Federated Learning towards the unknown: a deep dive into diabetic retinopathy prediction from real-world EHR structured data on unseen diabetic centers

Alessandro Cacciatore[1] (✉), Mariachiara Di Cosmo[2], Emanuele Frontoni[3], and Michele Bernardini[4]

[1] Department of Humanities, University of Macerata
a.cacciatore1@unimc.it
[2] The BioRobotics Institute, Sant'Anna School of Advanced Studies
mariachiara.dicosmo@santannapisa.it
[3] Department of Political Sciences, Communication and International Relations, University of Macerata
emanuele.frontoni@unimc.it
[4] Department of Theoretical and Applied Sciences, eCampus University
michele.bernardini@uniecampus.it

Abstract. The compatibility of Federated Learning (FL) models with unseen Out-Of-Federation (OOF) centers remains a critical yet underexplored challenge, particularly when dealing with heterogeneous data. To address this gap, this study proposes a data-driven approach to assess the feasibility of applying an FL model to OOF centers. The case study explored is the prediction of diabetic retinopathy from multiple real-world, highly heterogeneous electronic health records. An FL XGBoost model (FL-XGB) is trained across five in-federation (IF) centers, showing an average test Area Under the ROC Curve (AUC) of 75.27%. A novel metric, the OOF Applicability (OFA) predictor, is introduced to estimate whether FL-XGB could be safely applied to the 15 OOF centers. OFA combines statistical and learnable features from both IF and OOF centers and is used as a predictor for a regression model, employed to estimate the performance of FL-XGB (in terms of AUC) on OOF datasets. The regression model achieved a confidence of 76% in predicting AUC values, with a statistically significant p-value ($\ll 0.001$). The average discrepancy between the predicted and observed AUC values was 6%. Overall, FL-XGB shows robust performance on IF centers and the OFA predictor plays a crucial role in assessing its applicability to infer on unseen OOF centers. By providing statistically significant estimations, OFA effectively identifies OOF centers whose characteristics are too divergent from what the FL model can effectively manage. Our codes are available at https://github.com/geronimaw/OFA4FL.

Keywords: Federated Learning · Out-of-Federation · Electronic Health Records · Predictive Modeling · Diabetic Retinopathy

1 Introduction

Diabetic Retinopathy (DR) is a leading complication of diabetes and a major cause of vision impairment and blindness worldwide. It currently affects approximately 103 million diabetic patients, a number expected to rise to 161 million by 2045 [1]. Despite its severity, DR often goes undetected in its early stages, resulting in reactive rather than preventive treatments. Early detection is essential, as timely intervention can significantly lower the risk of severe vision loss and enhance the quality of life for those affected.

Machine Learning (ML) and Deep Learning (DL) techniques have shown great promise in DR diagnosis, particularly when applied to retinal fundus images and optical coherence tomography scans [2]. While these imaging modalities are widely adopted to identify disease markers and grade disease severity [3,4], Electronic Health Records (EHRs) remain underutilized in DR research, despite providing continuous and comprehensive insight into a patient's health journey. EHRs contain invaluable longitudinal data, such as demographics, clinical history, comorbidities, and routine laboratory results, which makes them suitable for early detection of DR risk and monitoring its progression [5]. Several studies have investigated ML models for predicting DR onset in diabetic patients using EHR data, and eXtreme Gradient Boosting (XGBoost) has consistently outperformed other classical ML models, such as logistic regression, support vector machines, artificial neural networks and random forests, in diabetic DR risk prediction [6,7,8,9,10]. Despite the promising results, these studies rely on the aggregation of data from multiple centers into a single repository, an approach that simplifies model training and is impractical in real-life scenarios due to privacy concerns and regulatory restrictions surrounding patient data. Moreover, centralizing data leads to less generalizable models, as they fail to capture the variability of data collected across diverse healthcare institutions, ultimately limiting their applicability.

Federated Learning (FL) [11] is a decentralized approach that offers the possibility to tackle these challenges by enabling multiple centers to collaboratively train a shared predictive model without exchanging data, thereby preserving patient privacy and adhering to data governance policies [12,13]. Despite these advantages, FL models face significant hurdles in real-world applications, particularly when dealing with Out-of-Federation (OOF) centers, whose data characteristics differ from those of the In-Federation (IF) centers and often exhibit high variability and heterogeneity in real-world scenarios. Therefore, this study is guided by the following research question:

Is it possible to determine whether an FL model can be reliably applied to OOF centers? How can we assess its compatibility with unseen data?

To address this question, we propose the novel Out-of-Federation Applicability (OFA) predictor, which evaluates whether an FL XGBoost model (FL-XGB) can be effectively used on data from an OOF center. OFA is a data-driven predictor that leverages a combination of statistical features (e.g., class imbalance, missing values) and latent features extracted via unsupervised DL techniques to quantitatively assess the FL model reliability to infer on OOF centers. Figure

1 shows the proposed framework, where FL-XGB is trained on IF centers and OFA assesses its compatibility with OOF centers.

The main contributions of this work are as follows:

- Introduction of the OFA predictor, a novel methodology integrating statistical and learnable features to determine the applicability of FL models to unseen centers.
- Introduction of the Out-of-Federation Suitability Score (OSS), a quantitative metric derived from the OFA outcomes, providing an interpretable measure of FL model applicability to OOF centers.
- Development of an FL framework based on XGBoost, called FL-XGB, designed and trained to predict DR risk from real-world EHR data collected from multiple diabetic centers.

To the best of our knowledge, this study is the first to introduce an approach to evaluate the compatibility of FL models with OOF centers, providing a data-driven framework to address privacy concerns, regulatory restrictions, and generalizability challenges in DR risk prediction from EHR data.

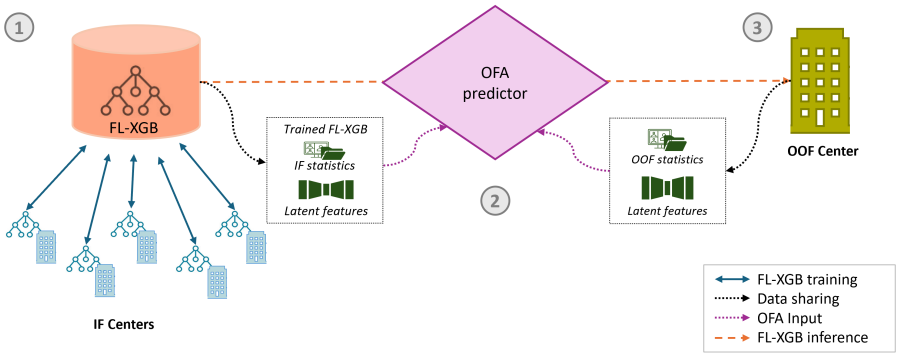


Fig. 1: Overview of the proposed framework for Diabetic Retinopathy (DR) risk prediction using a global XGBoost model (FL-XGB). (1) The process begins with the FL-XGB model being trained across five In-Federation (IF) centers, with blue arrows showing the flow of model updates. (2) The trained FL-XGB model and statistical features from IF and Out-Of-Federation (OOF) centers are fed to the novel Out-of-Federation Applicability (OFA) predictor to assess compatibility. (3) If the evaluation from the OFA predictor indicates sufficient compatibility, the FL-XGB model is applied to the OOF center for inference.

1.1 Related works

Although EHR data are essential for early-stage DR onset prediction and preventive care, their integration into predictive models remains largely unexplored.

While recent studies [6,7,8,9,10] demonstrated the potential of EHRs for DR risk prediction, they rely on centralized approaches, overlooking privacy concerns as well as the distributed nature of EHRs. The advent of FL has created new opportunities for collaborative learning with EHRs across multiple healthcare institutions, enabling predictive modeling for various clinical tasks while preserving patient privacy [14,15]. However, to the best of our knowledge, FL for EHR-based DR prediction has been explored in only one study [16], which optimizes a logistic regression model in an FL fashion on 22 centers. While this work achieved a sensitivity of 72% on real-world EHRs, it relied on dataset undersampling to balance classes, a strategy that likely oversimplified the task.

Outside of the DR and diabetes research domain, most studies involving FL in clinical tasks artificially create homogeneity in the federation by splitting a single dataset to simulate multiple FL clients. This approach facilitates experimentation but fails to reflect the complexity of real-world EHR collections [17]. To better handle heterogeneous data within the federation, strategies such as client selection [17] or client similarity metrics [18,19] have been introduced. However, these strategies often require preventive feature space transformation, which may reduce the interpretability and applicability of the models to real-world EHR data. Similarly, in the broader field of FL, literature primarily addresses data heterogeneity inside the federation [20,21] or between federations [22]. However, the challenge of FL model generalizability to OOF centers has been recently tackled by studies such as [23,24,25], in which the FL model is explicitly trained to generalize well across both seen and unseen environments.

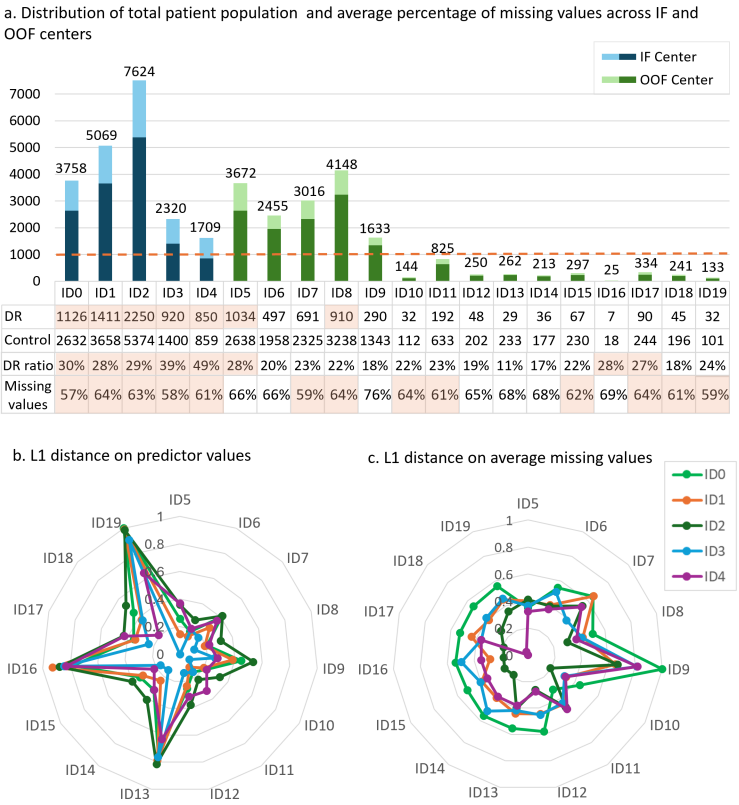
In light of these considerations, our work addresses a key gap in the field of FL in healthcare by effectively training an FL model on real-world EHR data despite their imbalance and heterogeneity across classes and centers. Furthermore, while previous studies focused on ways to train an FL model to increase its generalizability towards OOF data, our work investigates ways to predict whether a pre-trained FL model can be effectively applied to OOF data based on information such as class imbalance, dataset size, missing values, and latent features extracted in an unsupervised manner. This shifts the focus from generalization to compatibility estimation between a trained FL model and a new center.

2 Methods

2.1 Data and predictive task definition

The study leverages EHR datasets from 20 diabetic centers across Italy, containing patient records organized into three fields: i) demographic information (e.g., gender, age, diabetes duration), ii) pathology data (comorbidities), and iii) laboratory test results. Patients are classified into two groups — DR and control — based on specific observational temporal windows, following established preprocessing criteria [6,8]. The objective, framed as a binary classification task, is to predict the likelihood of DR development in diabetic patients. The classification is based on 62 predictors, including laboratory test results averaged

over the observational period, demographic information such as gender and age, and potential comorbidities. Missing data were handled using an extra-value imputation strategy.



	ID0	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10	ID11	ID12	ID13	ID14	ID15	ID16	ID17	ID18	ID19
DR	1126	1411	2250	920	850	1034	497	691	910	290	32	192	48	29	36	67	7	90	45	32
Control	2632	3658	5374	1400	859	2638	1958	2325	3238	1343	112	633	202	233	177	230	18	244	196	101
DR ratio	30%	28%	29%	39%	49%	28%	20%	23%	22%	18%	22%	23%	19%	11%	17%	22%	28%	27%	18%	24%
Missing values	57%	64%	63%	58%	61%	66%	66%	59%	64%	76%	64%	61%	65%	68%	68%	62%	69%	64%	61%	59%

Fig. 2: Statistical summary of the involved datasets. (a) Patient distribution across IF (blue) and OOF (green) centers, with total counts above the bars. The table details DR/control counts (also visually differentiated within the bars by hues, with lighter shades for DR), the DR ratio, and the percentage of missing data. IF selection criteria are marked: a 1000-patient threshold (dotted line), and DR counts $\geq$800, DR ratio $\geq$25%, and missing values $<$65% (red in the table). (b-c) Radar charts representing the L1 distances of each OOF center from the federation based on predictor values (b) and missing data percentages (c), with scales normalized from 0 to 1.

The diabetic centers are categorized into two groups, as shown in Fig. 2:

– *IF centers:* 5 centers (ID0-ID4) involved in the FL-XGB training and validation, selected based on data quality, volume, and diversity criteria: >1000

total patients, ≥800 DR patients, ≥25% DR patients, and <65% missing values.
- *OOF centers:* 15 centers (ID5-ID19), used for external validation, exhibiting high variability in patient numbers, predictor distributions, and missing data.

The inclusion criteria ensure that IF centers have sufficient data volume, diversity, and quality to support robust model training. As shown in Fig. 2a, the federation includes 20480 patients, while OOF centers collect data about 17648 patients. OOF centers exhibit a wide range of data characteristics, including extremely small numbers of patients and significantly high percentages of missing values. Preliminary analyses were carried out to assess the similarity between IF and OOF centers, focusing on the mean, median, 75th percentile, and range (minimum and maximum values) of predictors and percentages of missing values. In Fig. 2b-c, the normalized L1 distance between IF and OOF centers is depicted in terms of the 75th percentile of predictors and missing values, respectively. This variability reflects real-world heterogeneity, with notable differences from the federation in OOF centers such as ID13, ID16, and ID19, characterized by greater data sparsity and distribution discrepancies compared to IF centers.

2.2 Federated framework

To predict DR onset in diabetic patients, the FL framework uses XGBoost [26], a tree-based boosting algorithm known for its effectiveness in handling sparse, imbalanced and complex datasets, making it well-suited for EHR data [27]. This choice is further supported by multiple studies that demonstrate superior XGBoost performance over other ML models in EHR-based DR prediction task [6,7,8,9,10]. These strengths, combined with FL, allow FL-XGB to leverage heterogeneous datasets to train a generalized, privacy-preserving model applicable to diverse patient populations. Five IF centers collaboratively train the global model by sharing model updates (gradients) with a central server that aggregates them via Federated Averaging [11]. This ensures compliance with data governance policies and patient privacy regulations.

2.3 Out-of-Federation Applicability (OFA) Predictor

In a real-world scenario, a trained FL model can be made publicly available, enabling other centers to use it for testing purposes on the same task. However, factors such as data distribution, class imbalance, and missing values — which can vary significantly across real-world EHR datasets — can affect model performance on OOF data [28]. Taking these factors into account, we introduce the OFA predictor. Given a model trained on a federation of centers and an OOF center C_{OF}, OFA predicts the model's performance on C_{OF} in terms of the Area Under the ROC Curve (AUC). This prediction is based on a combination of properties of C_{OF} and metadata about the IF centers, grouped as follows:

- Statistical characteristics (S_c) of C_{OF}: Cardinality, class imbalance, and distributions of missing values and average predictors. The last two characteristics are compared with the corresponding distributions from each IF center, which are released as metadata.
- Prediction confidence (p_{pred}): The FL-XGB model's output on C_{OF}, serving as a measure of the alignment between C_{OF} and all IF centers.
- Latent representations (L_{AE}) of C_{OF}: The predictors in C_{OF} are projected into a low-dimensional latent space via an autoencoder and compared with the latent representations of each IF center, also released as metadata.

The metadata required from the federation can be shared while ensuring OFA's compliance with regulatory restrictions.

Formulation. OFA is based on the hypothesis that there is a (linear) correlation between the FL-XGB model's AUC performance on C_{OF} and a combination of the above-mentioned properties, X_{OFA}, defined as follows:

$$X_{OFA} = S_c \times L_{AE} \times p_{pred} = \left(\frac{\eta}{n} \times D_{L1,miss} \times D_{L1,feat}\right) \times D_{AE,feat} \times p_{pred} \tag{1}$$

where i) η is the imbalance ratio (negative/positive samples) in C_{OF}; ii) n is the cardinality of C_{OF}; iii) $D_{L1,feat}$ is the average L1 distance between C_{OF} and all the IF centers, calculated for the 75th percentile of their predictors (see Fig. 2b); iv) $D_{AE,feat}$ is same metric but calculated for centroids in a 6-dimensional latent space generated by an autoencoder (details in Sec. 2.4); v) $D_{L1,miss}$ measures the same distance in terms of missingness (see Fig. 2c); vi) p_{pred} is the average probability predicted by FL-XGB over all samples in C_{OF}. The variables composing X_{OFA} are selected for their relevance in capturing differences between the OOF center and the federation. The relationship between these variables and model performance was empirically validated through experimentation.

Finally, linear regression is used to estimate the relationship between X_{OFA} and AUC obtained from the FL-XGB model on C_{OF}:

$$AUC_{OFA} = m \times X_{OFA} + q \tag{2}$$

Where the parameters m and q need to be regressed from a set of OOF centers to fit the distribution at hand.

Scoring. To further refine decision-making, we define an OOF Suitability Score (OSS) based on the OFA-predicted AUC values (AUC_{OFA}). This score assesses whether applying the FL-XGB model to a specific OOF center C_{OF} is advisable, and it is computed by comparing AUC_{OFA} on C_{OF} against the average performance across IF centers ($\overline{AUC}_{IF}$), while also incorporating the predictive strength of OFA's linear regression, quantified by $r^2[\%]$. $OSS \in [0, 100]$ is defined by the formula:

$$OSS\% = r^2 \times \frac{100 - \left|AUC_{OFA} - \overline{AUC}_{IF}\right|}{100} \tag{3}$$

Higher values indicate greater confidence in the model's ability to generalize to the OOF center, and $OSS \geq 50\%$ indicates that the FL-XGB model can be safely applied to the OOF center. The average OSS, if computed on a significant number of OOF centers, is particularly meaningful for new centers that do not have annotations or the possibility to train their own model.

2.4 Implementation details

Table 1: Range of XGBoost hyperparameters used for training the DR prediction model. Regarding the scale for positive weight hyperparameter, η is defined as the number of negative samples over the number of positive samples. The objective function was also tuned among three different loss functions: Binary Cross Entropy (BCE), Weighted BCE (w-BCE) and Focal BCE (f-BCE). The hyperparameters α and γ marked with an asterisk (*) were only applied to f-BCE and w-BCE, respectively.

Hyperparameter	**Range**
Number of estimators	$\{3, 5, \mathbf{7}, 9, 12, 15, 25\}$
Maximum tree depth	$\{15, 25, \mathbf{50}, 75, 100\}$
Learning rate	$\{0.01, \mathbf{0.05}, 0.1, 1\}$
Subsample ratio	$\{0.3, \mathbf{0.5}, 0.7, 0.9, 1\}$
Column sample ratio	$\{0.3, 0.5, 0.7, 0.9, \mathbf{1}\}$
L2 regularization term (λ)	$\{1, 3, 5, 7, \mathbf{10}, 15\}$
Scale for positive class weight	$\eta \times \{0.5, \mathbf{1}, 3, 5, 10\}$
Imbalance parameter (α^*)	$\{1, \mathbf{2}, 3, 4, 5\}$
Focusing parameter (γ^*)	$\{1, 1.5, \mathbf{2}, 2.5, 3\}$
Objective function	{**BCE**, w-BCE, f-BCE}

FL-XGB. Following a thorough experimental evaluation, the optimal training configuration for DR prediction using FL-XGB is established after a 10-fold cross-validation and an extensive grid search to maximize $\overline{AUC}_{IF}$. A list of the optimal hyperparameters can be found in Table 1. Regardless of AUC value maximization, the column sample ratio, i.e., the number of randomly sampled columns among the 62 predictors, is set to 1. This choice is driven by the high inhomogeneity among centers and aimed to prevent weak learners in XGBoost from relying on unimportant predictors during training.

To address class imbalance, a weight for positive samples (*scale_pos_weight*) is applied, and Binary Cross Entropy (BCE) is selected as the objective function. Additionally, two other objective functions were tested in this study — namely, Weighted BCE (w-BCE) and Focal BCE (f-BCE), as they are designed to handle imbalanced datasets. The FL framework, implemented using the Flower platform

[29], involves training each model locally for five rounds before sending them to the central server for aggregation using Federated Averaging [11]. Feature importance is analyzed using the 'weight' strategy, emphasizing the frequency of predictors used to split nodes across the model's trees.

OFA. The optimal OFA predictor is determined by varying the variables involved in the computation of X_{OFA} and the relationship between them, as explained in the next Section. A critical aspect pertains to the design of the autoencoder architecture, which consists of two layers (16 and 32 neurons) in the encoder and specular layers for the decoder, with a latent space dimension of 6, determined from a grid search from the set [3, 6, 9, 12]. Hidden layers are activated by Leaky ReLU, while sigmoid is used for the output layer. The autoencoder is trained for 100 epochs using a batch size of 64, a learning rate of 0.001, and the Adam optimizer, with mean squared error as the loss function.

To evaluate the predictive relationship of OFA between X_{OFA} and AUC_{OFA}, all OOF centers are utilized except for ID11 and ID16, which are left out to test the model due to their unique characteristics in terms of sample size (25 for ID16) and missingness rate (76% in ID9), as shown in Fig. 2.

2.5 Ablation studies

To assess the robustness and generalizability of FL-XGB, we conducted numerous experiments. While performing a wide hyperparameters grid search as in Table 1, as Ablation Study 1 (AS1), we explored different loss functions (BCE, w-BCE, and f-BCE) to handle class imbalance and evaluated which of these loss functions was the most suitable for the task. Moreover, we evaluated different sets of IF centers in order to obtain the optimal federation in Ablation Study 2 (AS2). In particular, we tested federations with a varying number of centers (from 3 to 7 centers), including those that strictly adhere to the inclusion criteria (see Sec. 2.1) and those that marginally met these criteria (i.e., ID5 and ID8) to evaluate the impact of center diversity and data heterogeneity on model performance.

To evaluate the effectiveness of the OFA predictor, two ablation studies were carried out to provide insights into the obtained results and enhance the correlation between the designed X_{OFA} and the AUC predictions from the FL-XGB model. Ablation Study 3 (AS3) investigates whether non-learnable statistical features are more or less significant than deep representations extracted by an autoencoder in assessing the compatibility of a C_{OF} with the FL-XGB model. Specifically, we compared the OFA predictor with two alternative variants: OFA_{L1}, which excludes $D_{AE,feat}$ from Eq. 1, and OFA_{AE}, which excludes $D_{L1,feat}$ and only uses learnable latent representations of the predictors. Finally, Ablation Study 4 (AS4) aims to assess the impact of each variable on the relationship between X_{OFA} and FL-XGB's performance on C_{OF}, as well as the impact of other variables that were not included in the final X_{OFA} definition, such as the alignment between the feature importance vectors obtained from the FL-XGB model and those from a locally-trained XGBoost.

It is important to note the absence of comparisons with similar state-of-the-art approaches, due to the lack of directly comparable work in the literature. Instead, our focus has been on exploring and validating this novel approach within the context of our datasets and operational constraints.

2.6 Evaluation Metrics

To evaluate the performance of the FL-XGB model and the OFA predictor, AUC is selected as the main performance metric due to its effectiveness in providing a balanced evaluation of model discrimination. Additionally, Sensitivity ($Sens$) and Specificity ($Spec$) are used to gauge the accuracy in identifying DR and control patients, respectively.

To assess whether X_{OFA} correlates significantly with AUC predicted by FL-XGB, statistical measures such as p-value and r^2 are employed. If the regressed linear model (Eq. 2) shows $p > 0.05$ and/or $r^2 < 0.6$, it is discarded as statistically insignificant. Additionally, the effectiveness of the OFA prediction is measured via the difference, referred to as ΔAUC, between the predicted AUC_{OFA} and the actual AUC obtained by the FL-XGB model.

3 Results

Table 2 summarizes the performance of FL-XGB compared to locally-trained XGBoost models on IF centers. FL-XGB achieves an average $\overline{AUC}_{IF}$ of 75.27%, closely matching the 75.21% average AUC of the local models. Notably, FL-XGB improves $Sens$ to 69.51%, up from 66.46% in local models, enhancing the ability to correctly identify positive cases. However, this improvement in $Sens$ comes at the expense of $Spec$, which decreases to 68.43% compared to 86.24% in local models. Center ID3 demonstrates the most substantial improvement from the federation with respect to the locally-trained model, with a 4.57% increase in AUC and a 5.23% increase in $Sens$. The last two columns of Table 2 include results from AS1, which was meant to analyze which objective function yielded best results among BCE, w-BCE, and f-BCE. While the last two functions are specifically tailored for imbalanced datasets, BCE achieves the best performance, with an average AUC of 75.27%. FL-XGB trained with f-BCE achieves the highest $Sens$ score (84.45% against 69.51% from BCE), but this comes at the expense of $Spec$, which decreases to 40.79% (against 68.43% with BCE). Given the crucial importance of $Spec$, especially in healthcare, BCE is considered the optimal loss function.

Table 3 shows results from AS2, i.e., how different federation compositions impact FL-XGB performance across diverse clinical scenarios. $Conf_{IF3}$ includes centers ID0, ID1, and ID2; $Conf_{IF4}$ includes centers ID0, ID1, ID2, and ID4; $Conf_{IF5}$ is the federation introduced in Fig. 2 (ID0, ID1, ID2, ID3, and ID4); additionally, ID5 was included in $Conf_{IF6}$, and ID5 and ID8 were included in $Conf_{IF7}$. Both ID5 and ID8 meet three out of four inclusion criteria outlined in Sec. 2.1. The optimal federation results to be $Conf_{IF5}$.

Table 2: Performance comparison of local XGBoost models and Federated Learning (FL) with XGBoost model (FL-XGB) using different objective functions (BCE, w-BCE, and f-BCE, as per Ablation Study 1) is reported in terms of Area Under the Curve (*AUC*), Specificity (*Spec*) and Sensitivity (*Sens*). All metrics are expressed as a percentage (%).

IF Center	**Local XGBoost**			**FL-XGB (BCE)**			**FL-XGB (w-BCE)**			**FL-XGB (f-BCE)**		
	AUC	*Spec*	*Sens*	*AUC*	*Spec*	*Sens*	*AUC*	*Spec*	*Sens*	*AUC*	*Spec*	*Sens*
ID0	74.09	85.99	66.23	74.68	67.67	67.49	72.35	67.67	67.49	70.71	30.97	89.13
ID1	75.51	88.98	65.59	76.10	65.69	69.39	72.82	65.69	69.39	72.16	49.93	80.88
ID2	78.68	86.18	69.83	74.13	64.84	66.29	67.53	64.84	66.29	68.86	44.44	80.41
ID3	72.15	79.17	66.76	76.72	76.87	71.99	76.15	76.87	71.99	76.90	27.18	96.71
ID4	75.61	90.89	63.90	74.75	67.12	72.38	69.27	67.12	72.38	70.59	51.43	75.12
Average	75.21	86.24	66.46	**75.27**	**68.43**	69.51	71.62	68.42	69.50	71.85	40.79	**84.45**

Table 3: Results from Ablation Study 2 (AS2), in which different configurations with varying numbers of In-Federation (IF) centers are tested. Average test performance are reported in terms of Area Under the Curve (AUC), Specificity (Spec), and Sensitivity (Sens) for each configuration. Metrics are expressed as a percentage (%).

Configuration	**ID0**	**ID1**	**ID2**	**ID3**	**ID4**	**ID5**	**ID8**	*AUC*	*Spec*	*Sens*
$Conf_{IF3}$ (3 centers)	✓	✓	✓	×	×	×	×	71.19	38.78	79.81
$Conf_{IF4}$ (4 centers)	✓	✓	✓	×	✓	×	×	71.63	30.27	84.67
$\boldsymbol{Conf_{IF5}}$ (5 centers)	✓	✓	✓	✓	✓	×	×	**75.27**	**68.43**	**69.51**
$Conf_{IF6}$ (6 centers)	✓	✓	✓	✓	✓	✓	×	72.61	44.16	77.39
$Conf_{IF7}$ (7 centers)	✓	✓	✓	✓	✓	✓	✓	71.79	40.22	80.23

Extending our analysis from the federation to the unseen OOF centers, we assess how FL-XGB performs beyond the training environment. Table 4 presents the performance of the FL-XGB in terms of *AUC* for all OOF centers, across which the FL model scores an average *AUC* of 58.16%, lower than $\overline{AUC}_{IF}$ (75.27%) reported for IF centers. Particularly poor predictions are obtained for ID13 and ID19 (*AUC* of 22.22% and 33.33%, respectively), while FL performance on ID7 ($AUC = 76.08\%$) is stronger than $\overline{AUC}_{IF}$.

Table 4 also shows the results of the OFA predictor, which assesses the applicability of FL-XGB to OOF. The upper part of the table includes the centers involved in the OFA regression process, whereas the lower part shows the two centers (ID9 and ID16) used to test the regressed predictor's accuracy. For each OOF center, FL-XGB results are compared to those obtained from the OFA predictor and its variants, OFA_{L1} and OFA_{AE}. The OFA predictor demonstrates an average prediction error (i.e., ΔAUC) of 6.13% and an average *OSS* of 63.70%. Centers ID5, ID6, ID7, and ID8 exhibit the highest *OSS* values, around 72%: ID5 is correctly predicted by OFA ($\Delta AUC = -0.59\%$); ID7 and ID8 are slightly under and overestimated by OFA ($\Delta AUC = -7.63\%$ and $\Delta AUC = +7.50\%$, respectively); while ID6 is largely overestimated by OFA ($\Delta AUC = +17.15\%$). ID13 and D19 obtain the lowest *OSS* values (39.20% and 42.45%), which is due to the fact that, although FL-XGB performance on both OOF centers is correctly predicted by OFA ($\Delta AUC = +4.16\%$ and $\Delta AUC = -2.72\%$), these

Table 4: Performance comparison between FL-XGB and OFA predictor in its variants: OFA as fully defined in Eq. 1, OFA_{L1} using only statistical information, and OFA_{AE} based on autoencoder latent space features only (Ablation Study 3). The regression prediction error (ΔAUC) represents the average absolute difference between the AUC_{OFA} values and the AUC values scored by FL-XGB. Similarly, the OSS score is reported for each OFA variant. The bottom part of the table reports the performance calculated over the two OOF centers (ID9 and ID16) excluded from the OFA regression process. All metrics are expressed as a percentage (%).

OOF Center	**FL-XGB**	**OFA**		**OFA_{L1}**		**OFA_{AE}**	
	AUC	ΔAUC	OSS	ΔAUC	OSS	ΔAUC	OSS
ID5	69.26	-0.59	71.64	-0.30	68.77	+1.53	59.60
ID6	51.14	+17.15	71.35	+17.21	68.32	+18.81	59.08
ID7	76.08	-7.63	71.47	-7.79	68.28	-5.50	59.47
ID8	61.07	+7.50	71.56	+7.75	68.67	+9.50	59.47
ID10	48.72	+5.58	60.62	+8.21	59.94	-7.51	41.15
ID11	64.06	+2.74	70.20	+2.95	67.34	+2.71	57.10
ID12	57.00	+2.73	64.78	+1.91	61.39	-2.38	49.51
ID13	22.22	+4.16	39.20	+3.02	36.68	+12.01	36.79
ID14	56.94	-0.57	62.20	-1.27	59.01	-7.38	46.36
ID15	73.91	-10.51	67.60	-11.37	64.06	-12.21	53.93
ID17	73.33	-7.47	69.48	-9.78	64.80	-6.59	57.08
ID18	69.00	-10.36	63.94	-10.01	61.45	-12.38	50.76
ID19	33.33	-2.72	42.45	-0.54	42.22	+9.40	42.10
Average	58.16	**6.13**	**63.70**	6.32	60.84	8.30	51.72
ID9	68.37	-3.60	68.65	-	-	-	-
ID16	56.35	-2.61	60.19	-	-	-	-

two centers are the ones on which FL-XGB performs the worst ($AUC = 22.22\%$ and $AUC = 33.33\%$). For the remaining OOF centers, OSS values range between 60.62% and 70.20%, while ΔAUC values vary from -10.51% to +5.58%. The last columns of Table 4 report the results from AS3, i.e., the comparison between OFA and its variants, OFA_{L1} and OFA_{AE}. From this comparison, the complete OFA model demonstrates superior accuracy. It achieves a lower average ΔAUC of 6.13%, compared to 6.32% for OFA_{L1} and 8.30% for OFA_{AE}, indicating that the integration of both statistics and latent representations provides a more comprehensive assessment of the applicability of FL-XGB to C_{OF}. On ID9 and ID16, FL-XGB scores AUC values of 68.37% and 56.35%, while the OFA predictor proves quite accurate, with ΔAUC of -3.60% and -2.61%. The resultant OSS values are 68.65% for ID9 and 60.19% for ID16.

The performance of the OFA predictor across the OOF centers is also represented in Fig. 3, which illustrates the relationship between the designed X_{OFA} comprehensive variable and the actual AUC values scored by FL-XGB on OOF centers. The linear regression $f : X_{AUC} \rightarrow AUC$, shown as a black line, explains 76.70% of the variance (as indicated by r^2), with a statistically significant p-value ($\ll 0.001$), underscoring OFA's effectiveness in capturing the correlation

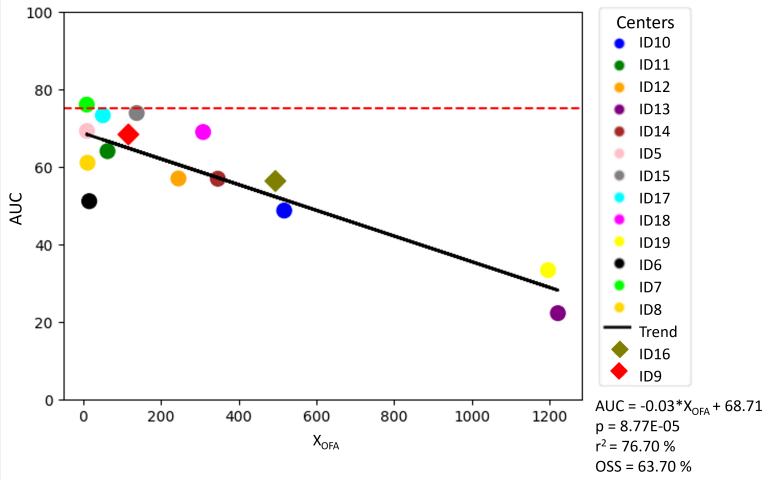


Fig. 3: Scatter plot illustrating the relationship between OFA predictions and actual AUC test values for OOF centers. The black line represents the linear regression between X_{OFA} and AUC predicted by FL-XGB. The regressed linear equation, r^2, $p-value$, and average OOF Suitability Score (OSS) are also reported. The dashed red line indicates the average performance of the FL-XGB model on IF centers ($\overline{AUC}_{IF}$ = 75.27%). All OOF centers are represented with a colored circle, except for ID9 and ID16 shown as diamonds, as they were excluded from the linear regression and used to evaluate the regressed OFA model.

between AUC performance from a pre-trained FL-XGB and the information about C_{OF} summarized in X_{OFA}. The effectiveness of the OFA predictor is further underscored in Table 5, which outlines the results from AS4, i.e., the impact of each variable in explaining the correlation between performance from FL-XGB on C_{OF} and X_{OFA}. The final row shows that X_{OFA}, as defined in Eq. 1, can explain up to 76.7% of the variance of the distribution shown in Fig. 3.

4 Discussion

Deploying FL models in real-world scenarios is challenging due to not only data variability across centers, but also internal variability [30]. Differences in demographics, healthcare standards, and acquisition protocols further increase heterogeneity, which impacts both training and deployment. While FL enables privacy-compliant collaborative learning, it does not automatically guarantee the same level of generalization to every center, whether within or outside the federation. Hence, rather than focusing on building a universally generalizable model, this study aims to explore the feasibility of deploying an existing FL model on data from unseen OOF centers. This perspective shifts the focus from generalization to compatibility estimation between a trained FL model and a new center.

Table 5: Results from Ablation Study 4 (AS4): impact of the variables involved in X_{OFA} in predicting the performance from FL-XGB on a given C_{OF}. The variables are: imbalance (η), cardinality (n), distances between features distributions ($D_{L1,feat}$ and $D_{AE,feat}$), distance between missing values distributions ($D_{L1,miss}$), and predicted probability (or confidence, p_{pred}). Each row is associated with the final r^2, i.e., the explained variance in the relationship between X_{OFA} and AUC values scored by FL-XGB on OOF centers.The last row corresponds to the definition of X_{OFA} in Eq. 1.

OFA variables	r^2 **(%)**
η / n	52.9
$\eta \times D_{AE,feat}/n$	54.0
$\eta \times D_{L1,feat}$ / n	70.6
$\eta \times D_{L1,feat} \times D_{AE,feat}$ / n	72.1
$\eta \times D_{L1,feat} \times D_{AE,feat} \times D_{L1,miss}$ / n	75.7
$\eta \times D_{L1,feat} \times D_{AE,feat} \times D_{L1,miss} \times p_{pred}$ / n	76.7

To investigate this, FL-XGB has been trained and tested on five IF centers. As shown in Table 2, its average performance across the federation ($\overline{AUC}_{IF}$ = 75.27%) closely matches that of locally trained models (75.21%). However, FL-XGB demonstrates greater consistency across centers, with a lower standard deviation in AUC (1.08 vs. 2.39 for local XGBoost). This stability is likely due to FL mitigating overfitting to center-specific data characteristics. Notably, FL-XGB also improves $Sens$ by +8.48%, a crucial advantage for DR early detection, where accurately identifying at-risk patients is essential.

Beyond the training phase, applying FL models to OOF centers introduces further complexity, as their data distributions may significantly differ from IF data. Notably, our goal is not to demonstrate that FL-XGB performs well on all OOF centers, but to provide a tool (the OFA predictor) that can predict whether an FL model is suitable for a given OOF dataset before deployment. FL-XGB's performance on OOF centers is, in many cases, comparable to that observed on the IF centers. Centers ID5, ID7, ID15, ID17, ID18, and ID9 share similar data characteristics with the IF centers (see Fig. 2), which results in strong performance from FL-XGB ($AUC > 69\%$) despite limitations such as the high volume of missing data or small and strongly unbalanced datasets. This suggests that FL-XGB itself has strong potential for real-world application, particularly valuable for centers that, due to limited data resources, could not train their own models. To estimate the performance of an FL model on OOF data, the OFA predictor has been developed and applied to the case study of DR risk prediction using real-world EHRs from 20 diabetic centers. OFA uses 13 OOF centers to regress a linear relationship between AUC from FL-XGB and X_{OFA}, a special variable designed to represent key features of an OOF center and its differences from the IF centers. As shown in Table 4, OFA effectively detects OOF centers that are incompatible with FL-XGB. For example, the poor performance of FL-XGB on ID13 and ID19 (AUC of 22.22% and 33.33%) is accurately identified by X_{OFA} (see Fig. 3), which can be seen as a function

that maps OOF centers based on their compatibility with FL-XGB. Additionally, OSS values discourage the use of FL-XGB on these unsuitable centers (39.20% for ID13 and 42.45% for ID19). On the other hand, for centers where $OSS \geq 60\%$, OFA maintains prediction errors ΔAUC within an 8-point margin and only a few cases reaching higher error rates, with a maximum ΔAUC of +17.15 points. Also, when tested on ID9 and ID16, centers left out of the regression process, OFA displays prediction errors ΔAUC within a tolerable range ($< 4\%$), proving its consistency on OOF data.

Table 5 shows the impact of each variable in explaining the correlation between performance from FL-XGB on C_{OF} and X_{OFA}. The basic variables η and n accounted for 52.9% of the AUC-X_{OFA} variance. The integration of the $D_{L1,feat}$ considerably improved the regression model, elevating r^2 to 70.6%, which is further refined by the introduction of $D_{L1,feat}$ and $D_{AE,feat}$ ($r^2 = 72.1\%$). This suggests that feature distribution statistics have a crucial role in explaining the difference between OOF and IF centers, as well as that latent features capture additional nuances not evident through direct statistical measures alone. Final X_{OFA} also includes $D_{L1,miss}$ and p_{pred}, reaching r^2 of 76.7%. This importance is further confirmed by comparing the performance of the two main OFA variants, OFA_{L1} (using only statistical information) and OFA_{AE} (leveraging only distances on latent space features), as reported in Table 4. Although the overall behaviors of the OFA variants are similar, both OFA_{L1} and OFA_{AE} tend to exhibit larger errors, especially for centers with moderate compatibility.Figure 3 illustrates the regression accuracy for the relationship between X_{OFA} and AUC scores from FL-XGB on the OOF datasets. The OFA predictor seems to capture the information needed to separate the datasets, aligning reasonably well with actual AUC scores. However, centers ID5, ID6, ID7 and ID8 present a challenge: they have similar X_{OFA} but exhibit substantial variation in FL-XGB performance, ranging from approximately 50% to 76% in AUC. This suggests that the variables included in X_{OFA} may not fully explain the relationship between X_{OFA} and FL-XGB's AUC.

It is worth mentioning that OFA's performance can be heavily influenced by FL-XGB's training parameters. In particular, XGBoost is often trained with a feature subsampling approach, in which only some predictors are randomly selected during model training. While this method reduces the risk of overfitting and enhances model reliability, it can lead to inconsistent outputs and suboptimal feature selection, particularly in datasets with extensive missing data or imbalanced features [31]. In datasets characterized by high sparsity, random feature selection can skew model reproducibility and reliability, depending on the predictors selected during training.To address this issue, we trained the FL-XGB model using all available features, avoiding random column subsampling for weak learners. This approach balances the trade-off between maximizing accuracy and ensuring robustness and reproducibility across diverse healthcare settings. Another important aspect concerns model interpretability and clinical validity. Future work will explore strategies to ensure that the estimated feature importance aligns with its actual clinical relevance.

We also plan to assess the generalizability of the OFA predictor on other medical use cases and multi-centric datasets. For instance, we will test it on EHRs from general practitioners [32] to evaluate its robustness in a different clinical context. In parallel, we will investigate alternative base models and FL strategies to verify the OFA stability under different ML configurations. While XGBoost was selected for its well-established effectiveness with tabular data, class imbalance, and missing values [6,7,8,9,10], exploring other models and learning paradigms may reveal complementary strengths or uncover limitations in the current implementation. Further directions include validating OFA in vertical FL settings and through multi-view analyses [33], as well as integrating differential privacy techniques [34] to enhance data protection and ensure privacy compliance in real-world deployments.

5 Conclusion

This study proposed an innovative OFA predictor which, by combining statistical and latent features of OOF data, demonstrated statistical significance in determining whether an FL model can be effectively applied to OOF centers, thus answering the research question in Sec. 1. Many clinical centers, due to their data characteristics or limited resources, cannot develop their own model, but could substantially benefit from accessing a robust FL model. The flexibility and privacy focus of the FL-XGB framework, combined with the OFA predictor's ability to evaluate its compatibility with OOF data, offer a valuable solution for DR screening across diverse clinical settings, potentially making a significant impact on diabetic preventive care. The experimental findings suggest how the OFA predictor may play a key role in ensuring the safe and effective deployment of FL models in OOF settings, offering a new paradigm for the adaptability and scalability of AI in clinical practice.

Disclosure of Interests. The authors did not receive support from any funding organization in the public, commercial, or not-for-profit sectors for the present work.

References

1. Tien-En Tan and Tien Yin Wong. Diabetic retinopathy: Looking forward to 2030. *Frontiers in Endocrinology*, 13:1077669, 2023.
2. Anila Sebastian, Omar Elharrouss, Somaya Al-Maadeed, and Noor Almaadeed. A survey on deep-learning-based diabetic retinopathy classification. *Diagnostics*, 13(3):345, 2023.
3. Ayoub Skouta, Abdelali Elmoufidi, Said Jai-Andaloussi, and Ouail Ouchetto. Deep learning for diabetic retinopathy assessments: a literature review. *Multimedia Tools and Applications*, 82(27):41701–41766, 2023.
4. Nurul Mirza Afiqah Tajudin, Kuryati Kipli, Muhammad Hamdi Mahmood, Lik Thai Lim, Dayang Azra Awang Mat, Rohana Sapawi, Siti Kudnie Sahari, Kasumawati Lias, Suriati Khartini Jali, and Mohammed Enamul Hoque. Deep learning in the grading of diabetic retinopathy: A review. *IET Computer Vision*, 16(8):667–682, 2022.

5. Archana Tapuria, Talya Porat, Dipak Kalra, Glen Dsouza, Sun Xiaohui, and Vasa Curcin. Impact of patient access to their electronic health record: Systematic review. *Informatics for Health and Social Care*, 46(2):194–206, 2021.
6. Michele Bernardini, Luca Romeo, Adriano Mancini, and Emanuele Frontoni. A clinical decision support system to stratify the temporal risk of diabetic retinopathy. *IEEE Access*, 9:151864–151872, 2021.
7. Cheng Yang, Qingyang Liu, Haike Guo, Min Zhang, Lixin Zhang, Guanrong Zhang, Jin Zeng, Zhongning Huang, Qianli Meng, and Ying Cui. Usefulness of machine learning for identification of referable diabetic retinopathy in a large-scale population-based study. *Frontiers in Medicine*, 8:773881, 2021.
8. Antonio Nicolucci, Luca Romeo, Michele Bernardini, Marco Vespasiani, Maria Chiara Rossi, Massimiliano Petrelli, Antonio Ceriello, Paolo Di Bartolo, Emanuele Frontoni, and Giacomo Vespasiani. Prediction of complications of type 2 diabetes: A machine learning approach. *Diabetes Research and Clinical Practice*, 190:110013, 2022.
9. Agnese Piersanti, Benedetta Salvatori, Piera D'Avino, Laura Burattini, Christian Göbl, Andrea Tura, and Micaela Morettini. Diabetic retinopathy detection: A machine-learning approach based on continuous glucose monitoring metrics. In *International Conference on e-Health and Bioengineering*, pages 763–773. Springer, 2023.
10. Xiaohua Wan, Ruihuan Zhang, Yanan Wang, Wei Wei, Biao Song, Lin Zhang, and Yanwei Hu. Predicting diabetic retinopathy based on routine laboratory tests by machine learning algorithms. *European Journal of Medical Research*, 30(1):1–19, 2025.
11. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.
12. Trung Kien Dang, Xiang Lan, Jianshu Weng, and Mengling Feng. Federated learning for electronic health records. *ACM Transactions on Intelligent Systems and Technology*, 13(5):1–17, 2022.
13. Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, 2018.
14. Wonsuk Oh and Girish N Nadkarni. Federated learning in health care using structured medical data. *Advances in Kidney Disease and Health*, 30(1):4–16, 2023.
15. Mohammad Moshawrab, Mehdi Adda, Abdenour Bouzouane, Hussein Ibrahim, and Ali Raad. Reviewing federated machine learning and its use in diseases prediction. *Sensors*, 23(4):2112, 2023.
16. Humayera Islam, Abu Mosa, et al. A federated mining approach on predicting diabetes-related complications: Demonstration using real-world clinical data. In *AMIA Annual Symposium Proceedings*, volume 2021, page 556, 2022.
17. Jiyoun Kim, Junu Kim, Kyunghoon Hur, and Edward Choi. EHRFL: Federated learning framework for heterogeneous EHRs and precision-guided selection of participating clients. *arXiv preprint arXiv:2404.13318*, 2024.
18. Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *IEEE Transactions on Information Theory*, 68(12):8076–8091, 2022.

19. Ahmed Elhussein and Gamze Gürsoy. Privacy-preserving patient clustering for personalized federated learnings. In *Machine Learning for Healthcare Conference*, pages 150–166. PMLR, 2023.
20. Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.
21. Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.
22. Yiqiang Chen, Wang Lu, Xin Qin, Jindong Wang, and Xing Xie. Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
23. Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.
24. A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.
25. Mengmeng Ma, Tang Li, and Xi Peng. Beyond the federation: Topology-aware federated learning for generalization to unseen clients. In *Forty-first International Conference on Machine Learning*, 2024.
26. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
27. Ping Zhang, Yiqiao Jia, and Youlin Shang. Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6):15501329221106935, 2022.
28. Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harmander Kaur. The use of electronic health records to support population health: a systematic review of the literature. *Journal of Medical Systems*, 42(11):214, 2018.
29. Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.
30. Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B Rawat, and Vladimir Vlassov. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 2023.
31. Jack Dunn, Luca Mingardi, and Ying Daisy Zhuo. Comparing interpretability and explainability for feature selection. *arXiv preprint arXiv:2105.05328*, 2021.
32. Michele Bernardini, Luca Romeo, Emanuele Frontoni, and Massih-Reza Amini. A semi-supervised multi-task learning approach for predicting short-term kidney disease evolution. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3983–3994, 2021.
33. Chenyang Ma, Xinchi Qiu, Daniel Beutel, and Nicholas Lane. Gradient-less federated gradient boosting tree with learnable learning rates. In *Proceedings of the 3rd Workshop on Machine Learning and Systems*, pages 56–63, 2023.
34. Bakary Dolo, Faiza Loukil, and Khouloud Boukadi. Early detection of diabetes mellitus using differentially private sgd in federated learning. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications*, pages 1–8, 2022.