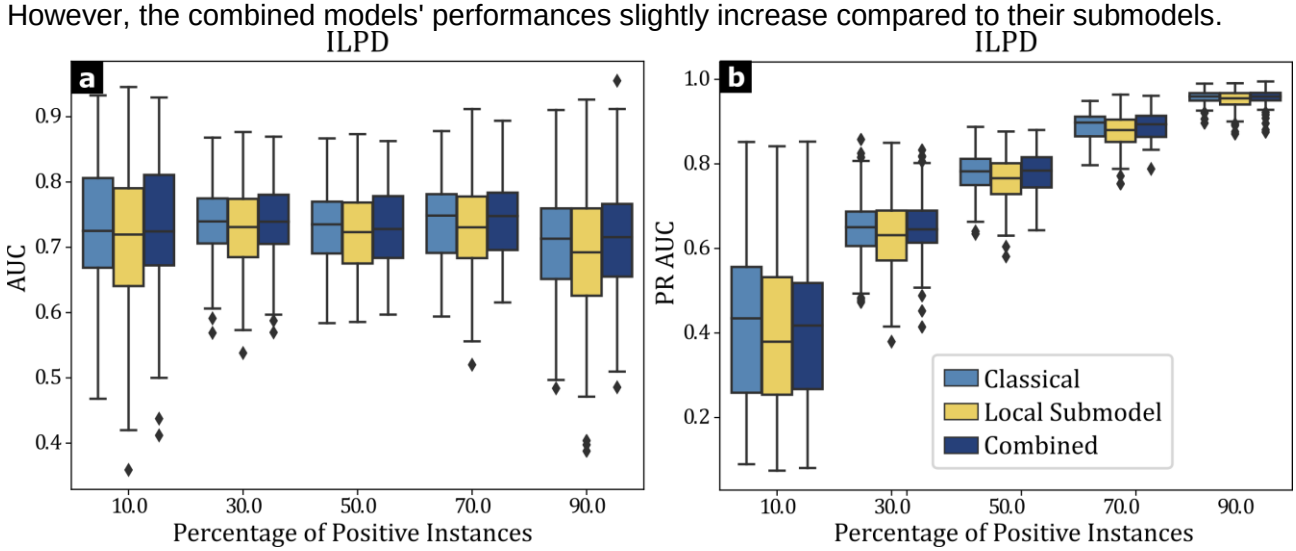

Figure 4. Comparison of the performance of the federated local and combined models with the classical model on differently unbalanced data sets. (a) The first box plot visualizes the AUC of models trained 10% - 90% positive samples respectively. (b) The second box plot depicts the corresponding area under the PR curve.

6 Open issues - Relevance for Explainability and Causability

In the subsequent sections, we demonstrated a comprehensive evaluation of federated ensemble learning models on different biomedical data sets and different challenges frequently occurring in clinical studies. However, it has been shown that besides model accuracy and robustness the most important factor for acknowledgement and acceptance of such technologies in clinical practice is Explainability and Causability of such models (Pearl 2009; Peters, Janzing, and Schölkopf 2017).

1. Why is Explainability relevant?

Explainability is at least as old as AI itself and rather a problem that has been caused by it. In the pioneering days of AI (Newell, Shaw, and Simon 1958), reasoning methods were logical and symbolic. These approaches were successful, but only in a very limited domain space and with extremely limited practical applicability. Early AI systems reasoned by performing some form of logical inference on human-readable symbols and were able to provide a trace of their inference steps. What do we mean by Explanation:

1) A peer-to-peer explanation as it is carried out among medical experts during medical reporting;
2) An educational explanation as it is carried out between teachers and students;
3) A scientific explanation in the strict sense of Karl Popper’s Theory of Science

2. Explainability in FeatureCloud

In **FeatureCloud** we deal with the first type of explanation. This is important because meanwhile not only performance but also transparency and interpretability are key issues, and recently the European Union emphasized robustness and explainability as the key issues for future AI (Hamon, Junklewitz, and Sanchez 2020). This is important because in the medical domain it is necessary to

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

understand the causality of learned representations (Pearl 2009; Peters, Janzing, and Schölkopf 2017).

Moreover, the explainability of AI could help to enhance the trust of medical professionals in future AI systems. Research towards building explainable-AI systems for application in medicine requires to maintain a high level of learning performance for a range of ML and human-computer interaction techniques. There is an inherent tension between ML performance (predictive accuracy) and explainability. Often the best-performing methods such as deep learning (DL) are the least transparent, and the ones providing a clear explanation (e.g., decision trees) are less accurate (Bologna and Hayashi, n.d.). Ensemble methods are ex-ante interpretable, therefore support re-traceability and interpretability. In particular, random forest offers a built-in ability for intuitive variable importance evaluation. The out-of-box-samples in the random forest algorithm allow to evaluate the importance of each variable for the prediction amongst all trained decision trees and therefore deliver robust measures such as the Gini-Index or the average increase of accuracy.

Currently, explanations of why predictions are made, or how model parameters capture underlying biological mechanisms are elusive. A further constraint is that humans are limited to visual assessment or review of explanations for a (large) number. So this still needs research.

3. Explainability in Medicine

We argue that in medicine explainable AI is urgently needed for many purposes including medical education, research, and clinical decision making (A. Holzinger 2018). If medical professionals are complemented by sophisticated AI systems and in some cases, future AI systems even play a huge part in the decision-making process, human experts must still have the means—on-demand —to understand and to retrace the machine decision process.
At the same time, it is interesting to know that while it is often assumed that humans are always able to explain their decisions, this is often not the case! Sometimes experts are not able to provide an explanation based on the various heterogeneous and vast sources of different information. Consequently, explainable-AI calls for confidence, safety, security, privacy, ethics, fairness, and trust (Kieseberg, Weippl, and Holzinger 2016), and brings usability (Andreas Holzinger 2005) and Human-AI Interaction into a new and important focus (Miller, Howe, and Sonenberg 2017). All these aspects together are crucial for applicability in medicine generally, and future personalized medicine, in particular (Hamburg and Collins 2010).

7 Conclusion

With recent developments in artificial intelligence and machine learning, tremendous opportunities for medical research are at our fingertips. However, up to now, limited data access due to security risks has hindered the field to gain from the full potential of computational methodology. This is particularly the case for machine learning based on patient data concerning data protection rules such as the GDPR.
Therefore, the **FeatureCloud** project pursues a federated privacy-by-design machine learning (FPDML) approach, which seeks to build a generalized model without access to a shared data basis (Gan, Lin, and Chao 2017) and therefore requires a fundamentally different architecture. Additionally, FPDML can account for issues arising through the complicated data ownership, privacy of the patients, and potentially ruining lawsuits. The privacy-by-design architecture ensures that the actual data never leaves the site and thus can be utilized with current infrastructure and also cannot be reconstructed by intercepting the exchanged models or by a fraudulent node in the distributed system.
While FPDML methods exist and are already frequently used in business applications and mobile apps (Konečný, Brendan McMahan, Yu, et al. 2016; Konečný, Brendan McMahan, Ramage, et al. 2016; McMahan et al. 2017), the deployment of the current state-of-the-art methodology to clinical settings is still in its infancy. **FeatureCloud** is aiming to address this.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

In order to show the applicability in healthcare and medicine, we performed a thorough benchmark of the efficacy of federated ensemble learning methods on four standard clinical data sets and compared the results with the classical centralized machine learning approach. Data generated by biomedical research is rather different in many aspects compared to data from other domains. The heterogeneity within and between data sets, in particular concerning ethical, legal, or social confounders, as well as imbalances with phenotype prevalence or cohort sizes pose challenges for machine learning in general (Cios and Moore 2002). In this study, we emulated three common challenging scenarios among clinical research, (i) a different number of subsets; (ii) different sizes of subsets; and (iii) different balances of classes, and evaluated the corresponding model performances. As shown in the result section, no significant differences can be observed between the accuracy of classical models trained on entire data sets in comparison to the distributedly trained combined models. Solely, models of increasingly small sizes of samples lead to small but significant differences (#subsets >2). Moreover, in most cases, the combined models outperformed the underlying submodels.
Therefore, in medicine where limited sample sizes and biased data sets often hinder the transition from research to clinical practice, federated privacy-by-design learning approaches aid to overcome these boundaries by enabling collaborations across institutes without the need for tedious paperwork and lengthy processes since no exchange of data is needed. Thus, tree-based federated ensemble learning methods represent a valuable option for the implementation of the various federated health apps that will be deployed in the **FeatureCloud** platform.

In the future, we will focus on the optimization of the federated models to account for the heterogeneity and noisiness of biomedical data. Furthermore, these methods routinely have to account for inhomogeneous data sites, meaning all data sites have their data objects stored with different feature sets (Kargupta et al. 1999). This is particularly often the case with medical data records. In particular, FPDML can help to overcome data biases such as socio-economic and ethnic confounders more or less prevalent in all clinical data sets. Therefore, TFEL-based architectures have the power to revolutionize both clinical research and practice and pave the way for the precision medicine of the 21st century (Hamburg and Collins 2010).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

9 References

Batra, Richa, Nicolas Alcaraz, Kevin Gitzhofer, Josch Pauling, Henrik J. Ditzel, Marc Hellmuth, Jan Baumbach, and Markus List. 2017. “On the Performance of de Novo Pathway Enrichment.” *NPJ Systems Biology and Applications* 3 (March): 6.

Best, J., H. Bilgi, D. Heider, C. Schotten, P. Manka, S. Bedreli, M. Gorray, J. Ertle, L. A. van Grunsven, and A. Dechêne. 2016. “The GALAD Scoring Algorithm Based on AFP, AFP-L3, and DCP Significantly Improves Detection of BCLC Early Stage Hepatocellular Carcinoma.” *Zeitschrift Fur Gastroenterologie* 54 (12): 1296–1305.

Bologna, Guido, and Yoichi Hayashi. n.d. “Characterization of Symbolic Rules Embedded in Deep DIMLP Networks: A Challenge to Transparency of Deep Learning.” *Journal of Artificial Intelligence and Soft Computing Research* 7 (4): 265–86.

Bösl, Andreas, Andreas Spitzmüller, Zerina Jasarevic, Stefanie Rauch, Silke Jäger, and Felix Offner. 2017. “MammaPrint versus EndoPredict: Poor Correlation in Disease Recurrence Risk Classification of Hormone Receptor Positive Breast Cancer.” *PloS One* 12 (8): e0183458.

Boulesteix, A. L., S. Janitza, and J. Kruppa. 2012. “Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics.” *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1072?casa_token=rQehYc8v68MAAAAA:ioUjiE6bgHxKzFH8SLDunomtflz3-p25TNNzYZLiJr_EgYkySqPiW_THx2deepIhlROaikKMVwplBkY.

Brisimi, Theodora S., Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. 2018. “Federated Learning of Predictive Models from Federated Electronic Health Records.” *International Journal of Medical Informatics* 112 (April): 59–67.

Calvert, Jacob, Nicholas Saber, Jana Hoffman, and Ritankar Das. 2019. “Machine-Learning-Based Laboratory Developed Test for the Diagnosis of Sepsis in High-Risk Patients.” *Diagnostics (Basel, Switzerland)* 9 (1). https://doi.org/10.3390/diagnostics9010020.

Center for Devices, and Radiological Health. 2019. “What Are Examples of Software as a Medical Device?” U.S. Food and Drug Administration. 2019. https://www.fda.gov/medical-devices/software-medical-device-samd/what-are-examples-software-medical-device.

Choudhury, Olivia, Aris Gkoulalas-Divanis, Theodoros Salonidis, Issa Sylla, Yoonyoung Park, Grace Hsu, and Amar Das. 2019. “Differential Privacy-Enabled Federated Learning for Sensitive Health Data.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1910.02578.

Cios, Krzysztof J., and G. William Moore. 2002. “Uniqueness of Medical Data Mining.” *Artificial Intelligence in Medicine* 26 (1-2): 1–24.

Constable, Scott D., Yuzhe Tang, Shuang Wang, Xiaoqian Jiang, and Steve Chapin. 2015. “Privacy-Preserving GWAS Analysis on Federated Genomic Datasets.” *BMC Medical Informatics and Decision Making* 15 Suppl 5 (December): S2.

Davis, Damek. 2015. “Convergence Rate Analysis of Primal-Dual Splitting Schemes.” *SIAM Journal on Optimization*. https://doi.org/10.1137/151003076.

Desautels, Thomas, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, et al. 2016. “Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach.” *JMIR Medical Informatics* 4 (3): e28.

Fatima, M., and M. Pasha. 2017. “Survey of Machine Learning Algorithms for Disease Diagnostic.” *Journal of Intelligent Learning Systems and*. https://www.scirp.org/html/1-9601348_73781.htm.

Gan, W., J. C. W. Lin, and H. C. Chao. 2017. “Data Mining in Distributed Environment: A Survey.” *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1216?casa_token=-ZDKMqVx5NAAAAAA:eAtXl-V5ONyou4kwAYq5ohee2qw27nRjfZP92TrPYvQIEyCN6z2GGmBwWvnQckXPmvdIinkHKhkj6Yo.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Hamburg, Margaret A., and Francis S. Collins. 2010. “The Path to Personalized Medicine.” *The New England Journal of Medicine* 363 (4): 301–4.
Hamon, Ronan, Henrik Junklewitz, and Ignacio Sanchez. 2020. “Robustness and Explainability of Artificial Intelligence.” *Publications Office of the European Union*. https://core.ac.uk/download/pdf/286448151.pdf.
Holzinger, A. 2018. “From Machine Learning to Explainable AI.” In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66.
Holzinger, Andreas. 2005. “Usability Engineering Methods for Software Developers.” *Communications of the ACM* 48 (1): 71–74.
Jeanquartier, Fleur, Claire Jean-Quartier, Max Kotlyar, Tomas Tokar, Anne-Christin Hauschild, Igor Jurisica, and Andreas Holzinger. 2016. “Machine Learning for In Silico Modeling of Tumor Growth.” *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-50478-0_21.
Kargupta, Hillol, B. Park, Daryl Hershberger, and Erik Johnson. 1999. “Collective Data Mining: A New Perspective toward Distributed Data Mining.” *Advances in Distributed and Parallel Knowledge Discovery* 2: 131–74.
Kieseberg, Peter, Edgar Weippl, and Andreas Holzinger. 2016. “Trust for the Doctor-in-the-Loop.” *ERCIM News* 104 (1): 32–33.
Konečný, Jakub, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. “Federated Optimization: Distributed Machine Learning for On-Device Intelligence.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1610.02527.
Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. “Federated Learning: Strategies for Improving Communication Efficiency.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1610.05492.
Lænkholm, Anne-Vibeke, Maj-Britt Jensen, Jens Ole Eriksen, Birgitte Bruun Rasmussen, Ann S. Knoop, Wesley Buckingham, Sean Ferree, et al. 2018. “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.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 36 (8): 735–40.
Lazarevic, Aleksandar, and Zoran Obradovic. 2001. “The Distributed Boosting Algorithm.” In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 311–16. KDD ’01. New York, NY, USA: Association for Computing Machinery.
Li, Mu, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. 2014. “Scaling Distributed Machine Learning with the Parameter Server.” In *11th ${USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}$ 14)*, 583–98.
Liu, Yang, Yingting Liu, Zhijie Liu, Yuxuan Liang, Chuishi Meng, Junbo Zhang, and Yu Zheng. 2020. “Federated Forest.” *IEEE Transactions on Big Data*. https://doi.org/10.1109/tbdata.2020.2992755.
Lorenzi, Marco, Boris Gutman, Paul M. Thompson, Daniel C. Alexander, Sebastien Ourselin, and Andre Altmann. 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*, 10160:1016016. International Society for Optics and Photonics.
McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, edited by Aarti Singh and Jerry Zhu, 54:1273–82. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR.
Miller, Tim, Piers Howe, and Liz Sonenberg. 2017. “Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1712.00547.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Ming, Jing, Eric Verner, Anand Sarwate, Ross Kelly, Cory Reed, Torran Kahleck, Rogers Silva, et al. 2017. “COINSTAC: Decentralizing the Future of Brain Imaging Analysis.” *F1000Research* 6 (August): 1512.

Newell, Allen, J. C. Shaw, and Herbert A. Simon. 1958. “Elements of a Theory of Human Problem Solving.” *Psychological Review*. https://doi.org/10.1037/h0048495.

Park, Byung-Hoon, and Hillol Kargupta. 2002. “Distributed Data Mining: Algorithms, Systems, and Applications.” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.7841.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Ramana, Bendi Venkata, M. Surendra Prasad Babu, and N. B. Venkateswarlu. 2012. “A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis.” *International Journal of Computer Science Issues (IJCSI)* 9 (3): 506.

Riemenschneider, Mona, Alexander Herbst, Ari Rasch, Sergei Gorlatch, and Dominik Heider. 2017. “eccCL: Parallelized GPU Implementation of Ensemble Classifier Chains.” *BMC Bioinformatics* 18 (1): 371.

Rousseaux, Sophie, Alexandra Debernardi, Baptiste Jacquiau, Anne-Laure Vitte, Aurélien Vesin, Hélène Nagy-Mignotte, Denis Moro-Sibilot, et al. 2013. “Ectopic Activation of Germline and Placental Genes Identifies Aggressive Metastasis-Prone Lung Cancers.” *Science Translational Medicine* 5 (186): 186ra66.

Schork, Nicholas J. 2015. “Personalized Medicine: Time for One-Person Trials.” *Nature* 520 (7549): 609–11.

Slodkowska, Elzbieta A., and Jeffrey S. Ross. 2009. “MammaPrint™ 70-Gene Signature: Another Milestone in Personalized Medical Care for Breast Cancer Patients.” *Expert Review of Molecular Diagnostics* 9 (5): 417–22.

Strecht, Pedro, João Mendes-Moreira, and Carlos Soares. 2014. “Merging Decision Trees: A Case Study in Predicting Student Performance.” In *Advanced Data Mining and Applications*, 535–48. Springer International Publishing.

Sundhar Ram, S., A. Nedić, and V. V. Veeravalli. 2012. “A New Class of Distributed Optimization Algorithms: Application to Regression of Distributed Data.” *Optimization Methods & Software* 27 (1): 71–88.

Sweeney, Latanya, Akua Abu, and Julia Winn. 2013. “Identifying Participants in the Personal Genome Project by Name (A Re-Identification Experiment).” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1304.7605.

Tuladhar, Anup, Sascha Gill, Zahinoor Ismail, Nils D. Forkert, and Alzheimer’s Disease Neuroimaging Initiative. 2020. “Building Machine Learning Models without Sharing Patient Data: A Simulation-Based Analysis of Distributed Learning by Ensembling.” *Journal of Biomedical Informatics* 106 (June): 103424.

Voigt, Paul, and Axel Von dem Bussche. 2017. “The Eu General Data Protection Regulation (gdpr).” *A Practical Guide, 1st Ed. , Cham: Springer International Publishing*. https://link.springer.com/content/pdf/10.1007/978-3-319-57959-7.pdf.

Wang, Shuang, Xiaoqian Jiang, Yuan Wu, Lijuan Cui, Samuel Cheng, and Lucila Ohno-Machado. 2013. “EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning.” *Journal of Biomedical Informatics* 46 (3): 480–96.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, et al. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics* 45 (10): 1113.

Wiwie, Christian, Irina Kuznetsova, Ahmed Mostafa, Alexander Rauch, Anders Haakonsson, Inigo Barrio-Hernandez, Blagoy Blagoev, et al. 2019. “Time-Resolved Systems Medicine Reveals Viral Infection-Modulating Host Targets.” *Systems Medicine (New Rochelle, N.Y.)* 2 (1): 1–9.

Wolberg, W. H., and O. L. Mangasarian. 1990. “Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (23): 9193–96.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. “Federated Machine Learning: Concept and Applications.” *ACM Trans. Intell. Syst. Technol.*, 12, 10 (2): 1–19.
Yu, Hwanjo, Xiaoqian Jiang, and Jaideep Vaidya. 2006. “Privacy-Preserving SVM Using Nonlinear Kernels on Horizontally Partitioned Data.” *Proceedings of the 2006 ACM Symposium on Applied Computing - SAC ’06*. https://doi.org/10.1145/1141277.1141415.

10 Table of acronyms and definitions

AI	Artificial Intelligence
AUC	*Area Under the receiver operating characteristic Curve*
BCD	breast cancer diagnosis
concentris	concentris research management GmbH
DL	Deep Learning
EHR	Electronic Health Records
FPDML	Federated Privacy-by-Design Machine Learning
GDPR	General Data Protection Regulation
GND	Gnome Design SRL
HCC	Hepatocellular Carcinoma
ICA	independent component analysis
ILPD	Indian Liver Patient Data
IVA	independent vector analysis
LTD	lung tumor diagnosis
ML	Machine Learning
MS	Milestone
MUG	Medizinische Universitaet Graz
Patients	In this deliverable, we use the term “patients” for all research subjects. In FeatureCloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus to increase readability, we simply refer to them as “patients”.
PR AUC	*Precision-Recall AUC*
RI	Research Institute AG & Co. KG
ROC	*receiver operating characteristic*
SBA	SBA Research Secure Business Austria GmbH
SDU	Syddansk Universitet
TCGA	The Cancer Genome Atlas
TFEL	Tree-based Federated Ensemble Learning
tSNE	t-diStributed Nonlinear Embedding
TUM	Technische Universitaet Muenchen
UM	Universiteit Maastricht
UMR	Philipps Universitaet Marburg
WP	Work package

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

11 Other supporting documents / figures / tables (if applicable)

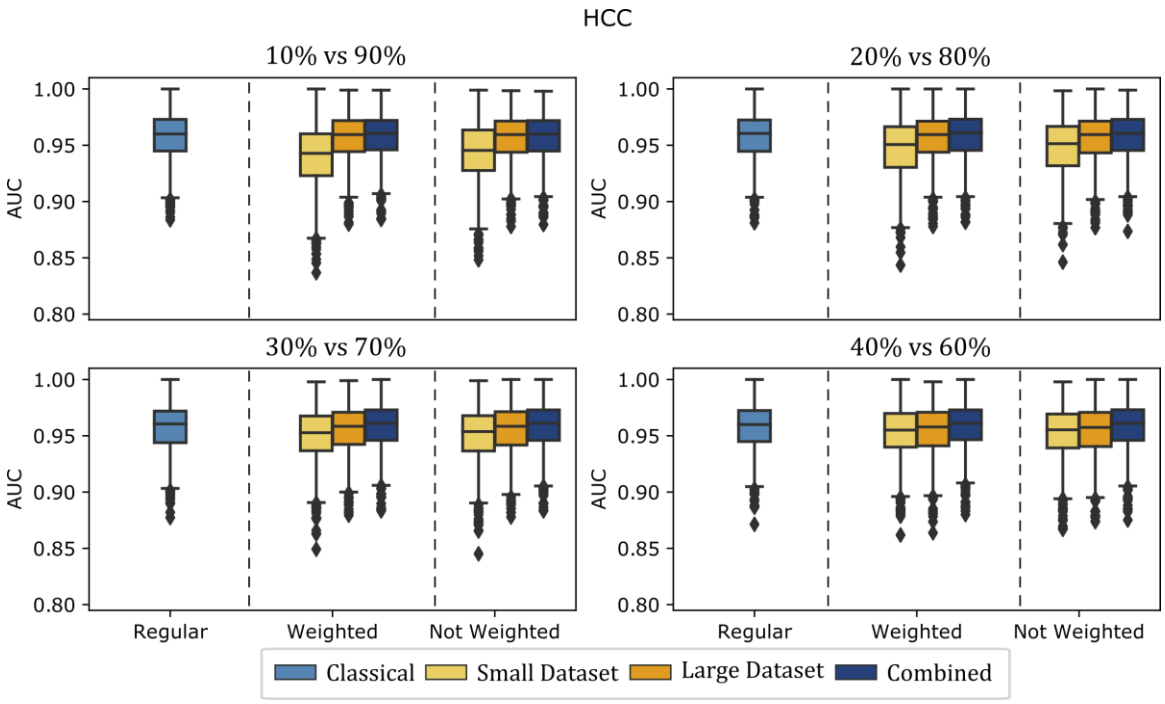

Figure 5. Example visualization of the results for the models trained on the HCC data set split by 10% / 90%, 20% / 80%, 30% / 70% and 40% / 60%. The boxplots show the performance of the models based on the smaller subsets (light yellow) and the larger subsets (dark yellow). Finally, we compare both the weighted and not weighted combined models.

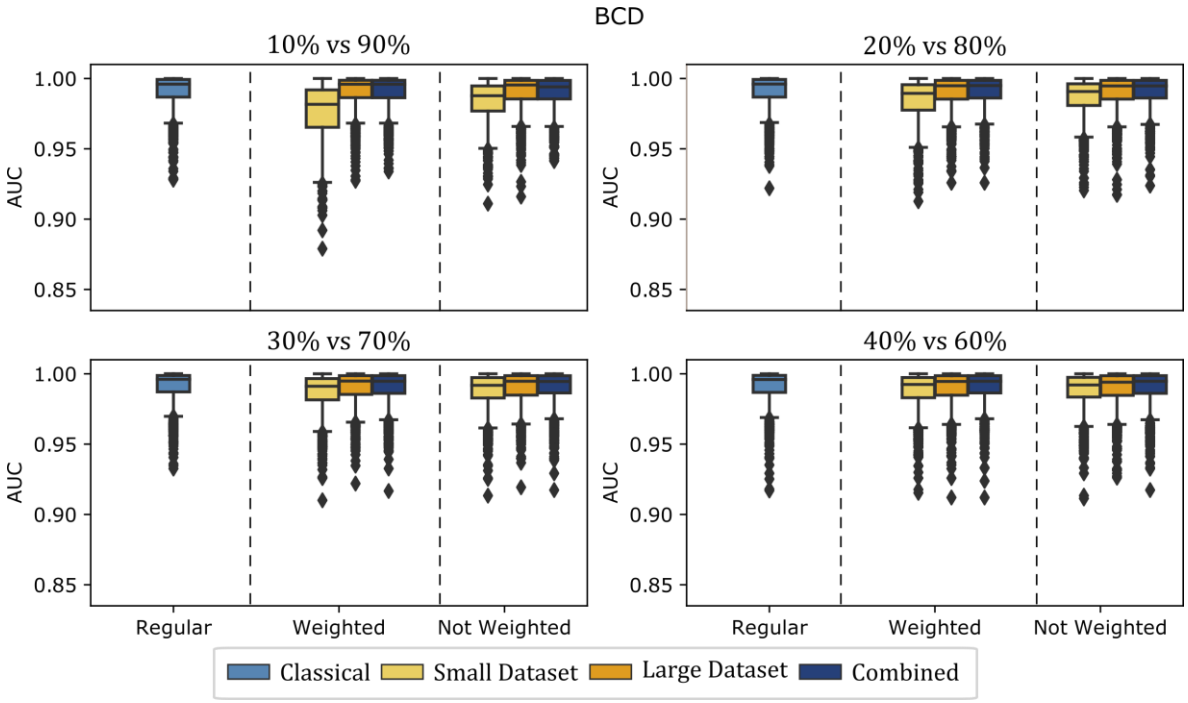

Figure 6. Example visualization of the results for the models trained on the BCD data set split by 10% / 90%, 20% / 80%, 30% / 70% and 40% / 60%. The boxplots show the performance of the models based on the smaller subsets (light yellow) and the larger subsets (dark yellow). Finally, we compare both the weighted and not weighted combined models.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

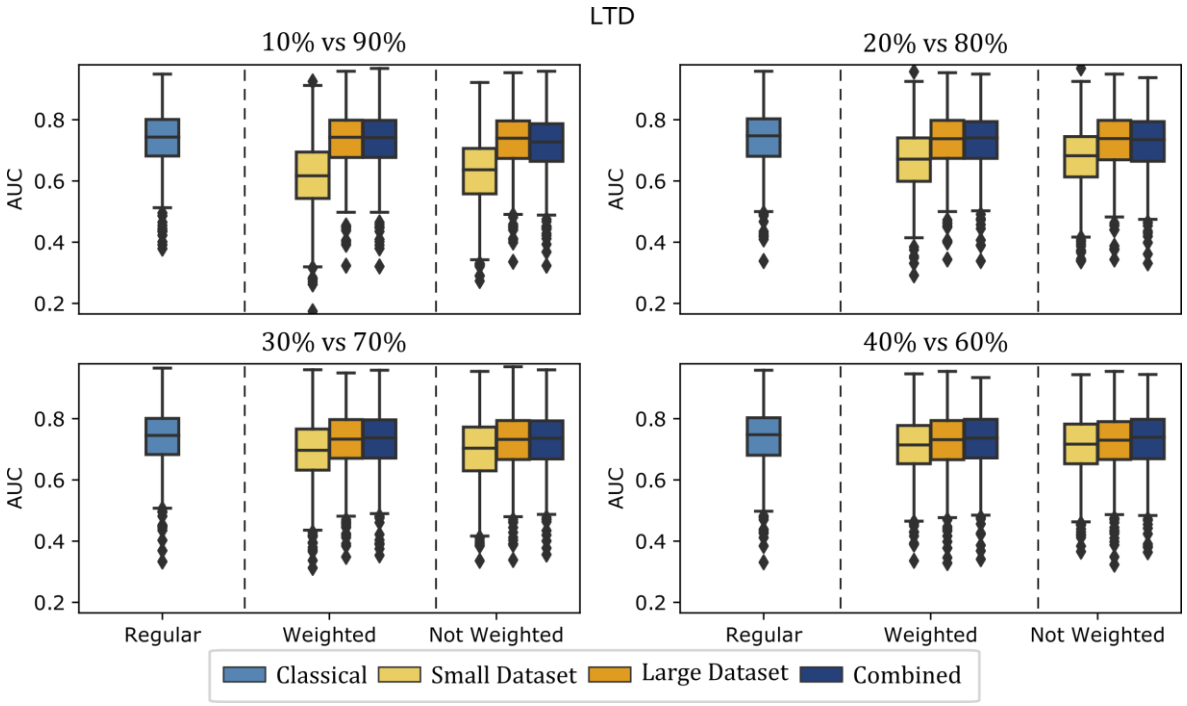

Figure 7. Example visualization of the results for the models trained on the LTD data set split by 10% / 90%, 20% / 80%, 30% / 70% and 40% / 60%. The boxplots show the performance of the models based on the smaller subsets (light yellow) and the larger subsets (dark yellow). Finally, we compare both the weighted and not weighted combined models.

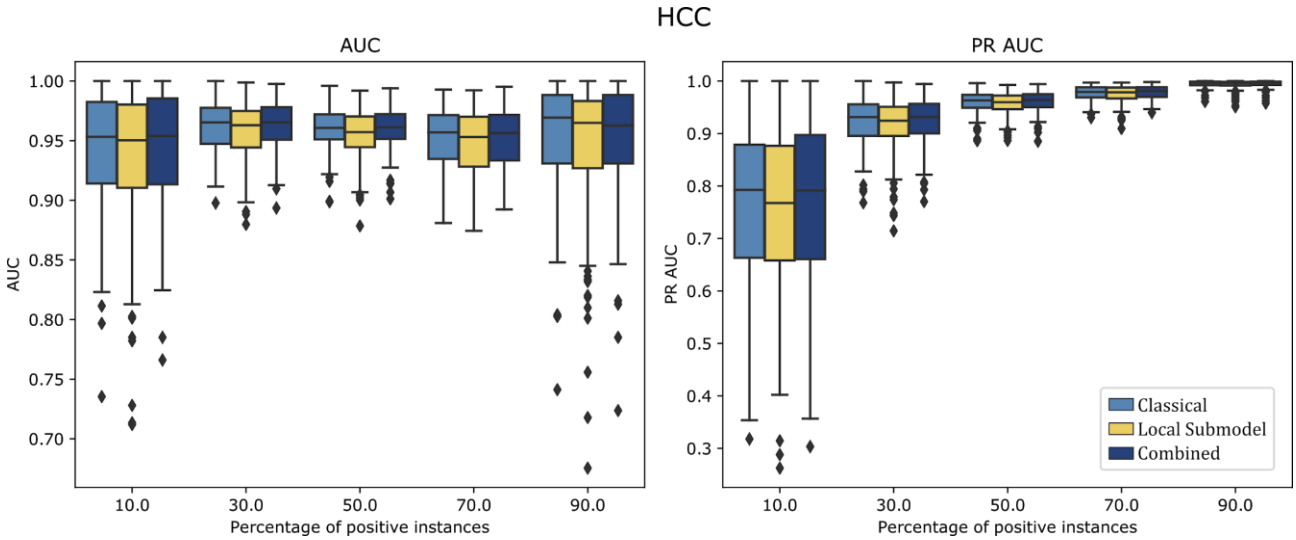

Figure 8. Comparison of the performance of the federated local and combined models with the classical model on differently unbalanced HCC data sets. The first box plot visualizes the AUC of models trained 10% - 90% positive samples respectively. The second box plot depicts the corresponding area under the PR curve.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

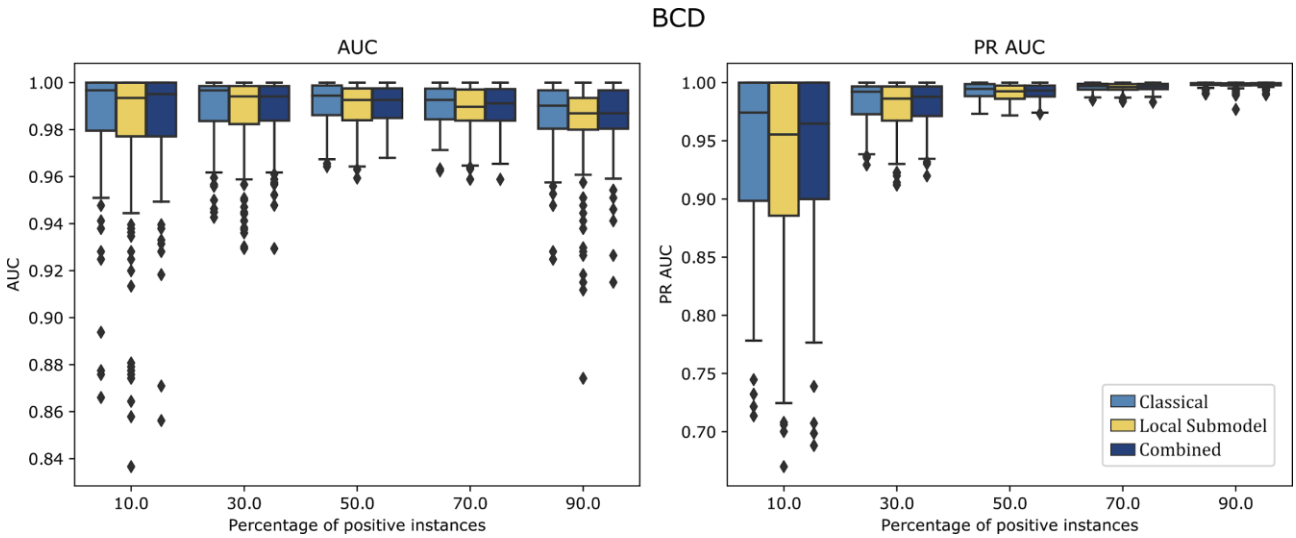

Figure 9. Comparison of the performance of the federated local and combined models with the classical model on differently unbalanced BCD data sets. The first box plot visualizes the AUC of models trained 10% - 90% positive samples respectively. The second box plot depicts the corresponding area under the PR curve.

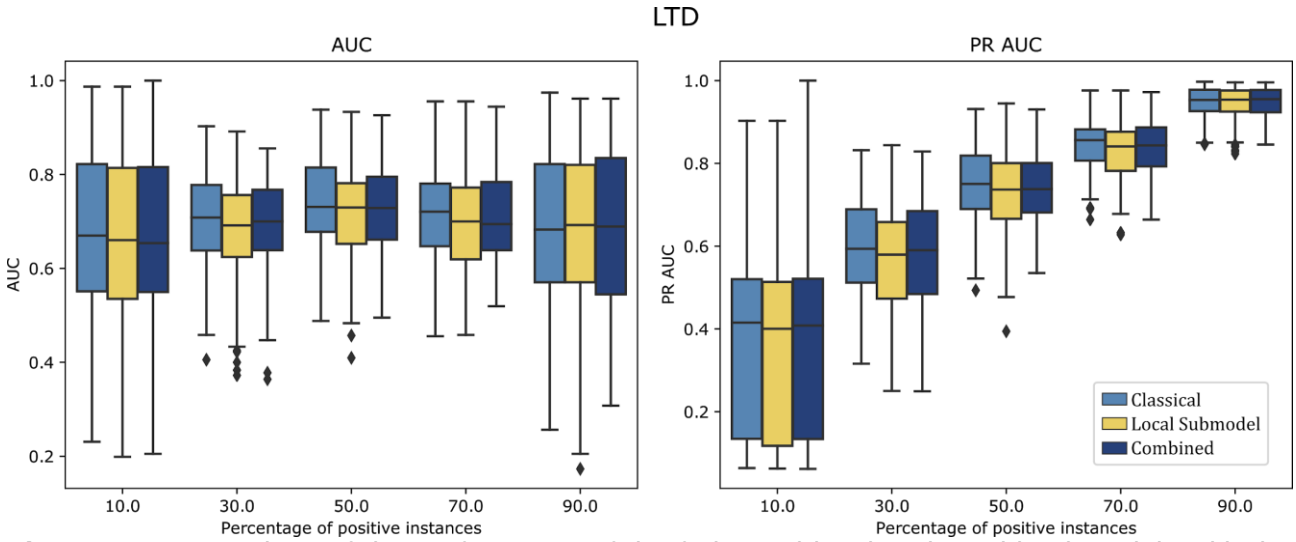

Figure 10. Comparison of the performance of the federated local and combined models with the classical model on differently unbalanced LTD data sets. The first box plot visualizes the AUC of models trained 10% - 90% positive samples respectively. The second box plot depicts the corresponding area under the PR curve.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.