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 D7.6
"Evaluation of federated vs. non-federated machine learning"

Work Package WP7
"Integrated FeatureCloud health informatics platform and App Store"

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	Christos Maramis, Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Jan Baumbach, University of Hamburg (UHAM)			
Deliverable	D7.6 - Evaluation of federated vs. non-federated machine learning			
Work Package	WP7 - Integrated FeatureCloud health informatics platform and App Store			
Date of Delivery	**Contractual**	30/06/2023 (M54)	**Actual**	30/06/2023 (M54)
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	07 UHAM			
Responsible Author	Mohammad Bakhtiari (UHAM)			
Keywords	Federated Learning, Non-Federated Learning, Machine Learning, Evaluation, Artificial Intelligence, AI, App Store, FeatureCloud			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	01/06/2023	First Draft	Mohammad Bakhtiari (UHAM)
V0.2	29/06/2023	Internal Review	Rudolf Mayer (SBA)
V1.0	30/06/2023	Final Version	Mohammad Bakhtiari (UHAM)

Actual effort in person-months (PMs)

Contributor (name and institution)	Invested resources (deliverable)	Overview of contributions
Mohammad Bakhtiari (UHAM)	2.0 PM	First draft and final version
Rudolf Mayer (SBA)	0.05 PM*	Internal Review

*This person dedicated a certain amount of time to FeatureCloud, but received no salary from the FeatureCloud budget (e.g. Professor, PI, Intern, Supervisor, Master/Bachelor student, etc.).

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

Table of Contents

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

1 Table of Acronyms and Definitions

BRCA	Breast Invasive Carcinoma
CV	Cross-Validation
D	Deliverable
DL	Deep Learning
DNA	Deoxyribonucleic Acid
GWAS	Genome-Wide Association Studies
f	federated
FL	Federated Learning
GTEx	Genotype-Tissue Expression (GTEx) project
HTTPS	Hypertext Transfer Protocol Secure
ILDP	Indian Liver Patient Dataset
ML	Machine Learning
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”.
RF	Random Forest
RI	Research Institute AG & Co. KG
RMSE	Root-Mean Squared Error
SBA	SBA Research Gemeinnützige GmbH
SHARE	Survey of Health, Aging and Retirement in Europe
SMPC	Secure Multi-Party Computation
SNP	Single Nucleotide Polymorphism
SVA	Surrogate Variable Analysis
TCGA	The Cancer Genome Atlas
UHAM	Universität Hamburg
WP	Work package

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

2 Objectives of the Deliverable based on the Description of Action (DoA)

This deliverable, evaluation of federated vs. non-federated machine learning, is tightly tied to the task 4 of WP 7, “Evaluation”, where we test federated machine learning using publicly available data and clinical trial data, comparing its performance to non-federated methods while ensuring data privacy and security. D7.6 is also related to the objective 4 of WP7, to implement automatized measures for evaluating the overall strategy of FeatureCloud by demonstrating that the performance of federated machine learning (in terms of accuracy) is comparable to the performance of traditional cloud-based approaches.

3 Executive Summary

This deliverable demonstrates the successful evaluation of federated applications in the FeatureCloud platform by comparing the results with corresponding non-federated models. Therefore, various applications like Flimma [1] (see sections 4.2), sPLINK [2] (see sections 4.1), Deep Learning [3] (see sections 4.3) are implemented as FeatureCloud apps (see sections 4.2.2, 4.1.2) and tested on different data. In the deliverable, we exemplary show it for the following datasets: The Cancer Genome Atlas (TCGA) Breast Cancer (BRCA) dataset [4] (see sections 4.2.1, and 5.2), the SHIP dataset [5] (see sections 4.1.1 and 5.1), and Survey of Health, Aging, and Retirement in Europe [6] (see section 5.3).

4 Introduction (Challenge)

Federated Learning (FL) emerged as a solution for respecting user privacy laws while enabling access to a wide range of distributed data for machine learning and data analysis approaches. Federated models are supposed to yield comparable results to the state of the art centralized approaches in a federated fashion while addressing various challenges like data heterogeneity, imbalance data, and at the same time requiring additional privacy enhancing technologies. In that regard, we designed and implemented federated applications inside the FeatureCloud ecosystem to test the comparable results to non-federated models in the FeatureCloud platform. Accordingly, in this deliverable, we will demonstrate federated machine learning, deep learning, and data analysis models, i.e., applications in FeatureCloud App Store, that achieve state-of-the art results in federated fashion on a variety of fields and data domains. We will cover three major peer-reviewed publications, sPLINK [2], Flimma [1], and FeatureCloud [3], to demonstrate the FeatureCloud platform is able to provide a solid solution for privacy concerns in federated fashion. In fact, in our endeavor to compare non-federated machine learning against corresponding federated applications, in detail. Meanwhile, we consider different simulated data heterogeneity levels and imbalanced data, and privacy considerations, to showcase the capabilities of FeatureCloud solution to address crucial challenges in Federated Learning.

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

5 Methodology

In this deliverable, we will provide an overview of a series of applications (based on peer-reviewed publications) that are integrated in the FeatureCloud platform for conducting federated collaboration to deliver results in various federated settings that are comparable to non-federated results. In the following subsections, we will demonstrate how sPLINK, Flimma, and in general FeatureCloud applied different methods while considering privacy issues in various fields (Table 1). We will summarize some of the applications that are implemented in the FeatureCloud ecosystem and manage to achieve results that are comparable to non-federated results, but in a federated fashion.

For evaluation of applications in the FeatureCloud platform, we have implemented a series of evaluation apps that calculate metrics based on model predictions that can, for example, be used to compare the federated vs non-federated results. As it is mentioned in Table 1, the apps Evaluation (Classification), Evaluation (Regression), and Evaluation (survival) are available in the App Store. Some applications have built-in model evaluation steps, such as the deep learning application.

5.1 sPLINK

sPLINK is a hybrid federated tool for privacy-aware, i.e., the raw data is not shared with third parties [2], genome-wide association studies (GWAS). sPLINK is initially implemented based on the HyFed [14] framework. HyFed (https://github.com/tum-aimed/hyfed) is a hybrid federated framework for privacy-preserving machine learning. It is designed to enhance the privacy of federated learning while maintaining the utility of the global model. HyFed provides developers with a generic API to develop privacy-enhanced algorithms and supports both simulation and federated operation modes.

sPLINK consists of four main components: a web application for configuring study parameters, a client for computing local parameters and sharing them, with additional noise added, a compensator for aggregating noisy values, and a server for computing global parameters. Unlike PLINK, sPLINK ensures privacy by keeping private data within each site and not revealing local parameter values to other parties. It is computationally efficient and supports multiple association tests. sPLINK offers advantages over meta-analysis approaches in terms of usability and robustness against data heterogeneity. It is easier to use, as clients only need to accept an invitation and select their datasets. It also produces consistent results even with imbalanced phenotype distributions or heterogeneous confounding factors, unlike meta-analysis tools that may lose statistical power in such scenarios. sPLINK uses Secure Multi-Party Computation (SMPC) for further enhancing the privacy while managing to deliver comparable results to centralized analysis and meta- analysis.

As it is shown in Figure 1, (1) The coordinator creates a new project through the WebApp component and (2) invites a set of cohorts to join the project; (3) the cohorts join the project and select the dataset using the client component. The project is started automatically, when all cohorts joined. The computation of the test results is performed in a an iterative manner, where the clients (4) obtain the global parameters from the server, (5) compute the local parameters, mask them with noise, and share the noise and noisy local parameters with the compensator and server, respectively; (6) the compensator aggregates the noise values and sends the aggregated noise to the server; the server calculates the global parameters by aggregating the noisy local parameters and the negative of the aggregated noise; (7) after the computation is done, the cohorts and coordinator can access the results. All communications are performed in a secure channel over HTTPS protocol. The cohorts can use Linux distributions, Microsoft Windows, or MacOS to run the client component.

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

Table 1. List of applications that were implemented and used in the FeatureCloud platform: All applications delivered comparable results to best achievable results by their non-federated counterparts [2].

Application	Type	Description
Ada boost	Machine learning	Classification model based on boosting trees
CACS forest	Machine learning	Random forest classifying patients into their CACS
Cox PH model	Survival analysis	Survival regression based on the “lifelines” library
Cross-validation	Preprocessing	Local splits for a k-fold cross-validation
Deep learning	Analysis	Deep neural networks implemented in PyTorch
Evaluation (Classification)	Evaluation	Evaluation with various classification metrics (e.g., accuracy)
Evaluation (Regression)	Evaluation	Evaluation with various regression metrics (e.g., mean squared error)
Evaluation (survival)	Evaluation	Evaluation of survival or time-to-event predictions
Flimma	Differential expression	Differential expression analysis based on limma-voom
Graph-guided random forest	Machine learning	Random forest classification, regression, and survival based on graphs
Kaplan-Meier	Estimator	Survival analysis, Survival function estimation, and log-rank test
Linear regression	Machine learning	Regression model
Logistic regression	Machine learning	Classification model
Nelson-Aalen estimator	Survival analysis	Hazard function estimation and log-rank test
Normalization	Preprocessing	Standardizing input data
One-hot encoder	Preprocessing	One-hot encoding for categorical variables
Random forest	Machine learning	Classification and regression model based on decision trees
Random survival forest	Survival analysis	Survival prediction based on scikit-survival
SVD	Machine learning	SVD for dimensionality reduction
sPLINK	GWAS	fGWAS based on PLINK
Survival SVM	Survival analysis	Survival prediction based on scikit-survival

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

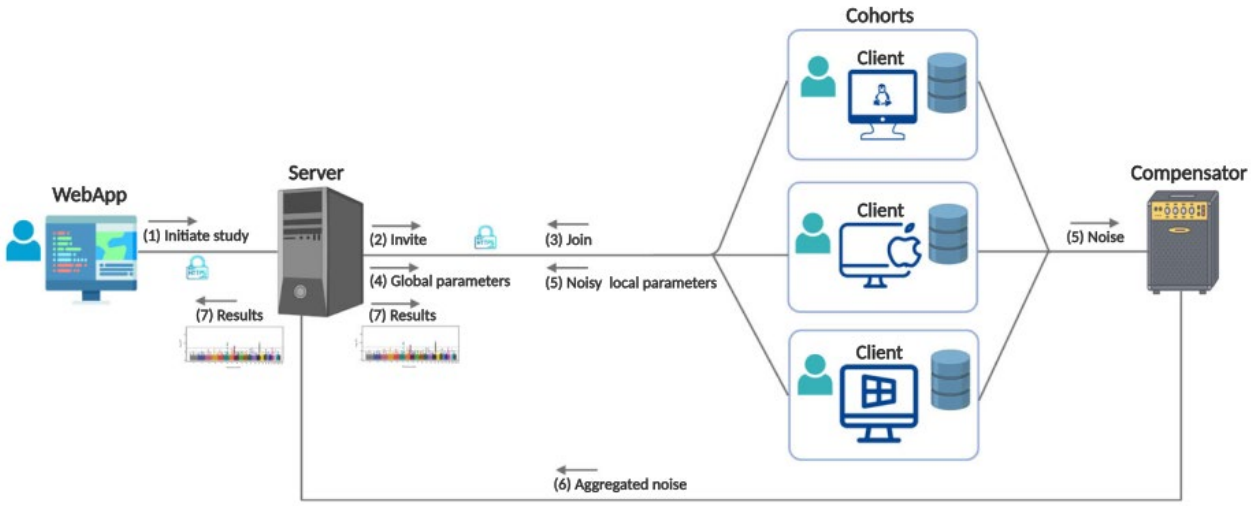


Figure 1. Architecture of sPLINK [2]

5.1.1 Datasets

sPLINK is applied on the SHIP dataset [5], accessible to researchers after completing a web-based request form and approval, the COPDGene dataset (http://www.copdgene.org/.), publicly available through dbGaP accession number phs000179.v1.p1., and the FinnGen [13] dataset, available for researchers by requesting access to the FinnGen Sandbox environment.

The SHIP dataset [5] refers to the Study of Health in Pomerania, a population-based cohort study conducted in Northeast Germany. The dataset includes information on various health-related factors, including genetics, lifestyle, and medical history. The FinnGen dataset is a large-scale research project that aims to identify genetic variants associated with various diseases and health-related traits in the Finnish population. It includes genomic data from over 500,000 individuals and is available for researchers by requesting access to the FinnGen Sandbox environment after completing training on how to deal with personal data and passing an exam about data security.

Table 2. Description of datasets used in sPLINK [2]. SNP, Single Nucleotide Polymorphism.

Dataset	# Samples	# SNPs	Adjustments	Phenotype
SHIP	3699	~5M	Sex, age, smoking status, daily alcohol consumption	SLA b, dichotomous (75th percentile, 934 cases, 2765 controls) SLA, quantitative, Mean ± SDc 1.23±0.3
COPDGene	5343	~600K	Sex, age, smoking status, pack years of smoking	COPD e, dichotomous, (2811 cases, 2532 controls) FEV1 f, quantitative, Mean ± SD 2.993±0.635
FinnGen	135,615	~ 1M	Sex and age	Hypertension, dichotomous, (34,257 cases, 101,358 controls)

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

Table 3. The SHIP case study [2].

Association test		Chi-square	Logistic regression	Linear regression
Split1	Sample size	229 | 712 | 941	229 | 712 | 941	941
	# of SNPs	5070067	5070067	5070067
Split2	Sample size	276 | 768 | 1044	276 | 768 | 1044	1044
	# of SNPs	5062964	5062964	5062964
Split3	Sample size	245 | 761 | 1006	245 | 761 | 1006	1006
	# of SNPs	5070192	5070192	5070192
Split4	Sample size	184 | 524 | 708	184 | 524 | 708	708
	# of SNPs	5077381	5077381	5077381
Aggregated	Sample size	934 | 2765 | 3699	934 | 2765 | 3699	3699
	# of SNPs	4878280	4878280	4878280

5.1.2 sPLINK FeatureCloud App

sPLINK is published as a certified app in FeatureCloud App Store (see Figure 2.), where it is documented how it can be used with sample data and in a federated collaboration. sPLINK app supports three algorithms: Chi-square, Linear Regression, and Logistic Regression.

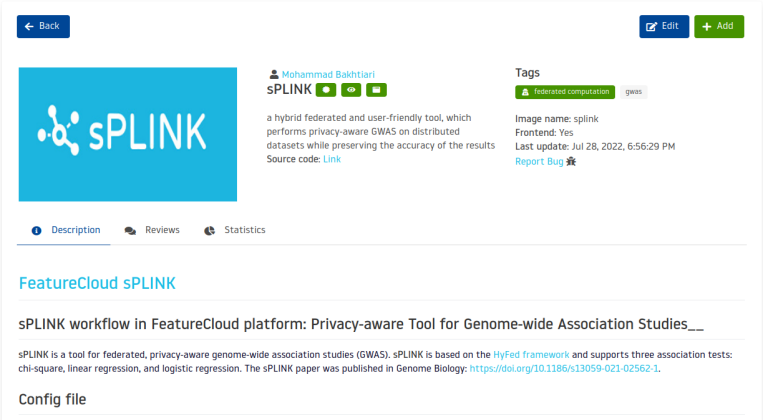


Figure 2. sPLINK app in the FeatureCloud App Store

For more information on how to use sPLINK with different algorithms, e.g., Chi-Square or Linear Regression, please visit the public GitHub repository at: https://github.com/FeatureCloud/fc-sPLINK

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

5.2 Flimma

Flimma [1] is a privacy-aware tool for differential expression analysis that implements a federated version of the limma-voom workflow. It operates on distributed cohorts without revealing sensitive data and uses a hybrid federated approach to hide local parameters from the server. Flimma was tested on two datasets, including a breast cancer expression dataset from TCGA [4] and a skin dataset from GTEx (https://gtexportal.org/home/datasets). It is robust to technical batch effects and models batch effects by adding binary covariates to the linear model.

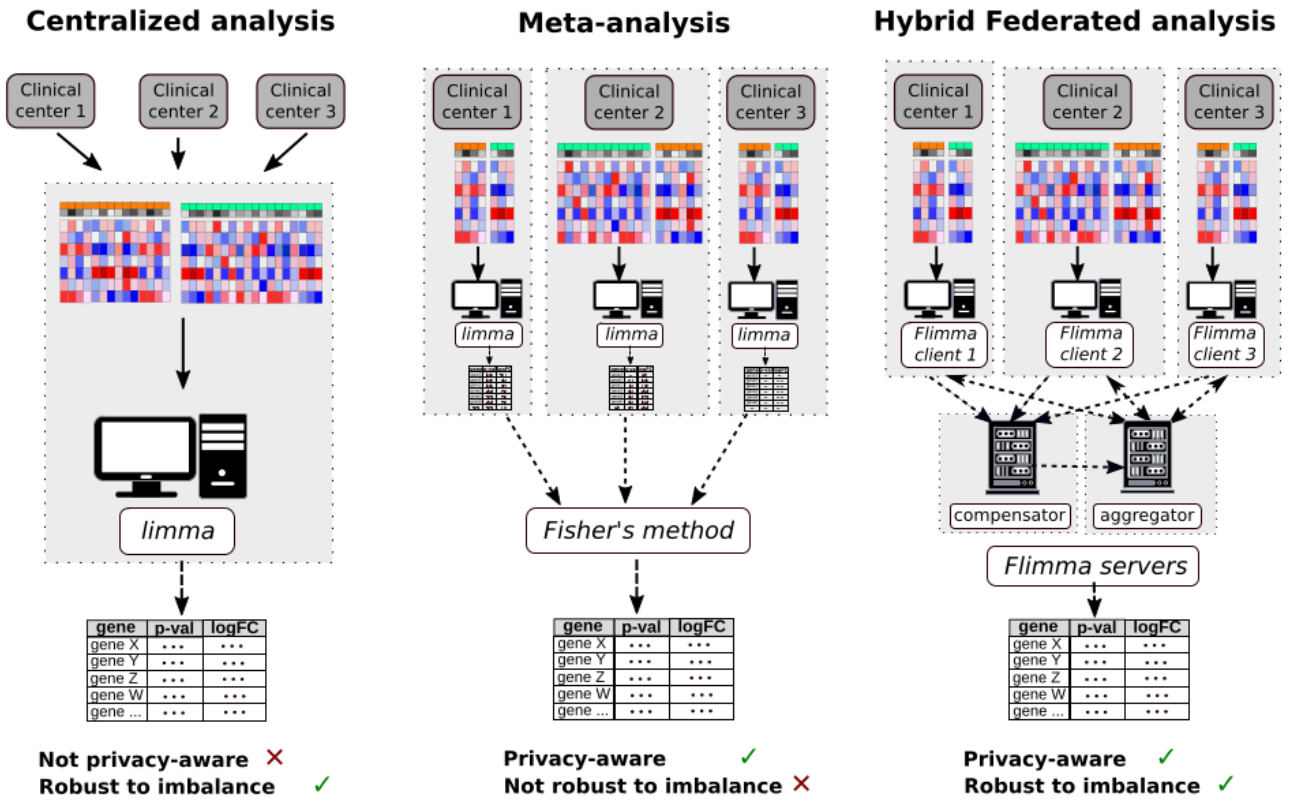


Figure 3. Gene expression analysis in case of multi-center studies. Bold arrows show the exchange of raw data, dashed arrows the exchange of model parameters or summary statistics. Gray areas highlight different physical locations [1].

As Figure 3. shows, Flimma used SMPC as a privacy enhancing technique for boosting privacy.

5.2.1 TCGA-Breast Cancer (BRCA) Dataset

The TCGA-Breast Cancer (BRCA) [4] dataset is a publicly available dataset that contains genomic, molecular, and histologic information on breast cancer. The dataset was extended to include additional histologic type annotations for a total of 1,063 breast cancers, and was analyzed to define transcriptomic and genomic profiles of six rare, special histologic types: cribriform, micropapillary, mucinous, papillary, metaplastic, and invasive carcinoma with medullary pattern. The dataset includes RNA-seq data, DNA copy number data, somatic mutation data, and histologic features data. The RNA-seq data was generated using Illumina HiSeq platform and is available in gene-level expression format. The dataset has been used to classify breast cancer into 12 consensus groups based on integrated genomic and histological features.

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

The DNA copy number data can be used to identify regions of the genome that are amplified or deleted in breast cancer. The somatic mutation data can be used to identify mutations that drive breast cancer development and progression. The histologic features data can be used to study the relationship between the molecular features of breast cancer and its histologic characteristics.

The TCGA-BRCA dataset has been used in various studies to identify new breast cancer subtypes and to understand the molecular mechanisms underlying breast cancer development and progression. For example, the dataset has been used to identify six rare, special histologic types of breast cancer and to define their transcriptomic and genomic profiles. The dataset has also been used to classify breast cancer into 12 consensus groups based on integrated genomic and histological features. These studies have improved our understanding of breast cancer and may lead to the development of new therapeutic approaches for this disease.

5.2.2 Flimma App in FeatureCloud App Store

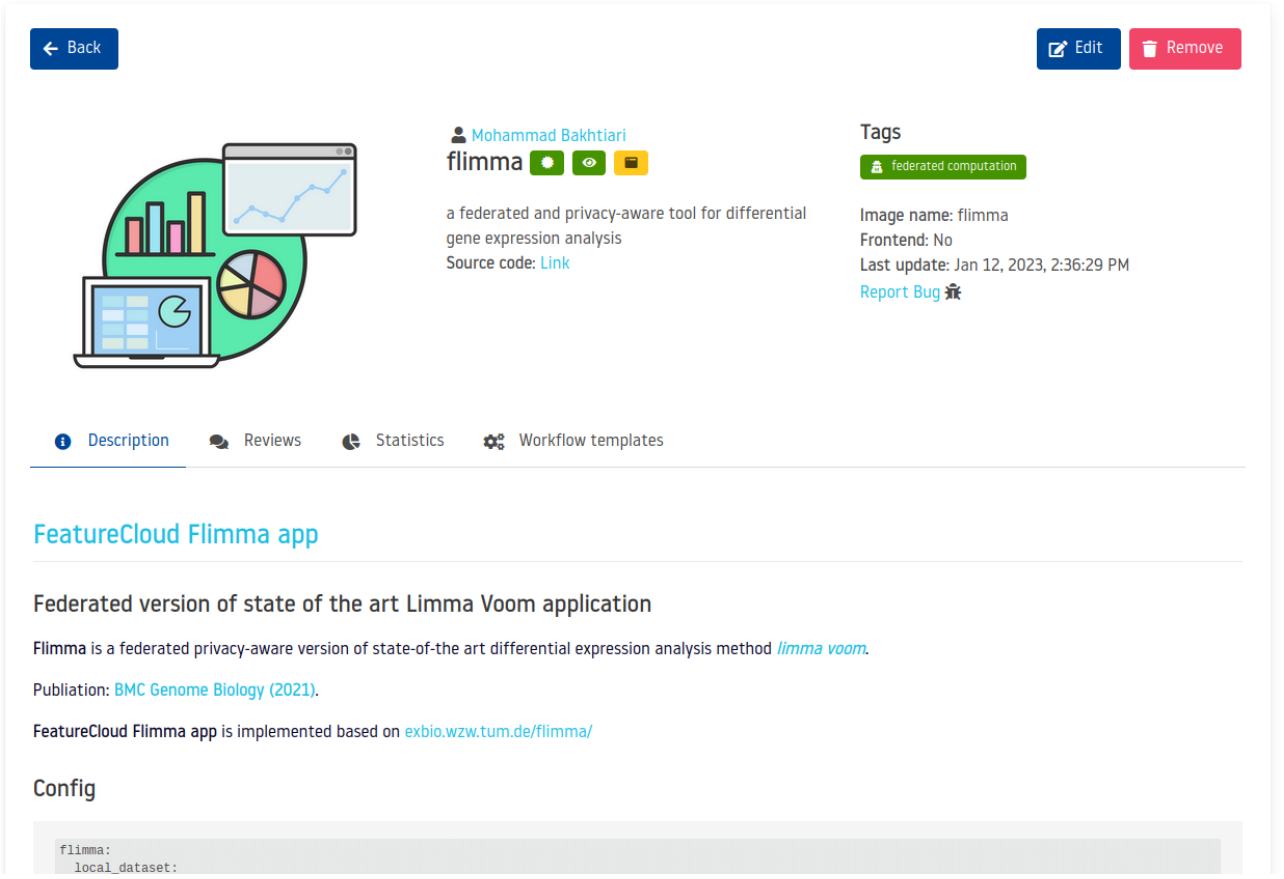


Figure 4. Flimma app

Flimma is an open source app which is available with documentation and sample data on GitHub: https://github.com/FeatureCloud/fc-flimma, and in the FeatureCloud App Store: https://featurecloud.ai/app/flimma

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

5.3 FeatureCloud

In this section we will cover applications that were implemented within FeatureCloud [3]. In general, the platform was tested using four different applications to achieve comparable results to centralized analysis in a federated fashion, considering federated learning challenges, e.g., data heterogeneity, imbalanced data, etc.

5.3.1 Federated Random Forest (RF)

We used the popular Random Forest (RF) classifier and RF regressor as the second algorithm for our evaluation. As an ensemble algorithm, RF can be easily federated in a naive manner [7]. Our implementation trains multiple classification or regression decision trees on the local primary data of each participant. The fitted trees are then transmitted to the coordinator and merged into a global RF. To account for the different number of samples for each participant, each of them contributes a portion of the merged RF proportional to the number of samples. To achieve a similar behavior as the centralized implementation, the size of the merged RF is kept constant, meaning that increasing the number of participants in turn decreases the number of required trees per participant. The federated computation occurs in three steps, each involving data exchange as follows: (1) participants indicate the number of samples and receive the total number of samples; (2) participants train the required number of trees, and the aggregator merges them into a global RF; and (3) participants receive the aggregated model to evaluate its performance on their data and share the results to obtain a global summary. As the aim is not to achieve the highest possible accuracy but to compare the federated version with the non-federated version, the hyperparameters were set to the default values of “sklearn”, namely, 100 decision trees, Gini impurity minimization as the splitting rule, and feature sampling equal to the square root of the features. Pre-pruning parameters such as maximum depth, minimum samples per node, and other constraints were not applied.

5.3.2 Federated Deep Learning (DL)

The federated deep learning (DL) application is based on the federated average algorithm [8]. In the training phase, an update of the weights and biases performed iteratively, where each iteration comprises the parameter aggregation performed in three steps as follows:

(1) the local weights and biases are computed by every participant individually and shared with the coordinator, (2) the coordinator averages the parameters and broadcasts them back to participants, and (3) the participants receive the new values of weights and biases and update the weights and biases of their model accordingly. The local weights and biases update is performed with the back-propagation algorithm, applied to data batches of a specified size. The neural network model architecture and training were implemented using the PyTorch library. The application enables the implementation of any DL architecture and provides a centralized version of a PyTorch code. The application also enables federated transfer learning to be applied to a pretrained model, whose specified layers are trained in the same federated fashion.

5.3.3 Federated Linear and Logistic Regression

For the implementation of the linear and logistic regression applications, the methods introduced by Nasirigerdeh et al [2] have been adapted from GWAS to a general ML use case. For linear regression, the local X^TX and X^TY matrices are computed by each participant individually, where X is the feature matrix and Y is the label vector. Then, they are sent to the coordinator, who aggregates the local matrices to the global matrices by adding them. Using these global matrices, the coordinator can calculate the beta vector (slope and intercept) through the federated method, and the beta vector, i.e., the estimated parameters via federated learning generally have very similar values to the estimated ones in centralized training Logistic regression was implemented as an iterative

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

approach (similar to the deep learning app). On the basis of the current beta vector, the local gradient and Hessian matrices of each participant are calculated and shared with the coordinator in each iteration. The coordinator aggregates the matrices again by adding them, updates the beta vector, and broadcasts it back to the participants. This process is repeated until convergence or the maximum number of iterations (prespecified for each execution) is achieved. Internally, the scikit-learn model API has been used to implement the applications [29,30]. In the performance evaluation, we used the default scikit-learn hyperparameters for the linear regression models. For logistic regression, the penalty was set to none; the maximum number of iterations was set to 10,000; and the “lbgs” solver was used to fit the models. These linear and logistic regression apps are each implemented as independent applications as this way they can be easily integrated into different workflow architectures with the FeatureCloud platform.

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

6 Results

6.1 sPLINK

In this section, we will provide an overview of the verification and comparison of sPLINK with the aggregated analysis conducted using PLINK, as well as with other existing meta-analysis tools (e.g., METAL and GWAMA). The analysis is performed on three different datasets: the SHIP dataset, COPDGene dataset, and FinnGen dataset.

In the SHIP dataset, which includes records of 3,699 individuals with serum lipase activity as the phenotype, sPLINK and PLINK are compared for logistic regression, chi-square test, and linear regression. The dataset is split into four parts using PLINK V1.9 [15] to simulate different cohorts, as shown in Table 3, and both tools calculate the same statistics for the association tests. The difference of SNP p-values between sPLINK and PLINK is found to be negligible, with a maximum difference of 0.162 attributed to floating-point precision inconsistencies. The correlation coefficient of p-values from both tools is high (0.99), indicating consistency. We investigate the overlap of significantly associated SNPs between sPLINK and PLINK. We consider a SNP as significant if its p-value is less than 5×10−8 (genome-wide significance). PLINK and sPLINK recognize the same set of SNPs as significant (Figure 5, d–f). Notably, the identified SNPs, e.g., rs8176693 and rs632111, lying in genes ABO (intronic) and FUT2 (3-UTR), respectively, have also been implicated in a previous analysis of this dataset [16].

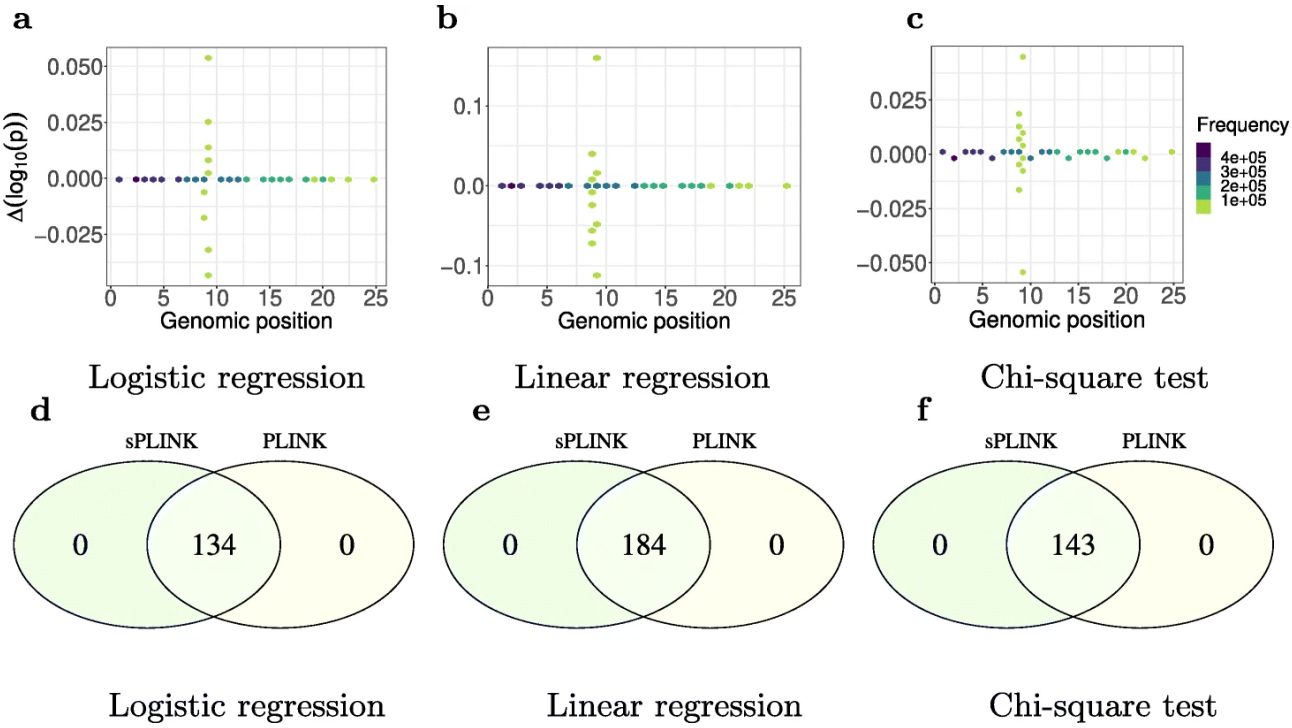
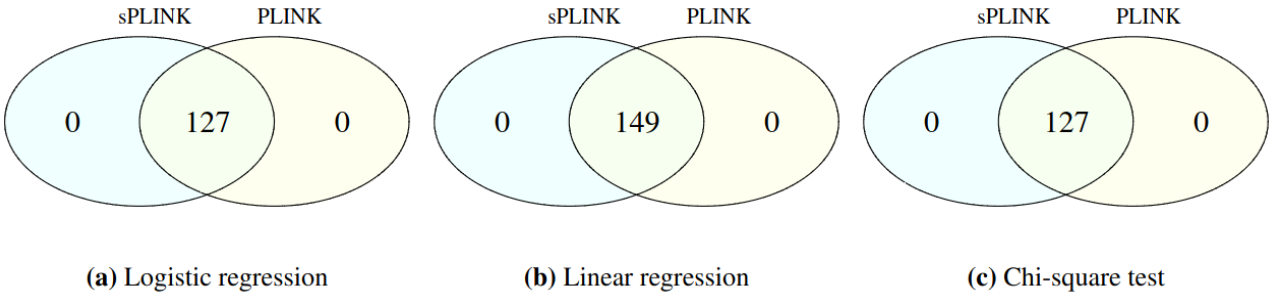


Figure 5. Δlog10(p-value) between sPLINK and PLINK as well as the set of SNPs identified by sPLINK and PLINK as significant for logistic regression (a, d), linear regression (b, e), and chi-square test (c, f), respectively. For most of the SNPs, the difference is zero, indicating that sPLINK gives the same p-values as PLINK. The negligible difference between p-values for the other SNPs can be attributed to differences in floating point precision. The spikes in some genomic positions are due to the strong association of the corresponding SNPs, which result in higher absolute error. sPLINK and PLINK also recognize the same set of SNPs as significant. [2].

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

Next, sPLINK is compared with PLINK, METAL, and GWAMA using the COPDGene dataset, which has an equal distribution of case and control samples, and the FinnGen dataset, which is much larger. In both cases, different phenotypes and confounding factors are considered. The comparison aims to assess the performance and accuracy of sPLINK in relation to other meta-analysis tools.



(a) Logistic regression **(b)** Linear regression **(c)** Chi-square test

Figure 6. The significant SNPs overlapped between sPLINK and PLINK for the SHIP case study considering Bonferroni significance threshold, which is ≈ 1 × 10−8 in our case. sPLINK and PLINK identify the same set of SNPs as significant [2].

Overall, the results indicate that p-values computed by sPLINK in a federated manner are comparable to those obtained from aggregated analysis using PLINK. Additionally, sPLINK shows promising results when compared to existing meta-analysis tools in different datasets, suggesting its effectiveness as an alternative for genetic association analysis.

6.2 Flimma

Flimma produces results in the form of a list of differentially expressed genes, along with their corresponding effect sizes, standard errors, t-statistics, and p-values. The results can be visualized using various methods such as volcano plots, heatmaps, and gene set enrichment analysis. Flimma also provides a statistical framework for assessing the significance of differential expression across multiple cohorts while accounting for batch effects and other sources of heterogeneity. Overall, Flimma provides a powerful and robust alternative to traditional meta-analysis methods for multi-center gene expression studies while enhancing patient privacy. In the following, we describe results in selected scenarios.

6.2.1 Imbalanced Scenario

Flimma has been tested on datasets with different levels of imbalanced data, where the fractions of target classes and the distributions of some covariates differed among cohorts. Flimma has been shown to perform well in both mildly and strongly imbalanced scenarios, where cohort sizes were unequal and related as 1:2:4 and 1:3:9, respectively. In addition, Flimma has been tested on the TCGA-BRCA dataset, where an imbalance of luminal and basal subtype frequencies was introduced. The results showed that Flimma was able to handle imbalanced data well and produced similar results to limma-voom in all tests.

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

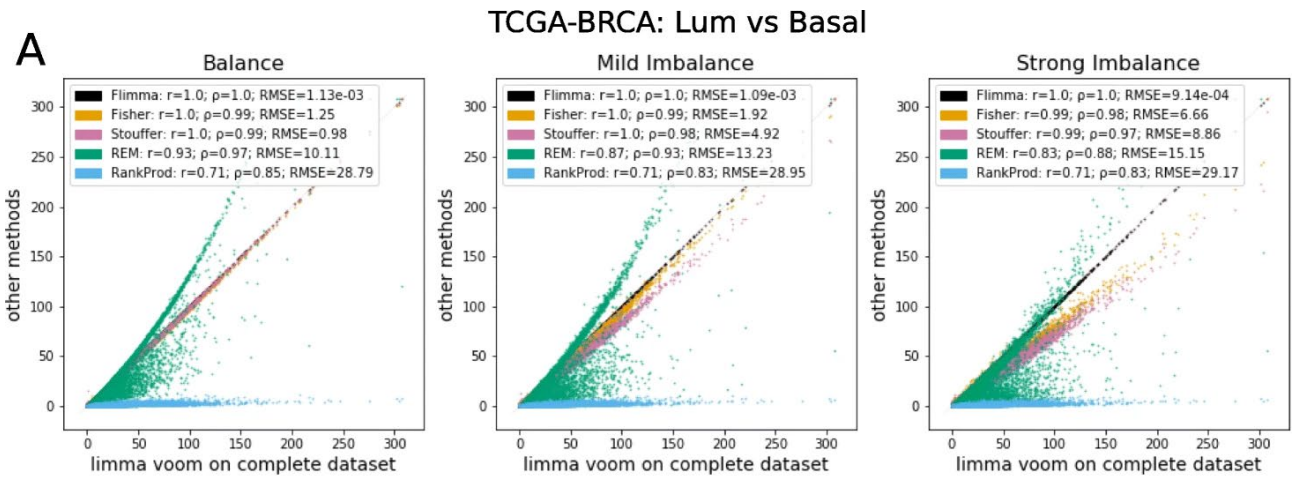


Figure 7. The comparison of negative log-transformed p-values computed by Flimma and meta-analysis methods (y-axis) with p-values obtained by limma on the aggregated dataset (x-axis) in three scenarios on GTEx skin datasets. Pearson correlation coefficient (r), Spearman correlation coefficient (ρ), and root-mean squared error (RMSE) calculated for each method are reported in the legend [1].

6.2.2 Performance on Top-ranked Genes

Flimma investigates the performance of different meta-analysis methods in identifying top-ranked genes that are significantly differentially expressed. The identification of top-ranked genes is important for research tasks such as biomarker discovery, where a small number of genes with large effect sizes are of interest. Flimma compares the performance of different meta-analysis methods in identifying top-ranked genes by altering the number of selected “top” differentially expressed genes after sorting by p-value. The results show that the gene rankings produced by all meta-analysis methods were comparable to the ranking produced by the aggregated limma-voom, and that Flimma performed well in identifying top-ranked genes across all scenarios tested.

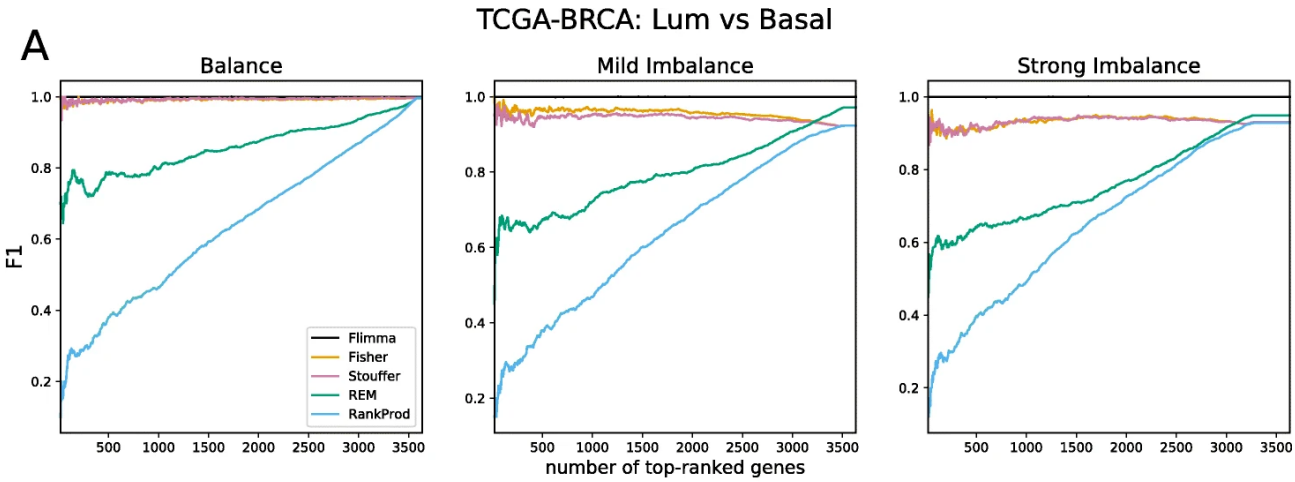


Figure 8. The dependency of the F1 score on the number of top-ranked genes considered to be differentially expressed. Genes were ranked in order of their negative log-transformed p-values decreasing and the number of top-ranked genes varied between 20 and 300 for GTEx Skin dataset with step 5 [1].

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

6.2.3 Performance in Presence of Batch Effects

Flimma addresses batch effects by including additional variables in the linear model to account for batch effects. This approach is known as the surrogate variable analysis (SVA) method. SVA estimates hidden factors that are correlated with batch effects and includes them in the linear model to adjust for these effects. Flimma also uses a modified version of the empirical Bayes method to estimate gene-specific variances, which helps to improve the accuracy of differential expression analysis in the presence of batch effects. Additionally, Flimma has been shown to be robust to batch effects, as demonstrated by its performance on publicly available breast cancer cohorts from GEO (https://maayanlab.cloud/archs4/) that were independently collected and sequenced at different laboratories and subjected to various experimental biases related to sample preparation, library construction, and sequencing platform.

Table 4. RMSE, precision, and recall comparison [1].

The number of cohorts		3	5	7	10	14
RMSE	Flimma	*0.0008*	*0.0007*	*0.0008*	*0.0017*	*0.0012*
	Fisher	0.94	1.82	2.53	3.86	5.37
	Stouffer	1.47	2.21	2.87	4.26	5.68
	REM	2.73	3.68	4.75	7.21	8.50
	RankProd	5.16	8.19	11.32	18.92	23.50
Precision	Flimma	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
	Fisher	0.85	0.88	0.90	0.93	0.95
	Stouffer	0.85	0.88	0.91	0.93	0.95
	REM	0.93	0.94	0.95	0.97	0.97
	RankProd	0.92	0.87	0.90	0.93	0.95
Recall	Flimma	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
	Fisher	0.92	0.95	0.95	0.96	0.97
	Stouffer	0.89	0.93	0.94	0.96	0.97
	REM	0.93	0.96	0.97	0.98	0.98
	RankProd	0.87	0.96	0.96	0.96	0.97

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

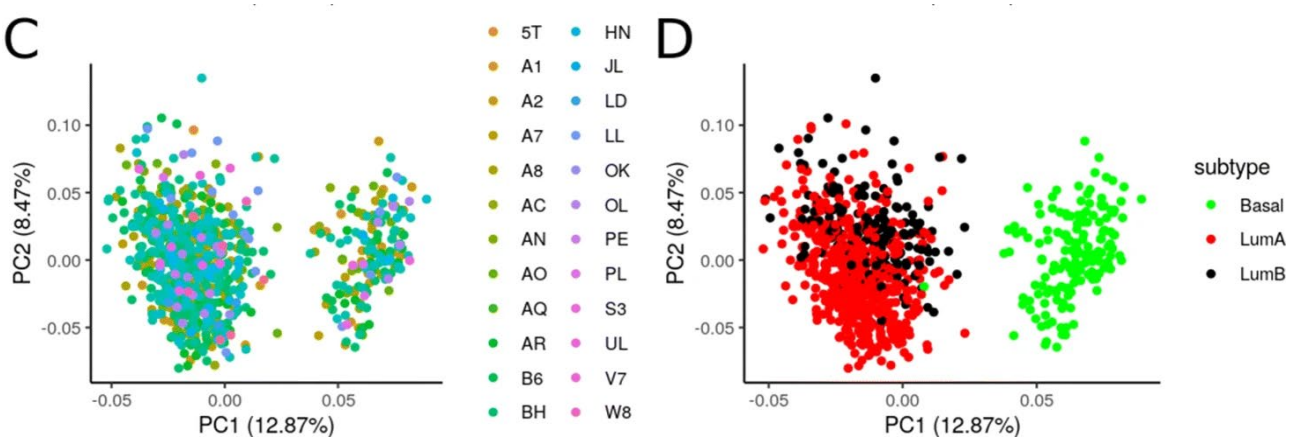


Figure 9. PCA projections [1]

Figure 9. shows the PCA projections computed and plotted by the proBatch R package of samples from three TCGA-BRCA cohorts. The samples are colored according to cohort and cancer subtype. The figure suggests that there is a clear separation between the different cohorts and cancer subtypes, indicating that gene expression patterns differ significantly between them. This observation supports the use of Flimma's federated approach for differential expression analysis, as it allows for the analysis of distributed cohorts without revealing sensitive data.

6.3 FeatureCloud

To evaluate the FeatureCloud, multiple workflows operating on different data sets were created. Except for DL, each workflow consists of a cross-validation (CV) application (10-fold CV), a standardization application, a model training application, and a final evaluation application. DL is evaluated on a 20% test set, as this is more common for big data to reduce the training time. Individual applications are data-type agnostic and are suitable for various applications. Classification analyses were performed on the Indian Liver Patient Dataset [9] with 579 samples and 10 features and the Cancer Genome Atlas Breast Invasive Carcinoma [10] data set with 569 samples and 20 features. For regression analyses, they were evaluated on the Diabetes [11][1] data set with 442 samples and 10 features and the Boston [12[2]] house prices data set with 506 samples and 13 features, both in the form provided by scikit-learn. Finally, for DL regression, we used a large data set from the Survey of Health, Aging, and Retirement in Europe [6], with 12 questionnaire variables and the target 12-item critical assessment of protein structure prediction quality of life score. After dropping samples with “Refusal” and “Don’t know” type values in those 12 variables and non-available 12-item critical assessment of protein structure prediction quality of life score, we were left with 42,894 (91.79%) out of 46,733 samples.

For each workflow, the central dataset is split into 5 participants with uneven data distribution. Participants 1, 2 and 3, and 4 and 5 each had 10% (4,289), 15% (6,434), and 30% (12,868) of the samples, respectively. The F1-score is used to evaluate the classification models and the root mean squared error for the regression models, as both are common metrics used to evaluate ML models.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html
[2] https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html

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

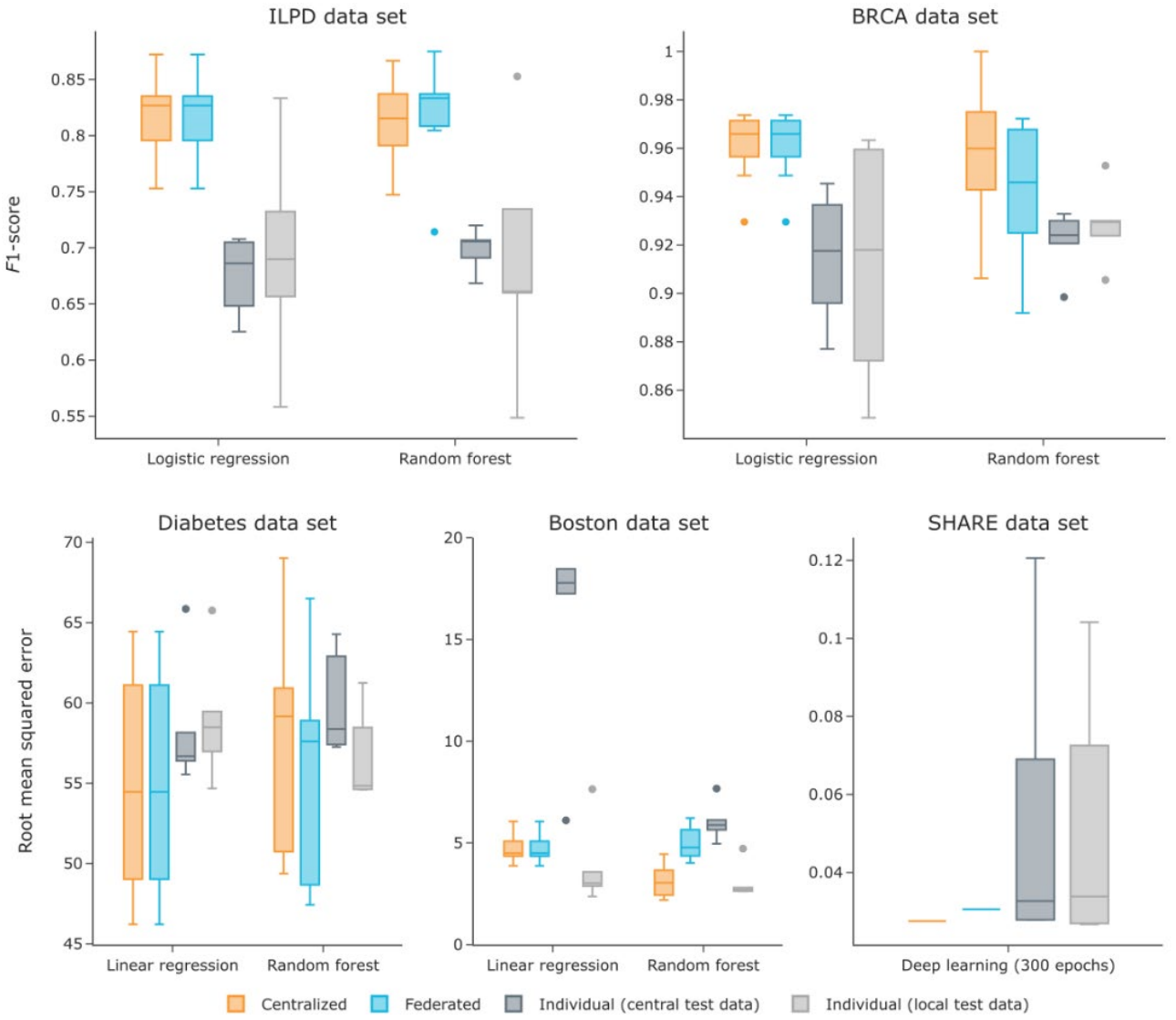


Figure 10. Performance evaluation of federated artificial intelligence methods. The box plots show the results of a 10-fold cross-validation for the different classification and regression models and data sets in multiple settings. Only the deep learning model was evaluated on a test set. The centralized results are shown in orange, the corresponding federated results in blue, and the individual results obtained locally at each participant in gray. Each model was evaluated on the entire test set (dark gray) such as the centralized and federated models and on the individual (local) parts of the test set (light gray). The federated logistic and linear regressions perform in identical fashion to their centralized versions, and the federated random forest and deep learning models perform in similar fashion to their centralized versions. BRCA: Breast Invasive Carcinoma; ILDP: Indian Liver Patient Dataset; SHARE: Survey of Health, Aging and Retirement in Europe.

The results showed that for logistic regression, linear regression, and random forest (RF) regression and classification models, the FeatureCloud workflow achieved performance identical or comparable to that of the centralized one, implemented with scikit-learn. However, due to the aggregation method and randomness in RF, identical results were not expected, and sometimes the federated RF performed even slightly better than the centralized approach.

The federated deep learning (DL) model trained in 300 epochs produced a very close root mean squared error compared to the centralized model. Additionally, the federated models were compared to individual models trained and evaluated by each participant using local test data. On average, the

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

local evaluation performance was worse than the federated models for classification. However, for regression models, the locally evaluated models sometimes outperformed the centralized model, although they didn't generalize well beyond the small test sets of individual participants.

The DL model evaluated on a larger test set performed more reliably than individual client models, which could have significantly worse results than the federated or centralized models. This emphasizes the effectiveness of FL, as it leverages more training and test data, leading to more generalized models. The passage also mentions that the RF application in FeatureCloud yields comparable results even when the data is non-independent and not identically distributed, outperforming the use of individual client data alone.

7 Conclusion

To summarize, the federated applications in the FeatureCloud App Store were tested in different domains to evaluate the performance against non-federated machine learning and data analysis models. We managed to not only achieve comparable results to centralized analysis in a federated fashion, but also designed federated scenarios to consider various federated challenges like data heterogeneity, imbalanced data in general. Besides, we applied privacy enhancing technologies like Secure aggregations techniques like SMPC on top of federated mechanisms to further enhance privacy preservation in FeatureCloud apps and platforms. Overall, the presented results in peer-reviewed publications demonstrate capabilities of FeatureCloud platform to provide a solid solution for privacy concerns in federated fashion.

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

8 References

[1] Zolotareva, O., Nasirigerdeh, R., Matschinske, J., Torkzadehmahani, R., Bakhtiari, M., Frisch, T., ... & Baumbach, J. (2021). Flimma: a federated and privacy-aware tool for differential gene expression analysis. Genome biology, 22(1), 1-26.

[2] Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Frisch, T., List, M., Späth, J., ... & Baumbach, J. (2022). sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. Genome Biology, 23(1), 1-24.

[3] Matschinske, J., Späth, J., Nasirigerdeh, R., Torkzadehmahani, R., Hartebrodt, A., Orbán, B., ... & Baumbach, J. (2021). The featurecloud AI store for federated learning in biomedicine and beyond. arXiv preprint arXiv:2105.05734.

[4] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015(1), 68-77.

[5] Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., ... & Hoffmann, W. (2011). Cohort profile: the study of health in Pomerania. International journal of epidemiology, 40(2), 294-307.

[6] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3 (pp. 265-284). Springer Berlin Heidelberg.

[7] Hauschild, A. C., Lemanczyk, M., Matschinske, J., Frisch, T., Zolotareva, O., Holzinger, A., ... & Heider, D. (2022). Federated Random Forests can improve local performance of predictive models for various healthcare applications. Bioinformatics, 38(8), 2278-2286.

[8] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

[9] Set, M. D. (2016). UC Irvine Machine Learning Repository URL: http://archive. ics. uci. edu/ml/datasets. Mushroom (28.03. 2016).

[10] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In Biomedical image processing and biomedical visualization (Vol. 1905, pp. 861-870). SPIE.

[11] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression.

[12] Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of environmental economics and management, 5(1), 81-102.

[13] FinnGen Consortium. (2021). FinnGen documentation of R4 Release.

[14] Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Baumbach, J., Rueckert, D., & Kaissis, G. (2021). HyFed: A hybrid federated framework for privacy-preserving machine learning. *arXiv preprint arXiv:2105.10545*.

[15] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

[16] Weiss, F. U., Schurmann, C., Guenther, A., Ernst, F., Teumer, A., Mayerle, J., ... & Lerch, M. M. (2015). Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. Gut, 64(4), 646-656.

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