`*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Federated Learning for the pathogenicity annotation of genetic variants in multi-site clinical settings

Nigreisy Montalvo[1], Francisco Requena[2], Emidio Capriotti[3] & Antonio Rausell[1,5,*]

[1] Université Paris Cité, INSERM UMR1163, Imagine Institute, Clinical Bioinformatics Laboratory, Paris, F-75006, France.

[2] Department of Physiology and Biophysics, Weill Cornell Medicine, Institute for Computational Biomedicine, Englander Institute for Precision Medicine, New York, NY, 10021, USA

[3] Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, 40126, Italy.

[4] Computational Genomics Platform, IRCCS University Hospital of Bologna, Bologna, 40126, Italy.

[5] AP-HP, Necker Hospital for Sick Children, Fédération de Génétique et Médecine Génomique, Service de Médecine Génomique des Maladies Rares, Paris, F-75015, France.

*To whom correspondence should be addressed

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Rare diseases collectively affect 5% of the population. However, fewer than 50% of rare disease patients receive a molecular diagnosis after whole genome sequencing. Supervised machine Learning is a valuable approach for the pathogenicity scoring of human genetic variants. However, existing methods are often trained on curated but limited central repositories, resulting in poor accuracy when tested on external cohorts. Yet, large collections of variants generated at hospitals and research institutions remain inaccessible to machine-learning purposes because of privacy and legal constraints. Federated learning (FL) algorithms have been recently developed enabling institutions to collaboratively train models without sharing their local datasets.

Results: Here, we present a proof-of-concept study evaluating the effectiveness of federated learn-ing for the clinical classification of genetic variants. A comprehensive array of diverse FL strategies was assessed for coding and non-coding Single Nucleotide Variants as well as Copy Number Variants. Our results showed that federated models generally achieved com-parable or superior performance to traditional centralized learning. In addition, federated models reached a robust generalization to independent sets with smaller data fractions as compared to their centralized model counterparts. Our findings support the adoption of FL to establish secure multi-institutional collaborations in human variant interpretation.

Availability: All source code required to reproduce the results presented in this manuscript, implemented in Python, is available under the GNU General Public License v3 at https://github.com/RausellLab/FedLearnVar.

Contact: antonio.rausell@institutimagine.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Rare diseases include a diverse range of more than 6172 clinical conditions that collectively affecting 4-5% of the population (Nguengang Wakap *et al.*, 2020). Approximately 70% of rare diseases are considered to be of genetic origin (Licata *et al.*, 2023). These are caused by high-impact germinal DNA defects including single-nucleotide variants (SNV) or large duplicated or deleted chromosomal regions referred to as Copy Number Variants (CNVs). Whole genome sequencing (WGS) has become a first-line genetic test for the diagnosis of rare diseases in the health system as it provides a comprehensive view of genomic alterations (Austin-Tse *et al.*, 2022). Yet, the assessment of the approximately 5 million genetic variants typically found in a single individual genome is still challenging, and only an average of 40% of patients receive a molecular diagnosis (Stranneheim *et al.*, 2021; Turro *et al.*, 2020).

Supervised machine learning (ML) has become a prominent bioinformatics strategy for the pathogenicity annotation of genetic variants, thanks to its ability to uncover complex patterns among genomic features (Eilbeck *et al.*, 2017; Zhu *et al.*, 2020). Here, models are trained on collections of genetic variants previously annotated as pathogenic or benign. Most state-of-the-art supervised ML scores for pathogenic variants prediction have been trained on curated genotype-phenotype databases, which encompass diverse sets of variant types, genes, diseases and phenotypic annotations (Brookes and Robinson, 2015). However, such reference databases contain only a fraction of all pathogenic variants identified to date across clinical and research institutions, due to sensitivity and privacy concerns. In the context of rare Mendelian diseases, even a single, ultra-rare genetic variant associated with a distinctive phenotype can, by itself, constitute identifiable information. Therefore, submission to public repositories typically requires explicit patient consent, approval from institutional ethics committees, and compliance with diverse national regulations, among other steps. As a consequence, currently available repositories of clinically annotated genetic variants remain limited in both size and heterogeneity, often resulting in supervised ML scores that generalize poorly to newly encountered patients or disease cohorts. (Bromberg *et al.*, 2024). Such limitations are even more pronounced in the case of genetic variants affecting non-coding genomic regions, i.e. those outside protein-coding sequences, whose functional and clinical impacts are still largely uncharacterized (Caron *et al.*, 2019; Ellingford *et al.*, 2022).

Multi-institutional collaboration can increase genetic and clinical data size and diversity, and therefore, improve the generalization of ML models for variant assessment. However, regulatory policies, such as General Data Protection Regulation (GDPR) for European Union (Hoofnagle *et al.*, 2019) and Health Insurance Portability and Accountability Act (HIPAA) in the United States of America (Annas, 2003), preclude genomic and clinical data sharing, as it represents threats to patient privacy. In genetic research, for example, data leaks can have harmful consequences for patients, such as genetic discrimination or data misuse (Bonomi *et al.*, 2020; Wan *et al.*, 2022).

Federated Learning (FL) has emerged as a promising solution for data-private collaboration in medicine and health (Sheller *et al.*, 2020; Sadilek *et al.*, 2021; Adnan *et al.*, 2022). In contrast to collaborative data sharing (CDS), where institutions need to centralize their local datasets for model training, FL proposes keeping the data decentralized and learning a consensus model, by aggregating locally-computed updates. In the traditional implementation, the clients (e.g., hospitals or health-research institutions) are coordinated by a central server, which defines and maintains a global ML model. At each round of FL, the server sends a copy of the model parameters to the clients for local training, and aggregates the local updates to derive a new global model, which is used in the subsequent round (**Figure 1**). This process is repeated until a maximum number of rounds is achieved or a different stopping condition is met.

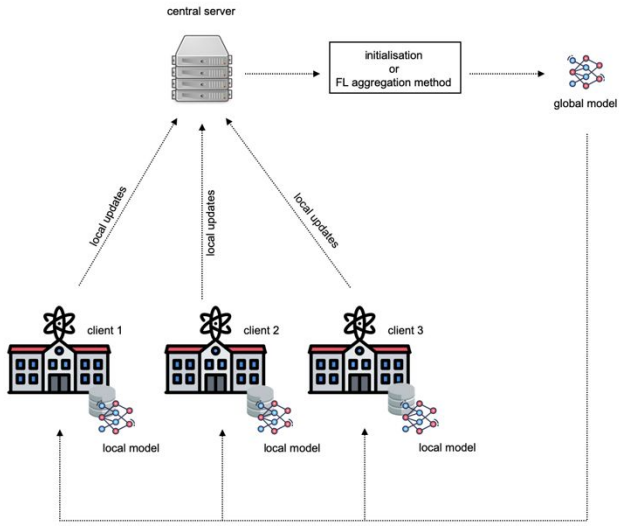


Fig. 1. General overview of Federated Learning training. A central server orchestrates the collaborative training of a global machine learning model across the clients. Instead of raw data, only model parameters are exchanged between the server and clients.

In recent years, FL has proven effective for secure genomic data sharing. Nasirigerdeh et al. (Nasirigerdeh *et al.*, 2022) presented sPLINK, a tool for the federated learning implementation of collaborative genome-wise association studies (GWAS). Experiments showed that sPLINK was robust against different sources of data heterogeneity, including the distribution of phenotypes and confounding factors. Raimondi et al. (Raimondi *et al.*, 2023) proposed a FL solution for multi-site exome-based risk prediction of Crohn's disease patients. The authors leveraged three public databases containing case and control Whole Exome Sequencing data to simulate a FL setting involving variable numbers of data owners. The results showed that the FL model improved the accuracy of models trained locally, even when the FL model was trained across 10 data owners who held very small datasets. More recently, Kolobkov et al. (Kolobkov *et al.*, 2024) proposed a simulated FL study for phenotype-from-genotype prediction and ancestry-from-genotype prediction on UK Biobank and the 1000 Genome Project. The authors showed that FL models were almost as accurate as the CDS model, and outperformed considerably the local models. To the best of our knowledge, however, studies evaluating the efficacy of FL for the pathogenicity annotation of genetic variants are currently lacking.

In this paper, we propose a proof-of-concept FL study for the pathogenicity annotation of genetic variants across independent institutions without raw data exposure. By leveraging the submitter information associated with each genetic variant in the publicly available ClinVar database (Landrum *et al.*, 2014), we mimicked three realistic multi-institutional collaborations for the clinical assessment of human genetic variants corresponding to three major types: coding SNVs, non-coding SNVs, and deletion CNVs. We then evaluated a comprehensive array of diverse FL strategies incorporating alternative network-based models, FL aggregation algorithms, local optimizers, client participation rates, FL server and client learning rates, collectively accounting for 1344 different settings (**Supplementary Table 1**). For each variant type, we systematically compared the performance of FL against that of CDS and

single-institutional models. Model performance was assessed through cross-validation on the training set as well as on two additional independent test sets, allowing us to evaluate the generalization capabilities of the classifiers across the three case studies. Further experiments evaluated model robustness to client dropouts as well as model behavior under identically or non-identically distributed features. Our study showed that overall, the performance of federated models is generally superior to local models, and can reach comparable or superior results to CDS.

2 Methods

Genetic variants data collection.

Coding and non-coding SNVs were extracted from ClinVar (version December 2020) as described in Capriotti and Fariselli 2023 (Capriotti and Fariselli, 2023). To adapt the dataset to a federated learning setting, we performed the following modifications (**Supplementary Figure 1**): We first split variants into two non-overlapping subsets: 1) SNVs reported before January 1st 2020 (referred to as *SNVs-Before-2020-01*), and 2) SNVs reported after January 1st (referred to as *SNVs-After-2020-01*). For coding SNVs, the multi-institutional training set was composed of SNVs from institutions having at least 1K coding SNVs in *SNVs-2020-01*, which resulted in 6 independent silos, with sizes ranging from 30,815 to 1,417 variants (**Supplementary Table 2**). Each institution subset was randomly downsampled in order to obtain a 1:1 balanced set of pathogenic and benign variants. In addition, we obtained 2 balanced independent test sets: a first set consisting on the remaining coding SNVs from *SNVs-2020-01* (10,378 variants) and a second set containing the coding SNVs from *SNVs-After-2020-01* (2,838 variants) respectively. In the case of non-coding SNVs, we obtained a multi-institutional training set formed by 8 institutions having at least 100 non-coding SNVs in *SNVs-Before-2020-01*, with sizes ranging from 2,288 to 101 (**Supplementary Table 3**). Similarly to coding SNVs, random downsampling was performed to obtain a 1:1 balanced set of pathogenic and benign variants and two independent test sets obtained containing 5,534 and 472 non-coding SNVs respectively.

In the case of CNVs, a high-confidence non-redundant set of pathogenic and benign deletion CNVs was obtained as described in Requena et al. (Requena *et al.*, 2022). Pathogenic and likely pathogenic deletion CNVs were obtained from ClinVar (version October 2021). Benign CNVs were obtained from reference databases and matched by genomic length with the pathogenic CNVs in ratio 1:1 [3]. Similarly to coding and non-coding SNVs, we split the dataset into two non-overlapping subsets: 1) deletion CNVs reported before January 1st 2021 (referred to as *CNVs-Before-2021-01*), and 2) deletion CNVs reported after January 1st (*CNVs-After-2021-01)*. For the purpose of this study, we considered benign deletion CNVs that had the same submission date as the corresponding pathogenic variant to which they were matched (see above). We derived a multi-institutional dataset by taking the deletion CNVs belonging to those institutions having at least 80 CNVs in *CNVs-Before-2021-01*, resulting in 8 independent silos, each with a 1:1 balanced set of pathogenic and benign variants, with sizes ranging from 6,936 to 82 variants (**Supplementary Table 4)**. We used the remaining deletion CNVs from *CNVs-Before-2021-01* to form the first independent test set, accounting for a total of 682 samples. A second independent test set was formed with the 96 samples from *CNVs-After-2021-01*. The largest silo in the obtained multi-institutional dataset held around 90% of the total data, while each of the remaining silos individually accounted for approximately 1-2%. In order to avoid bias during training, we decided to train centralized and federated models across the smaller clients and compare their performance with the model trained on the largest client.

Variant features annotation.

Both coding and non-coding SNVs were annotated with a total of 60 features following (Capriotti and Fariselli, 2023), including: i) 25 values representing the 5-nucleotide window sequence centered on the mutated position (5 times 5 possible nucleotides: A, C, G, T, N), ii) 10 values representing the conservation scores of 100-species (PhyloP100) and 470-species alignments (PhyloP470) of the five-nucleotide window sequence centered on the mutated position (Pollard *et al.*, 2010). In this study, we also annotated SNVs by using an additional 25 values mapping the conservation scores of 3-species (PhyloP3), 4-species (PhyloP4), 7-species (PhyloP7), 17-species (PhyloP17), and 20-species (PhyloP20) to the five nucleotide window positions.

CNVs were annotated with 38 features, grouped into gene-based and region-based features, as described in (Requena *et al.*, 2022). CNV gene-based features involved genes for which at least one base pair overlapped with the CNV genomic coordinates, based on Ensembl Gene (version 103), and using the GRCh37.p13 human reference genome. Genes were annotated using the following features: i) the probability of loss-of-function intolerance of the gene (pLI version 2.1.1, (Karczewski *et al.*, 2020)); ii) the loss-of-function observed/expected upper bound fraction (LOEUF score, version 2.1, (Karczewski *et al.*, 2020)); iii) the probability of being tolerant to both heterozygous and homozygous loss-of-function variants (pNull, version 2.1.1, (Karczewski *et al.*, 2020)); iv) the constrained coding region (CCR) score (Havrilla *et al.*, 2019); v) the enhancer domain (EDS) score (Wang and Goldstein, 2020); vi) predictions of haploinsufficiency or triplosensitivity, based on a meta-analysis of rare CNVs from 753,994 individuals (Collins *et al.*, 2022); vii) ohnolog genes, as reported in the OHNOLOGS database (version 2, (Singh and Isambert, 2020)), considering only pairs labeled as strict (highly reliable); viii) genes encoding transcription factors, according to data from the FANTOM consortium; ix) fitness cost due to gene inactivation, based on a genome-wide CRISPR-based score (CRISPR score, (Wang *et al.*, 2014)); x) involvement of the proteins encoded by the genes in a protein complex (as extracted from hu.MAP 2.0 (Drew *et al.*, 2021), with only proteins labeled with extremely high confidence selected); xi) mean and minimum gene expression across 54 tissues, obtained from the median transcripts per million (TPM) expression levels for each gene, provided by the Genotype-Tissue Expression Project (GTEx, version 8, (GTEx Consortium, 2013)), in which all GTEx tissues were considered; xii) mean PhastCons 46-way placental score (Siepel *et al.*, 2005), xiii) the CpG density of the promoter regions identified as 2 kb upstream and downstream from the transcription start site (TSS), defined as the first nucleotide of the transcript, according to previous work (Boukas *et al.*, 2020); and xiv) six network-based gene/protein features extracted from a protein-protein interaction network: degree, PageRank, and shortest path to proteins associated with haploinsufficient and triplosensitive genes respectively (Requena *et al.*, 2022). Gene-based features were transformed into categorical variables ("0" or "1" coding for the absence or presence, respectively, of at least one gene with the corresponding categorical feature), or quantitative variables encoding the maximum or minimum of the corresponding feature across the genes mapping within the CNV, except for minimum expression, the shortest path to haploinsufficient genes, and the shortest path to triplosensitive genes, for which the minimum was applied.

CNV region-based features considered included: i) the percentage of the CNV covered by each of the following six types of regulatory regions:

open chromatin regions, transcription factor binding sites (TFBS), promoters, promoter flanking regions, CTCF sites, and enhancers, identified on the H1 human embryonic stem cell line (H1-hESC) and obtained from the Ensembl Regulatory Build (version 2019-11-01, (Zerbino *et al.*, 2015)); ii) the maximum recombination rate (Halldorsson *et al.*, 2019), CADD score (version 1.6, (Kircher *et al.*, 2014)) and GERP scores (Davydov *et al.*, 2010) across the CNV genomic interval, with the scores previously summarized by their maximum value within non-overlapping 100 base pairs (bp) sliding windows; iii) Maximum gene density across the 1 megabase (Mb) sliding windows overlapping the CNV genomic interval; iv) four features encoding the presence or absence of CNV overlap (i.e. at least a 1 bp overlap) with the following regions of biological interest: a) human accelerated regions (HARs) (Capra *et al.*, 2013), b) lamina-associated domains (LADs) (Guelen *et al.*, 2008), c) ultra-conserved non-coding elements (UCNEs) (Dimitrieva and Bucher, 2013)), and d) structural variant (SV) hotspot regions (Ebert *et al.*, 2021); and v) two features encoding the distance (in Mb) to the centromere and the closest telomere regions respectively; genomic coordinates were retrieved from the UCSC Genome Browser "Gap" track (Haeussler *et al.*, 2019).

Federated learning settings

In this study we focused on a cross-silo and horizontal FL setting. Cross-silo refers to a small number of participants, typically organizations, such as genetic testing companies and research institutions in our context. The organizations are always available for local training and can participate in each round of FL training. The term horizontal indicates that clients' datasets share the same feature space, but hold different data samples. As illustrated in **Figure 1**, one round of FL encompasses the following steps: 1) The central server initializes a global ML model. In the experiment the models were initialized with random weights. 2) The central server then selects a subset of clients and broadcasts them a copy of the global model. Since we train neural network-based models, communicating the model refers to exchanging model weights. 3) The clients then use the local dataset for optimizing the received model through Stochastic Gradient Descent (SGD) for a predefined number of epochs. In the experiments, the number of local epochs was fixed to 10. 4) After completing the local training, the clients send the local updates (local models) to the central server. 5) The central server uses a FL aggregation method to combine the local updates into a new global model. Steps 2–5 are repeated for a predefined number of rounds, which was set to 200 in our experiments, or until a stopping condition is met.

Throughout the FL process, clients and central server kept the same neural network to ensure consistency in the learning process and allow for seamless aggregation of local updates from multiple clients. Different aggregation algorithms can be used by the server to integrate the local updates into a unified ML model. In this study we benchmarked the following FL aggregation algorithms: (i) FedProx (Li *et al.*, 2018) is a re-parametrization of Federated Averaging (FedAvg) (McMahan *et al.*, 2017), specifically designed to address statistical heterogeneity among clients. In FedAvg, the central server randomly selects a subset of clients for local optimization and aggregates their local models using a weighted averaging approach, where the weights are taken proportional to the clients' training dataset sizes. FedProx extends FedAvg by introducing a regularization term to the local training loss, denoted as μ, which constrains the local updates to be close to the global model. Notably, FedAvg is a special case of FedProx, when $\mu = 0$ and Stochastic Gradient Descent (SGD) is the local optimizer. (ii) FedAdagrad (Reddi *et al.*, 2020) is a federated version of the adaptive optimizer Adagrad (Lydia, A. A. and Francis, F. S.). The central server adjusts the learning rate for each model parameter based on the historical gradient information, therefore improving model convergence on non-IID data. (iii) FedAdam (Reddi *et al.*, 2020) is an adaptation of the adaptive optimizer Adam (Kingma and Ba, 2014) to the FL setting. The central server adjusts the learning rate for each model parameter by using the first moment (the mean) and the second moment (the uncentered variance) of the gradients. (iv) FedYogi (Reddi *et al.*, 2020) is a federated version of the adaptive optimizer Yogi (Zaheer *et al.*, 2018), which proposes modifications to the update rule of the second moment in Adam to improve model stability in non-convex optimization scenarios. In the literature, FedAdagrad, FedAdam, and FedYogi are considered FedOpt variants.

Neural network-based models.

Two neural network-based models were evaluated in this study: (i) a prototypical and well known Multilayer Perceptron architecture (MLP) (Popescu *et al.*, 2009), and (ii) a Shallow Neural Decision Forest (sNDF) (Kontschieder *et al.*, 2015), for which provide here a short background for non-familiar readers. sNDF is a supervised ML model that combines elements of neural networks and decision trees. The main goal of sNDF is unifying the representation learning capabilities of deep neural networks with the divide-and-conquer mechanism of decision trees. To that end, sNDF implements first a deep convolutional neural network (CNN) for feature learning from raw input data, and then uses the learned representation for training a forest of stochastic and differentiable decision trees. Similar to Random Forest (Breiman, 2001), each decision tree is trained using a subset of features. The final prediction of the forest is taken as the average of individual decision tree predictions. In this approach, the model takes the original feature vectors as input and feeds them into a forest of stochastic and differentiable decision trees, allowing for end-to-end training. More specifically, the original feature vectors are followed by a block of fully connected layers, with output units f_n (Kontschieder *et al.*, 2015). Each f_n is associated with a decision (split) node in a tree, with decision function $d_n(x) = \sigma(f_n(x))$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The output of d_n is interpreted as the probability of routing the sample to the left or right subtree. Once the sample reaches a leaf node l, the prediction of the tree is given by the class-distribution π_l. Further details on the training of the model can be found in (Kontschieder *et al.*, 2015).

Hyperparameter tuning and model training and evaluation.

To prevent, by design, downstream contamination of training and testing variants, we split training variant datasets by chromosome as proposed in (Capriotti and Fariselli, 2023; Requena *et al.*, 2022; Sharo *et al.*, 2022). More precisely, for each variant type and model configuration we trained a bundle of 23 classifiers, so that variants in a given chromosome were predicted using the classifier trained on variants from the remaining chromosomes. For example, we predicted CNVs in chromosome 2 by using a model trained on CNVs in chromosomes 1, 3, 4, etc. We refer to this approach as *leave-one-chromosome-out* training. We followed the guidelines in (Ogier du Terrail *et al.*, 2022) for hyper-parameter tuning by grid-search. We first tuned the hyperparameters of centralized ML models for further comparison with the federated version. For centralized MLP and sNDF we tuned the hyperparameters related to model architecture, batch size, learning rate, and optimizer. Specifically, we trained a 3-layer MLP and considered the following hyperparameters for this model: 1) number of neurons in the hidden layer $\in \{3, 4, 5, 6, 7, 8, 9\}$, 2) batch size $\in \{4, 8, 16, 32\}$, 3) learning rate $\in \{0.1, 0.01, 0.001, 0.0001\}$, 4) optimizer: SGD (with a momentum of 0.9) and Adam (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$). We fixed weight decay to 0.0001. For sNDF we considered the same hyper-parameter sets for batch size, learning rate, optimizer and weight

decay. We also explored the following hyperparameters related to model architecture: 1) number of trees ∈ {3, 5, 10, 15}, depth of the trees ∈ {3, 6, 9}, feature rate ∈ {0.5, 0.6, 0.7, 0.8, 0.9} (referring to the rate of original features used for training each individual tree in the forest). We used the best hyperparameters found per model (based on area under ROC curve) for training single-site models as well. The architecture, batch size, and weight decay of the best centralized models was reused for training the federated versions, as well as for the centralized non-differentiable models (random forest and XGboost, for which equivalent sNDF hyperparameters were applied).

We optimized the FL-related hyperparameters that, for practicality, we divided into general, and FL aggregation method-specific. Among general hyperparameters we considered: 1) local learning rate ∈ {0.1, 0.01, 0.001, 0.0001}, 2) local optimizer: SGD (with momentum 0.9) and Adam (with $\beta1 = 0.9$ and $\beta2 = 0.999$), and 3) client rate ∈ {0.5, 1.} (referring to the number of clients selected randomly for local training). For FedProx we also consider proximal term $\mu \in \{0.5, 0.1, 0.01, 0.001, 0.0001, 0\}$. For FedAdam, FedAdagrad, and FedYogi we also considered: 1) server learning rate ∈ {1, 0.1, 0.01, 0.001}, 2) $\tau \in \{0.0001, 0.01, 0.1\}$, representing the adaptivity parameter. For FedAdam and FedYogi we fixed $\beta1 = 0.9$ and $\beta2 = 0.999$. In the case of the training of federated MLP we also considered the inclusion (or exclusion) of local batch normalization layers. **Supplementary Table 5** outlines the parameters retained for CDS and FL models for the pathogenicity annotation of coding SNVs, non-coding SNVs, and CNVs. Hyperparameters of CDS and FL models were tuned with 10-fold collaborative cross-validation (CCV) on the entire training data. As outlined in (Linardos *et al.*, 2022), CCV proposes coordinating cross-validation across participants in the federation, so that each institution splits its local dataset into training and validation sets. Therefore, the final validation set is the aggregation of each institutional validation set. To ensure fairness in the experiments, we used the same splits for comparing CDS and FL approaches. We performed CCV with 3 different seeds for model weight initialization. All models were trained to 200 epochs, with early stopping if the best model did not improve for 20 consecutive epochs. After finding the best parameters, we trained centralized (CDS), single-institutional (local), and FL models on the entire training set, and evaluated them in the two held-out independent test sets. In this case, we repeated the experiments with 30 different seeds for model weight initialization to obtain a more accurate estimation of model performance.

Generation of identically-distributed settings for evaluation in federated learning models.

The study of the potential impact of identical or non-identical feature distributions in the FL process was carried out here following a strategy described in (Ogier du Terrail *et al.*, 2022), where identically distributed (IID) partitions were generated from the original training set. In order to do so, we took the complete training set and randomly distributed the variants across clients while respecting their original sample size. Then we trained the FL models on such new data partitions for 30 seeds for model initialization. We repeated the random data partition 100 times to obtain a histogram with the distribution of median AUC ROC values obtained for each training set.

Implementation

We implemented all the experiments using Flower (Beutel *et al.*, 2020), a Python library that enables the federation of several ML frameworks, including PyTorch. The experiments were performed on a high-performance computing (HPC) cluster, consisting of 18 computed nodes, each equipped with between 16 and 24 CPUs.

3 Results

3.1 Realistic scenarios of collaborative training for genetic variant assessment across independent institutions

In this work we evaluated the performance in the classification of human genetic variants as pathogenic or non-pathogenic of a diverse set of supervised learning approaches trained in a Federated Learning (FL) manner across multiple institutions, and compared it to the collaborative data sharing (CDS) and individual-institution model counterparts **(Methods)**. To that aim, we took advantage of the publicly-available ClinVar database, which reports the pathogenic classification of genetic variants primarily submitted by genetic testing laboratories and research institutions. ClinVar database allows the association of each variant to its submitter institution, which is leveraged in this study for mimicking a realistic multi-institutional collaboration with or without data sharing. Three major types of genetic variants were independently evaluated: coding SNVs, non-coding SNVs and CNVs. For each variant type, we defined a multi-institutional dataset for training FL and CDS models, as well as two independent test sets for assessing the performance of the classifiers (**Methods** and **Supplementary Table 1**). **Figure 2** shows the distribution of the training data across the institutions considered for each case study. Exact size of silos in the multi-institutional training set and in two independent test sets for each case studies are provided in **Supplementary Tables 2 - 4**.

The uneven distribution of variants observed across the participant institutions naturally provided different FL scenarios. In the case of coding and non-coding SNVs, the fraction of contributed instances showed a large heterogeneity across institutions, with contributors ranging from a maximum of 64.8% and 45.5% to a minimum of 3% and 2%, respectively (**Figure 2**). In contrast, CNVs presented a high concentration of variants in a single institution (90%) as compared to a panel of 7 small contributors whose contribution ranged between 1.2% and 2.2%. Thus, the first two cases were well adapted for a comparative analysis between the CDS and FL models involving all institutions. However, the third case appeared better suited for an evaluation comparing the performances of the models trained on the largest contributor against those resulting from the collaboration among the smaller ones, either in a CDS or a FL manner.

3.2 Performance assessment of alternative federated learning model architectures

For each of the case studies described above we implemented two neural network-based models: a simple Multilayer Perceptron (MLP) as the first model (Popescu *et al.*, 2009) using ReLU activation functions, and a machine learning model combining neural networks and decision trees, named Shallow Neural Decision Forest (sNDF) (Kontschieder *et al.*, 2015) (**Methods**). For either the MLP or the sNDF models, we explored alternative FL strategies depending on the type of aggregation algorithm and the client rate used. The aggregation algorithm, defined as the mechanism used by the central server to combine the local updates into a unified ML model (**Figure 1**), plays a key role in FL, as it directly impacts model accuracy and convergence. Consequently, we benchmarked four FL aggregation algorithms, namely FedProx, FedAdagrad, FedAdam and FedYogi (Li *et al.*, 2018; Reddi *et al.*, 2020). Further details about the specificities of each method are provided in **Methods**. Second, we

explored different values for FL client rate – the percentage of clients considered in a given round for local training by randomly selecting the clients. In addition, we evaluated the impact of the inclusion or exclusion of local batch normalization layers during the training of federated MLP, since several studies have proposed its use for improving FL model performance under non-independent and non-identically distributed (non-IID) data conditions (Li *et al.*, 2021; Andreux *et al.*, 2020).

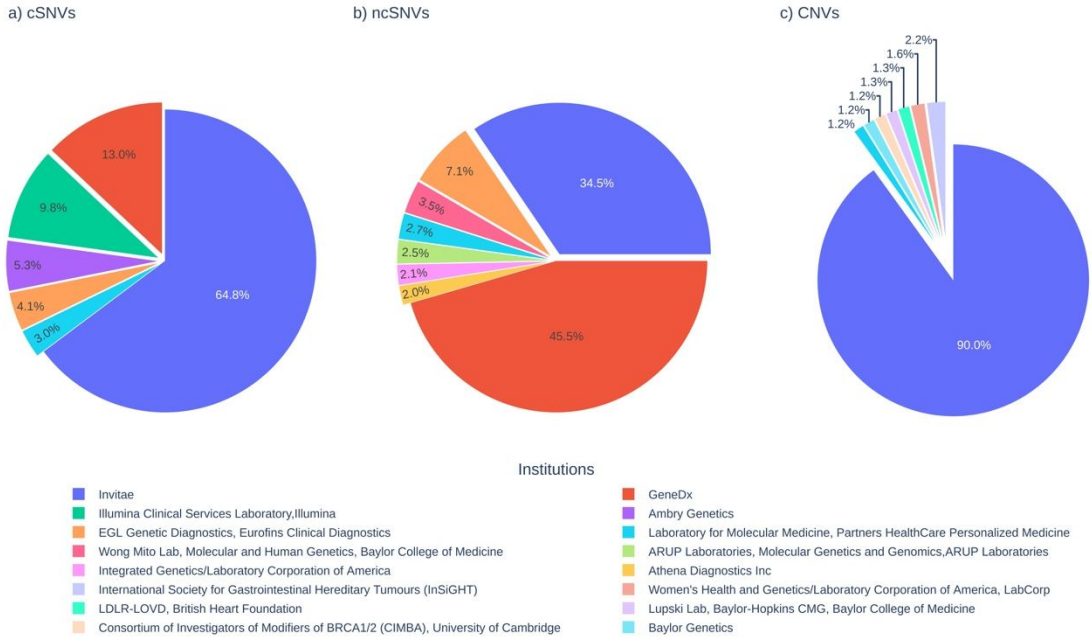


Fig. 2. Distribution of the training data across institutions considered in the experimentation. Distributions are shown for a) coding Single Nucleotide Variants (cSNVs), b) non-coding Single Nucleotide Variants (ncSNVs), and c) deletion Copy Number Variants (CNVs). Detailed information about the institutions included in the multi-institutional coding SNVs, non-coding SNVs and CNV dataset derived from the ClinVar database is provided in Supplementary Table 2-4, respectively. Tables list the names and locations of the contributing institutions, along with the original dataset size and the number of pathogenic and benign samples associated with each institution.

We assessed model performance through cross-validation on the training set as well as on two additional independent test sets, allowing us to evaluate the generalization capabilities of the classifiers (**Methods**). The two independent test sets were designed to consider two complementary aspects: 1) variants belonging to cohorts from sites not having participated in the FL process, and 2) variants reported later than those variants used for FL training (**Supplementary Table 6**). The different FL settings evaluated are summarized in **Supplementary Table 1**. Additional details on the collection of datasets, feature annotations, ML models used for training, and FL aggregation strategies are provided in **Methods**.

Supplementary Figure 2 displays the area under the receiver operating characteristic (AUC ROC) curve obtained through cross-validation on the training set for the three types of genetic variants considered and across the different settings evaluated. **Supplementary Table 5** outlines the best hyperparameters retained for each evaluation setting. The results showed that FedProx generally outperformed the alternative FL aggregation strategies evaluated. In addition, we observed that a FL client rate of 50% generally led to improved performance across all FL aggregation algorithms. Finally, for the MLP models, the inclusion of a batch normalization layer was generally detrimental. Overall, the two alternative learning strategies, MLP and sNDF, were similarly competitive when using the optimal FL settings, i.e., FedProx optimization, a 50% client rate, and no batch normalization layer. Results observed in the two independent tests confirmed these general trends (**Supplementary Table 6**).

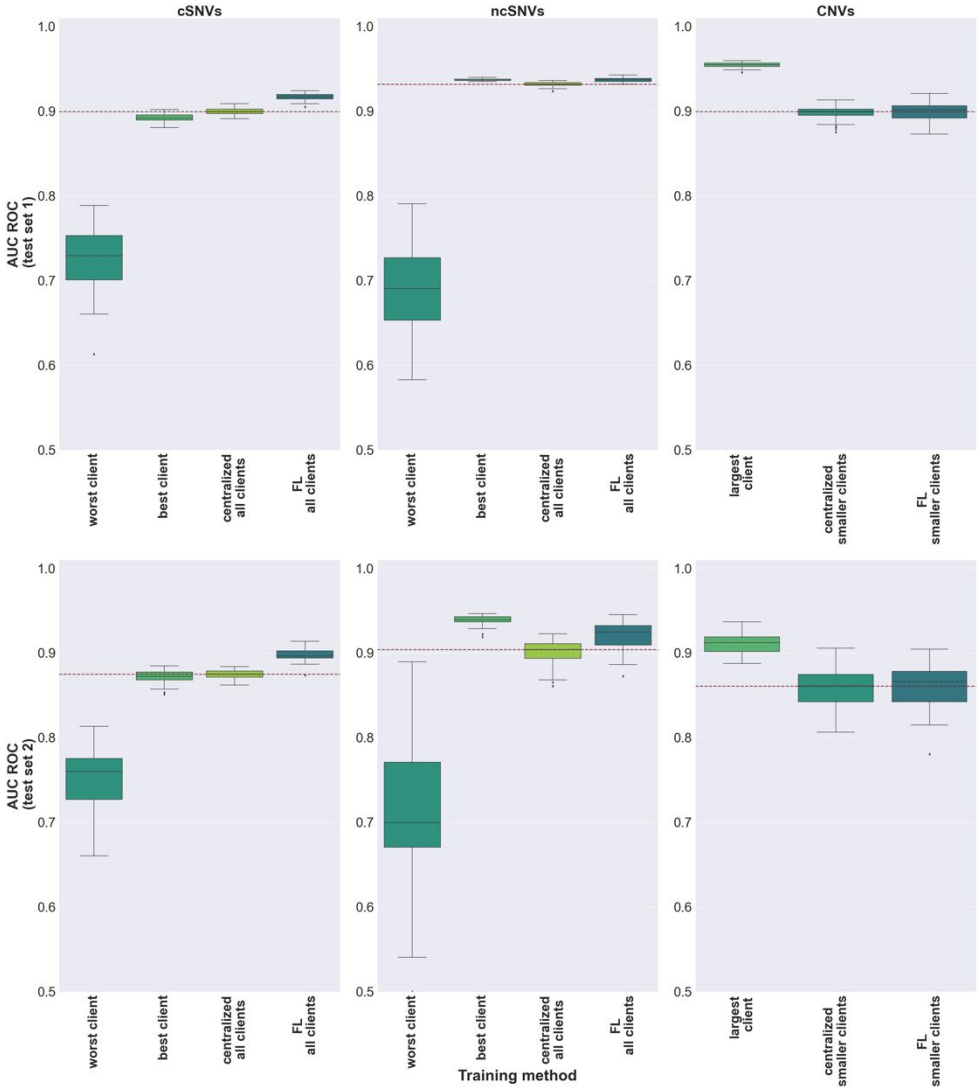
3.3 Comparative performance evaluation among local-client, centralized, and federated models on two independent test sets.

The assessment of the federated learning models through cross-validation on the training set allowed us to identify the optimal FL parameters for each type of genetic variant and learning method considered (MLP and sNDF), in terms of aggregation strategy, client rate, and batch normalization in the case of MLP (**Methods** and **Supplementary Table 5**). We then compared the performance of such optimal FL models against models obtained either from individual-client models or from centralized model counterparts, i.e., trained on aggregated client data. The AUC ROC values obtained on the two independent test sets for the three types of genetic variants considered are represented for the MLP and sNDF models, respectively, in **Figures 3** and **Supplementary Figure 3,** with complete details reported in **Supplementary Tables 7–9**.

In the case of coding and non-coding SNVs, the results showed that the federated learning models were consistently either comparable or superior to the centralized model counterparts. FL models were particularly

remarkable in the case of MLP (**Figure 3**), where they significantly outperformed the centralized models in the two independent sets evaluated both for coding and non-coding SNVs (Wilcoxon rank sum test p-value <1.5e-05, **Supplementary Table 10**). Individual client models however showed a large heterogeneity, illustrating the risks of non-collaborative settings. Thus, while some reached competitive performance, others led to significantly inferior results (**Supplementary Tables 7–8**).

Fig. 3. Performance of local, centralized, and federated MLP models on two independent test sets. Top and bottom panels show, respectively, the results on the first and second independent sets (Methods). In the case of coding SNVs and non-coding SNVs, the performance of the worst and best individual client models as well as the pooled-clients models are displayed. The performances of the remaining client models are reported in **Supplementary Tables 7–9**. In the case of CNVs, the performance of the largest client was compared against the centralized and federated learning models of the smaller clients. Boxplots in the panels represent the distribution of AUC ROC values obtained upon 30 different random seeds for model



weight initialization. To ease comparison across models, a dotted red line represents the median value obtained for the centralized models.

As previously introduced, in the case of CNVs the performance of the largest client model was compared against the centralized and FL models across the smaller clients. Here, the largest-client ranked first across all scenarios evaluated, as expected from the sharp difference in the training set size (**Figure 2**). However, despite this limitation, the cooperation across small clients through FL led to competitive results which, in the case of MLP, were not significantly different from their centralized versions (**Figure 4**). Importantly, the federated learning models based on differentiable algorithms explored in this study (MLP and sNDF) achieved results comparable to the state-of-the-art performance of non-differentiable models such as Random Forest and XGBoost, which were trained on centralized data and for which federated learning implementations have not yet been developed (**Supplementary Figure 4** and **Supplementary Table 11**). Considered together, these results showed the beneficial impact of collaboratively training supervised models across different scenarios for the classification of human genetic variants while respecting data privacy constraints.

3.4 Measuring the robustness of centralized and federated models against client dropouts

We then evaluated to what extent the performance of centralized and federated models would have been affected in the eventual case that a given client had not participated in the collaborative process. **Figures 4** and **Supplementary Figure 5** allow comparison of such events on the two learning algorithms evaluated, MLP and sNDF, respectively, for the two independent sets evaluated. In the case of coding and non-coding SNVs, the results showed that the federated learning models were consistently more resilient to client dropouts than the centralized model counterparts (**Supplementary Tables 12 – 13**). FL models were particularly robust in the case of MLP, where they were not significantly affected for most of the client drops, and in sharp contrast to the performance drops observed in the centralized settings (**Figure 4** and **Supplementary Figure 5**). In the case of CNVs, less consistent trends were observed upon client dropouts, reflecting stochastic sampling effects as a result of the small sample size on both the centralized and federated settings. Thus, while MLP models were generally robust to client dropouts in both the centralized and FL settings, sNDF showed less resilient to the absence of clients. Considered together, our results suggest that federated learning models need comparatively smaller training datasets than their centralized model counterparts in order to generalize adequately to independent datasets.

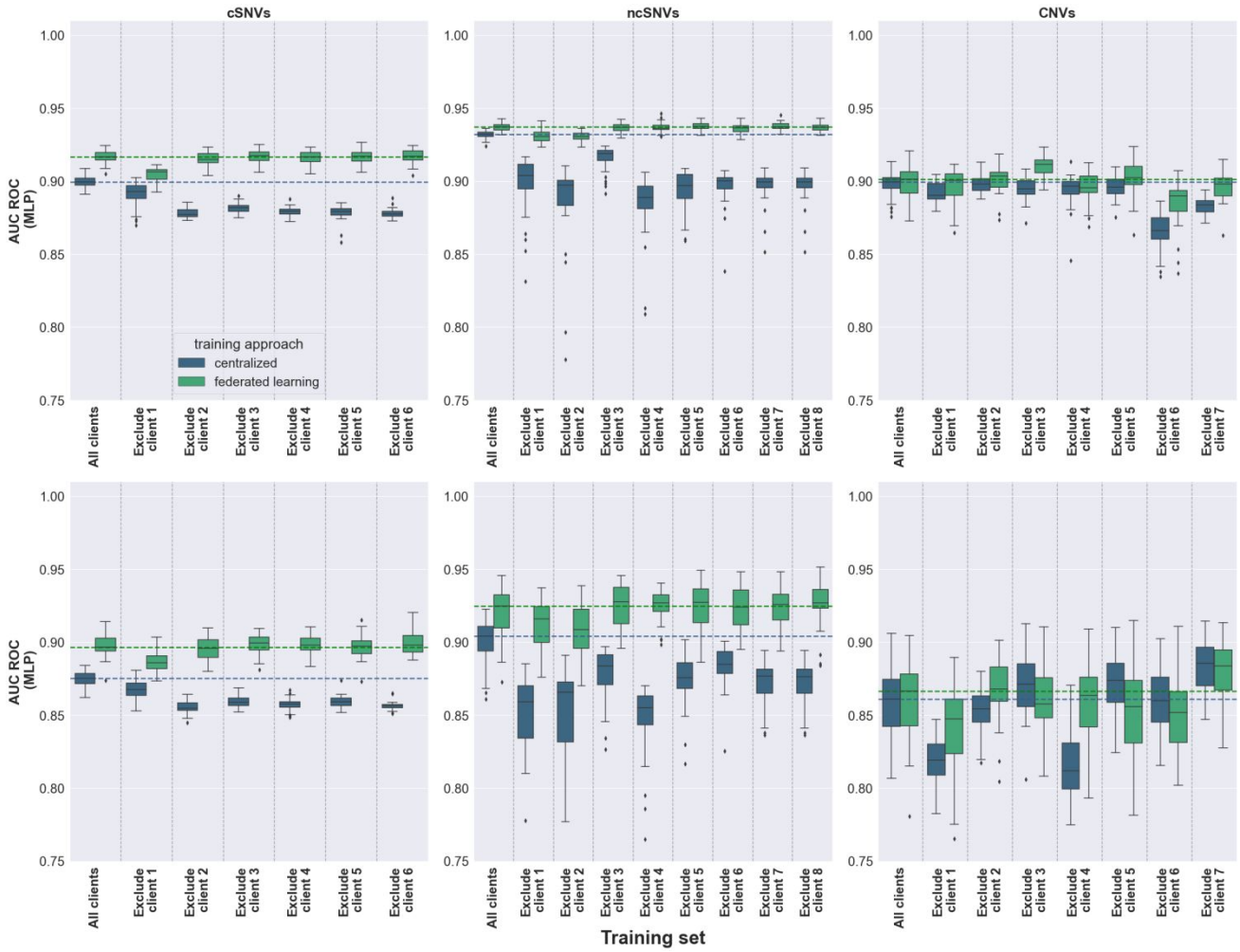


Fig. 4. Performance of centralized and federated MLP models upon client dropouts on two independent test sets. Top and bottom panels show, respectively, the results obtained on the first and second independent test (Methods). Boxplots in the panels represent the distribution of AUC ROC values obtained upon 30 different random seeds for model weight initialization. Centralized and FL models are colored in blue and green, respectively. Each panel represents the AUC ROC values obtained considering all clients, as well as excluding one client at a time. To ease comparison across models, dotted lines represent the median values obtained for the centralized (blue) and federated (green) models considering all clients.

3.5 Assessing the similarity of centralized and federated models

In the previous sections we showed that centralized and federated models led to overall comparable performances across diverse settings and scenarios, and this in spite of their different training process. Here we further inquired whether such results reflected a convergence of both

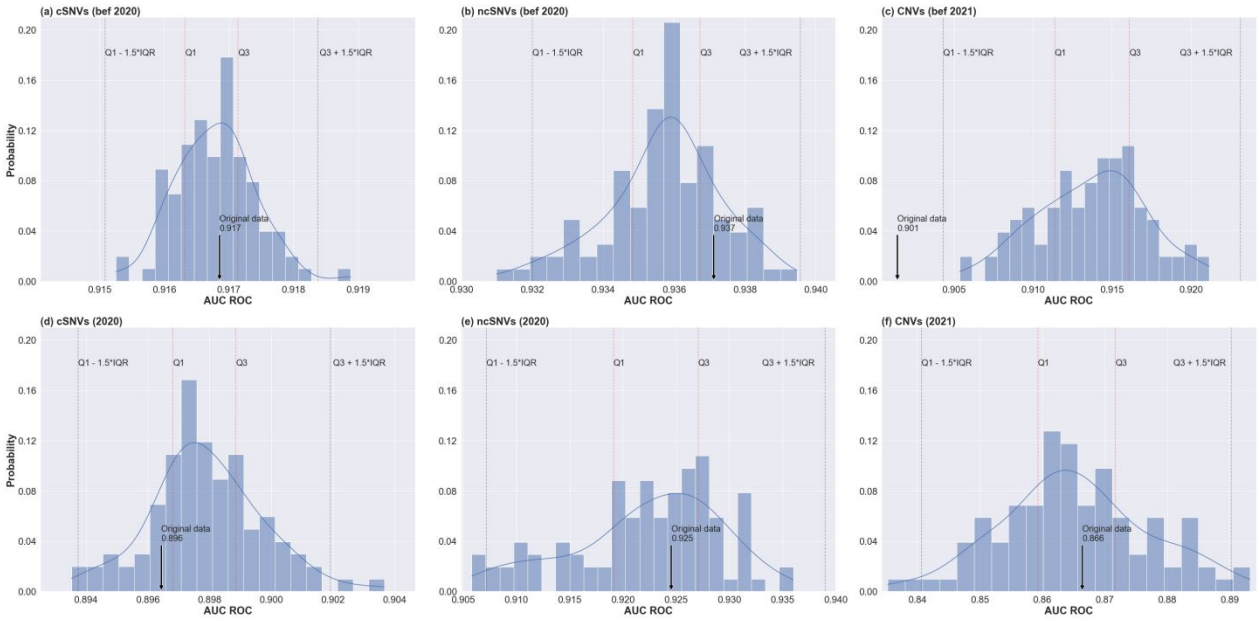
approaches into largely redundant models, not only in terms of global accuracy, but also in terms of their classification outcomes for each specific genetic variant evaluated. To address this question, we evaluated the Spearman rank correlation between the pathogenicity scores for each genetic variant produced by the centralized and the federated model, for each evaluated setting (**Supplementary Table 14**). The observed rank correlation values ranged between 0.92 and 0.97, with p-values < 2.16e-132 for comparisons involving coding and non-coding SNVs, suggesting that models learned in a centralized and federated learning manner were largely equivalent for individual variant assessment. In the case of CNVs, the similarities between the two approaches were also statistically significant across all settings (p-values < 1.36e-25) yet with lower rank correlation values, ranging between 0.82 and 0.87. Such values suggest a larger divergence between the centralized and federated approaches as compared to the SNV scenarios, probably reflecting stochastic sampling variance across clients in the federated learning process due to the smaller clients' sample size.

3.6 Analyzing the influence of genetic variant feature distribution on federated learning performance

Motivated by the previous results, we further investigated whether the comparable or superior performances observed in FL as compared to centralized models were driven by underlying identical or non-identical distributions of genetic variant features across the clients participating in the training process. It has been hypothesized that FL models may generalize better than their centralized counterparts in the presence of non-identical feature distributions across clients, due to an implicit regularization effect introduced by model averaging, similar to the benefits achieved through dropout (McMahan *et al.*, 2017). To explore such possibility, we first inspected the distribution of variants in the training set across an unsupervised low-dimensional representation obtained from the variants feature matrices (**Supplementary Figure 6**, **Methods**). In the case of coding and non-coding SNVs, variants were homogeneously distributed across the UMAP space both in terms of their clinical label (pathogenic or benign) and of their institutional submitter. In the case of CNVs, however, pathogenic variants formed distinctive clusters, which generally reflected their client of origin (**Supplementary Figure 6**). Those observations suggest that CNV variants, but not coding and non-coding SNVs, might have a non-identical feature distribution across participant Institutions involved in the FL experiments.

To further investigate this possibility, we then generated a pool of 100 synthetic random partitions of the initial training set across the same number of virtual clients, while respecting their original sample size, and trained FL models using the optimal hyperparameters on the generated partitions (**Methods**). For execution time limitations, these experiments were conducted exclusively on the MLP model. **Figure 5** shows the histograms representing the distribution of median AUC ROC values achieved by the FL models trained in the 100 newly generated training sets, and evaluated on the two independent test sets. In the case of coding and non-coding SNVs, the results showed that the actual median AUC ROC values obtained by the original model on the two independent test sets were not significantly different than those obtained by models where variants were randomly distributed across clients (z-scores ranging from -1.05 to 0.80, with normal p-values > 0.05). These observations reflect that both coding and non-coding Single Nucleotide variants may not exhibit strong non-identical feature distributions. Conversely, for CNVs, the original AUC ROC value was significantly lower than the random distribution in the case of the first held-out test set (z-score = −3.74, normal p-value = 0.74), while no significant difference was observed for the second test set. We hypothesize that, in this case, the limited sample size of the second test set (96 samples) might have reduced the statistical power to detect differences in performance. Nonetheless, the results observed on CNVs are in line with the observations from the low-dimensional representations, both reflecting that the actual CNV data partition across clients presented non-identically distributed features, probably accentuated by the small clients' data size.

The previous results suggest that the advantages observed for federated models throughout this work in terms of superior and more robust performance, would be driven by the iterative learning process across disjoint and partly represented datasets (with client rates < 100%) rather than by non-identical feature distribution across clients. As a corollary of this result, it may be proposed that, even when a large aggregated training dataset is available, mimicking a virtual federated learning through



random splits across synthetic clients could lead to performance improvements.

Fig. 5. Distribution of median AUC ROC values obtained by the FL models trained on 100 randomly generated training sets. Top and bottom panels illustrate the histogram of the FL models evaluated, respectively, on the first and second independent test set for coding SNVs (left), non-coding SNVs (middle), and CNVs (right). For each histogram we delimit: the interquartile range (IQR = Q3 − Q1), representing the range within which the middle 50% of the data lies; the lower bound (Q1 − 1.5 × IQR) and upper bound (Q3 + 1.5 × IQR), for the identification of outliers; and the median area under the ROC curve (referred to as the original data) achieved by the FL model trained on the original data split.

4 Discussion

In this work we carried out a proof-of-concept study evaluating whether Federated Learning (FL) might be an effective collaborative machine learning strategy for the pathogenicity annotation of human genetic variants in the context of clinical genomics for rare diseases. To this end, we leveraged the publicly available database ClinVar to mimic realistic multi-institutional collaborations for the training of supervised ML models, with and without data sharing. Our experiments showed that, in most cases, FL achieved competitive or superior performance compared to collaborative data sharing (CDS) approaches, while also outperforming single-institutional models for the majority of participants. We also demonstrated that FL generally exhibited greater robustness to the removal of a participant's data from the training set compared to CDS approaches. Such results suggest that FL needs relatively smaller datasets than CDS approaches to generalize adequately to unseen datasets. Our study thus supports FL as a beneficial approach for training supervised ML models for the clinical classification of genetic variants across multiple institutions, while respecting data privacy constraints.

Notably, when using Multilayer Perceptron as a learning algorithm, FL outperformed CDS approaches in the clinical assessment of coding as well as of non-coding SNVs, despite the distributions of genetic variant features across the clients participating in the training process being identical. Such a result may initially seem counterintuitive, considering that a CDS approach benefits from complete data access and centralized optimization. However, our findings align with previous studies observing a similar trend in biomedical applications. For example, Linardos et al. (Linardos *et al.*, 2022) observed that a FL model for cardiac magnetic resonance imaging (MRI) diagnosis outperformed its centralized counterpart. The authors hypothesized that this behavior could be attributed to the model averaging process in each round of FL, which may have had a stabilizing effect, leading to improved performance across different model initializations. Ogier du Terrail et al. (Ogier du Terrail *et al.*, 2022) presented a FL benchmark using seven multi-institutional datasets, simulating diverse realistic healthcare settings covering different tasks and modalities. Authors found that FL outperformed its centralized model counterpart when the ML models were linear and when the training set was low-dimensional, tabular data. Based on these studies, we speculate that FL outperformed the CDS approach in our case due to model averaging and to the characteristics of the training set, which consisted of low-dimensional tabular data. Interestingly, the performance gap between FL and CDS was more pronounced for MLP, a simpler model with fewer parameters, as compared to sNDF. This suggests that federated optimization may become more challenging as model complexity and number of model parameters increases.

The results obtained through cross-validation (**Figure 3**) revealed additional interesting findings. First, the introduction of batch normalization layers to local MLP models had a negative impact, damaging the performance of all FL aggregation strategies evaluated. While some studies suggested that the use of local batch normalization layers can enhance the performance of FL under conditions of data heterogeneity (Li *et al.*, 2021; Andreux *et al.*, 2020), other studies point to potential drawbacks, especially when there is a great mismatch between local and global model statistics (Wang *et al.*, 2024). It is worth noting that these studies focused, however, on computer vision applications often involving deep learning architectures. In the context of our study, our results suggest that incorporating batch normalization layers into shallow models, such as a 3-layer MLP, can be counterproductive. Second, using only 50% of clients for local training, randomly selected in each round, generally resulted in better performance compared to using all clients (**Supplementary Figure 1**). This observation suggests that, by training on a different subset of clients in each round, the FL model may avoid overfitting to specific clients' data, improving its generalization capabilities. Finally, a corollary of our work is that, even in scenarios with full data access to large centralized genetic variant collections, mimicking FL training across virtual clients with randomly generated data splits can lead to model improvements over their centralized counterparts.

Supervised learning approaches for the clinical assessment of genetic variants have extensively relied on centralized curated repositories such as ClinVar (Landrum *et al.*, 2014) and Decipher (Firth *et al.*, 2009), among others. While these are valuable resources for the community, they necessarily exclude sensitive information from the carrier individuals, including complete genomic data, clinical history and relevant phenotypic information. Thus, as our proof-of-concept study is based on such type of centralized repositories, we could only evaluate the beneficial aspects of FL models for variant classification that rely exclusively on genomic features of the variants, without considering patients' clinical signs. Yet, our study serves as an incentive for the implementation of a FL approach in real-word settings that could benefit from the incorporation of additional data modalities also requiring data privacy, such as electronic health records and medical imagine. However, since our study is based on the ClinVar database, it primarily focuses on germline variants and Mendelian diseases, and further research is needed to assess whether our findings extend to more complex disease contexts. An additional limitation of our study is the lack of explicit control for population genetic diversity, as ethnicity information is not systematically reported in ClinVar. Thus, future implementations of federated learning in clinical genomics should account for ancestry-related information to ensure comparable model performance across diverse populations.

In our experiments we assumed that clients and server are honest and trusted, which means that no party is attempting to disrupt or manipulate the FL training process. However, a real implementation of FL would need to take into consideration further privacy-preserving guarantees against potential malicious attacks, such as attempts to reconstruct the original data from model updates (Geiping *et al.*, 2020). Homomorphic encryption (Acar *et al.*, 2018) and Differential Privacy (Ziller *et al.*, 2024) can be used to secure parameter exchange in FL. Homomorphic encryption allows computation to be performed on encrypted data, allowing the clients to encrypt the model updates before sending them to the server. The central server then aggregates the encrypted local updates, returning an encrypted global model to the clients for decryption. Although Homomorphic encryption offers strong privacy protections, current implementations are computationally expensive, slowing down the FL training process. Differential Privacy, on the other hand, adds noise to local updates before sending them to the server for model aggregation. However, it is crucial to carefully choose the amount and the shape of noise to be added, since

excessive noise can degrade model performance, while insufficient noise may not provide sufficient privacy protection.

Our work extends to genetic variant assessment the scope of biomedical applications for which Federated Learning has proven to be an effective strategy for implementing collaborative machine learning approaches under data privacy constraints. These findings represent a major novelty in the field of clinical genomics and we expect them to encourage the adoption of FL to establish secure multi-institutional collaborations for human variant interpretation. This, in turn, would lead to more robust ML models that benefit not only from larger datasets but, most importantly, from more diverse datasets covering a wider range of genetic conditions, genetic backgrounds, and clinical manifestations.

Availability

All code required to reproduce the results presented in this manuscript is available under the GNU General Public License v3. The code is accessible via a *GitHub* repository at https://github.com/RausellLab/FedLearnVar and a snapshot of the code is archived with doi: https://doi.org/10.5281/zenodo.16029049

Acknowledgements

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

Funding

The Laboratory of Clinical Bioinformatics of the Imagine Institute, headed by A.R. was partly supported by the French National Research Agency (ANR) 'Investissements d'Avenir' Program [ANR-10-IAHU-01 and ANR-21-PMRB-0004, FACE.S-4-KIDS project, "FACE and SKULL for Key Innovative Data Science]; by the European Rare Diseases Alliance (ERDERA) programme funded by the European Union's Horizon Europe research and innovation programme under grant agreement N°101156595; and by the French government as part of the "Important Project of Common European Interest" (IPCEI) Cloud call of the France 2030 programme (E2CC - AI4RDP - AI for Rare Diseases Pathogenicity project). N.M. was partly supported by the French National Research Agency (ANR) 'Investissements d'Avenir' Program [ANR-10-IAHU-01], the JANSSEN HORIZON Fonds de dotation, and by the French government as part of the "Important Project of Common European Interest" (IPCEI) Cloud call of the France 2030 programme (E2CC - AI4RDP - AI for Rare Diseases Pathogenicity project). EC was supported by the Italian Ministry of Health - PNRR-MR1-2022-12376067 Multiomic strategies to implement the diagnostic workflow of rare diseases.

Conflict of Interest: none declared.

Author's contribution

Conceived and designed the experiments: NM, AR. Performed the experiments: NM. Analyzed the data: NM, AR. Contributed materials/analysis tools: NM, FR, EC, AR. Wrote the paper: NM, EC, AR.

References

Acar,A. *et al.* (2018) A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Comput. Surv.*, **51**. https://doi.org/10.1145/3214303

Adnan,M. *et al.* (2022) Federated learning and differential privacy for medical image analysis. *Sci Rep*, **12**, 1953.

Andreux,M. *et al.* (2020) Siloed Federated Learning for Multi-Centric Histopathology Datasets. arXiv. https://doi.org/10.48550/arXiv.2008.07424

Annas,G.J. (2003) HIPAA regulations - a new era of medical-record privacy? *N Engl J Med*, **348**, 1486–1490.

Austin-Tse,C.A. *et al.* (2022) Best practices for the interpretation and reporting of clinical whole genome sequencing. *npj Genom. Med.*, **7**, 27.

Beutel,D.J. *et al.* (2020) Flower: A Friendly Federated Learning Research Framework.

Bonomi,L. *et al.* (2020) Privacy challenges and research opportunities for genomic data sharing. *Nat Genet*, **52**, 646–654.

Boukas,L. *et al.* (2020) Promoter CpG Density Predicts Downstream Gene Loss-of-Function Intolerance. *The American Journal of Human Genetics*, **107**, 487–498.

Breiman,L. (2001). *Machine Learning*, **45**, 5–32.

Bromberg,Y. *et al.* (2024) Variant Effect Prediction in the Age of Machine Learning. *Cold Spring Harb Perspect Biol*, **16**, a041467.

Brookes,A.J. and Robinson,P.N. (2015) Human genotype–phenotype databases: aims, challenges and opportunities. *Nat Rev Genet*, **16**, 702–715.

Capra,J.A. *et al.* (2013) Many human accelerated regions are developmental enhancers. *Phil. Trans. R. Soc. B*, **368**, 20130025.

Capriotti,E. and Fariselli,P. (2023) PhD-SNPg: updating a webserver and lightweight tool for scoring nucleotide variants. *Nucleic Acids Res*, **51**, W451–W458.

Caron,B. *et al.* (2019) NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol*, **20**, 32.

Collins,R.L. *et al.* (2022) A cross-disorder dosage sensitivity map of the human genome. *Cell*, **185**, 3041-3055.e25.

Davydov,E.V. *et al.* (2010) Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol*, **6**, e1001025.

Dimitrieva,S. and Bucher,P. (2013) UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res*, **41**, D101-109.

Drew,K. *et al.* (2021) hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol Syst Biol*, **17**, e10016.

Ebert,P. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372**, eabf7117.

Eilbeck,K. *et al.* (2017) Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet*, **18**, 599–612.

Ellingford,J.M. *et al.* (2022) Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med*, **14**, 73.

Firth,H.V. *et al.* (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*, **84**, 524–533.

Geiping,J. *et al.* (2020) Inverting gradients - how easy is it to break privacy in federated learning? In, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc., Red Hook, NY, USA.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580–585.

Guelen,L. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.

Haeussler,M. *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*, **47**, D853–D858.

Halldorsson,B.V. *et al.* (2019) Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, **363**, eaau1043.

Havrilla,J.M. *et al.* (2019) A map of constrained coding regions in the human genome. *Nat Genet*, **51**, 88–95.

Hoofnagle,C.J. *et al.* (2019) The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, **28**, 65–98.

Karczewski,K.J. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.

Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization. arXiv;. https://doi.org/10.48550/arXiv.1412.6980

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, advance online publication.

Kolobkov,D. *et al.* (2024) Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project. *Front Big Data*, **7**, 1266031.

Kontschieder,P. *et al.* (2015) Deep Neural Decision Forests. In, *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, pp. 1467–1475. https://doi.org/10.1109/ICCV.2015.172

Landrum,M.J. *et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*, **42**, D980-985.

Li,T. *et al.* (2018) Federated Optimization in Heterogeneous Networks. arXiv. https://doi.org/10.48550/arXiv.1812.06127

Li,X. *et al.* (2021) FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. ICLR 2021.

Licata,L. *et al.* (2023) Resources and tools for rare disease variant interpretation. *Front. Mol. Biosci.*, **10**, 1169109.

Linardos,A. *et al.* (2022) Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Sci Rep*, **12**, 3551.

Lydia, A. A. and Francis, F. S. Adagrad—an optimizer for stochastic gradient descent. *International Journal of Information and Computing Science*, **6**, 566–568.

McMahan,B. *et al.* (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In, Singh,A. and Zhu,J. (eds), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, pp. 1273–1282.

Nasirigerdeh,R. *et al.* (2022) sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. *Genome Biol*, **23**, 32.

Nguengang Wakap,S. *et al.* (2020) Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*, **28**, 165–173.

Ogier du Terrail,J. *et al.* (2022) FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. In, Koyejo,S. *et al.* (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 5315–5334.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110–121.

Popescu,M.-C. *et al.* (2009) Multilayer perceptron and neural networks. *WSEAS Trans. Cir. and Sys.*, **8**, 579–588.

Raimondi,D. *et al.* (2023) Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohn's disease patients. *Sci Rep*, **13**, 19449.

Reddi,S. *et al.* (2020) Adaptive Federated Optimization. arXiv. https://doi.org/10.48550/arXiv.2003.00295

Requena,F. *et al.* (2022) CNVscore calculates pathogenicity scores for copy number variants together with uncertainty estimates accounting for learning biases in reference Mendelian disorder datasets. medRiv https://doi.org/10.1101/2022.06.23.22276396

Sadilek,A. *et al.* (2021) Privacy-first health research with federated learning. *NPJ Digit Med*, **4**, 132.

Sharo,A.G. *et al.* (2022) StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants. *The American Journal of Human Genetics*, **109**, 195–209.

Sheller,M.J. *et al.* (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*, **10**, 12598.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034–1050.

Singh,P.P. and Isambert,H. (2020) OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res*, **48**, D724–D730.

Stranneheim,H. *et al.* (2021) Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*, **13**, 40.

Turro,E. *et al.* (2020) Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, **583**, 96–102.

Wan,Z. *et al.* (2022) Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet*, **23**, 429–445.

Wang,T. *et al.* (2014) Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, **343**, 80–84.

Wang,X. and Goldstein,D.B. (2020) Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *The American Journal of Human Genetics*, **106**, 215–233.

Wang,Y. *et al.* (2024) Why Batch Normalization Damage Federated Learning on Non-IID Data? *IEEE Trans. Neural Netw. Learning Syst.*, 1–15.

Zaheer,M. *et al.* (2018) Adaptive methods for nonconvex optimization. In, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18. Curran Associates Inc., Red Hook, NY, USA, pp. 9815–9825.

Zerbino,D.R. *et al.* (2015) The ensembl regulatory build. *Genome Biol*, **16**, 56.

Zhu,C. *et al.* (2020) Computational Approaches for Unraveling the Effects of Variation in the Human Genome and Microbiome. *Annu. Rev. Biomed. Data Sci.*, **3**, 411–432.

Ziller,A. *et al.* (2024) Reconciling privacy and accuracy in AI for medical imaging. *Nat Mach Intell*, **6**, 764–774.

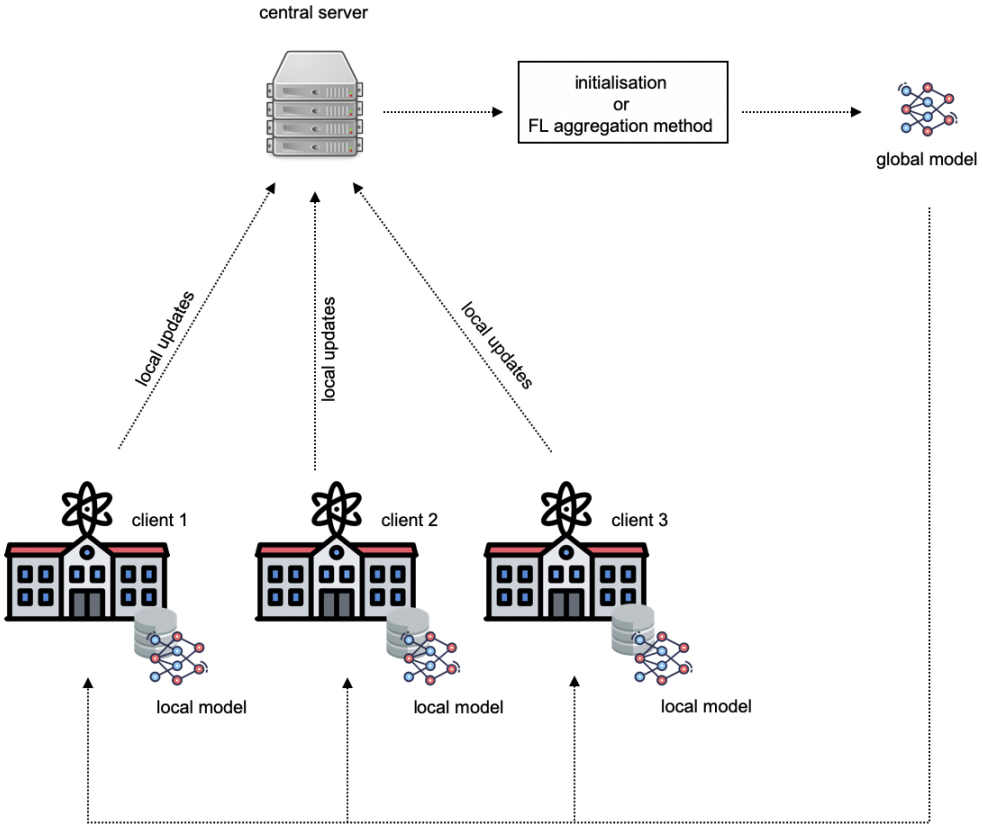


Figure 1

93x80mm (300 x 300 DPI)

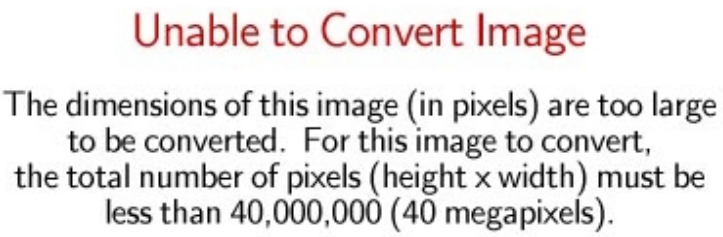


Figure 2

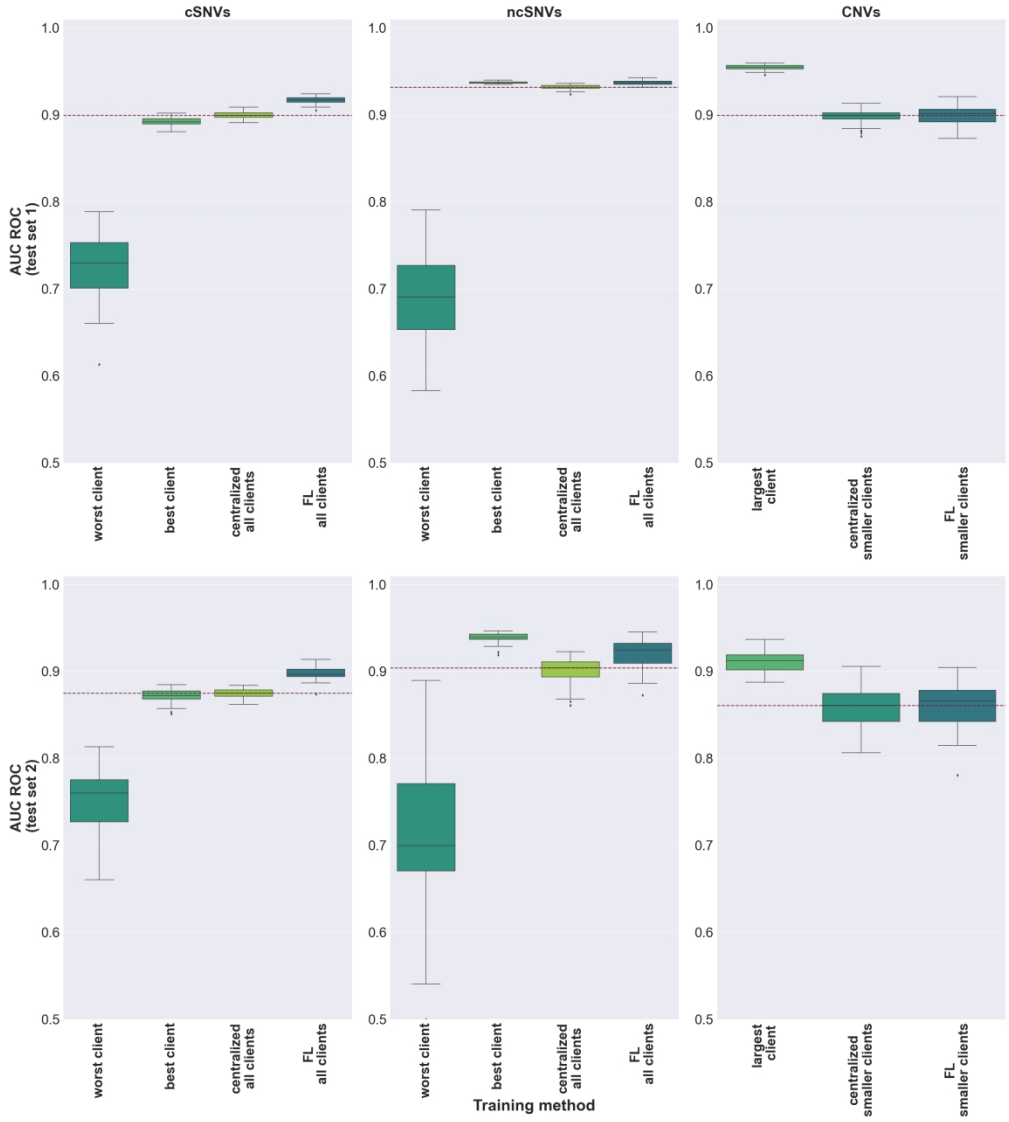


Figure 3

241x269mm (300 x 300 DPI)

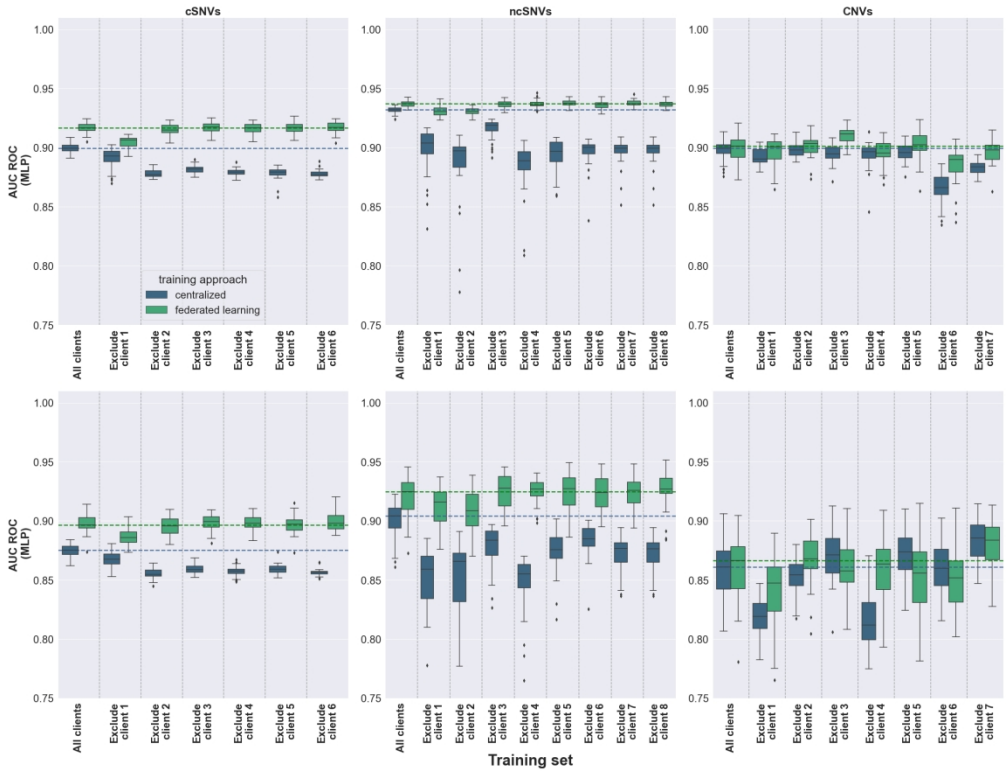


Figure 4

195x149mm (300 x 300 DPI)

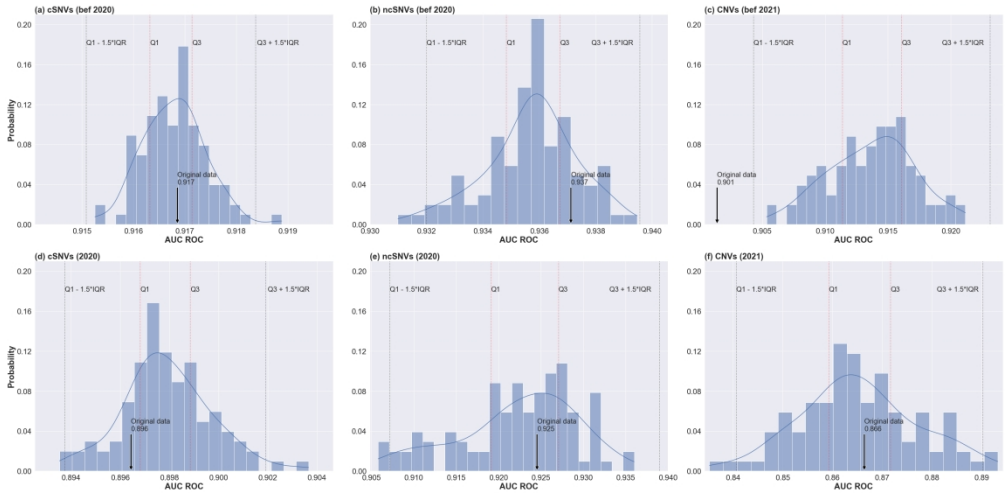


Figure 5

244x119mm (300 x 300 DPI)