Review

Explainable AI Approaches in Federated Learning: Systematic Review

Titus Tunduny, MSc; Bernard Shibwabo, PhD

School of Computing & Engineering Sciences, Strathmore University, Nairobi, Kenya

Corresponding Author:
Titus Tunduny, MSc
School of Computing & Engineering Sciences
Strathmore University
PO Box 59857 – 00200
Nairobi
Kenya
Phone: 254 728778002
Email: ttunduny@gmail.com

Abstract

Background: Artificial intelligence (AI) has, in the recent past, experienced a rebirth with the growth of generative AI systems such as ChatGPT and Bard. These systems are trained with billions of parameters and have enabled widespread accessibility and understanding of AI among different user groups. Widespread adoption of AI has led to the need for understanding how machine learning (ML) models operate to build trust in them. An understanding of how these models generate their results remains a huge challenge that explainable AI seeks to solve. Federated learning (FL) grew out of the need to have privacy-preserving AI by having ML models that are decentralized but still share model parameters with a global model.

Objective: This study sought to examine the extent of development of the explainable AI field within the FL environment in relation to the main contributions made, the types of FL, the sectors it is applied to, the models used, the methods applied by each study, and the databases from which sources are obtained.

Methods: A systematic search in 8 electronic databases, namely, Web of Science Core Collection, Scopus, PubMed, ACM Digital Library, IEEE Xplore, Mendeley, BASE, and Google Scholar, was undertaken.

Results: A review of 26 studies revealed that research on explainable FL is steadily growing despite being concentrated in Europe and Asia. The key determinants of FL use were data privacy and limited training data. Horizontal FL remains the preferred approach for federated ML, whereas post hoc explainability techniques were preferred.

Conclusions: There is potential for development of novel approaches and improvement of existing approaches in the explainable FL field, especially for critical areas.

Trial Registration: OSF Registries 10.17605/OSF.IO/Y85WA; https://osf.io/y85wa

(JMIR AI 2026;5:e69985) doi: 10.2196/69985

KEYWORDS

explainable artificial intelligence; federated learning; explainable federated artificial intelligence; privacy-preserving machine learning; model interpretability

Introduction

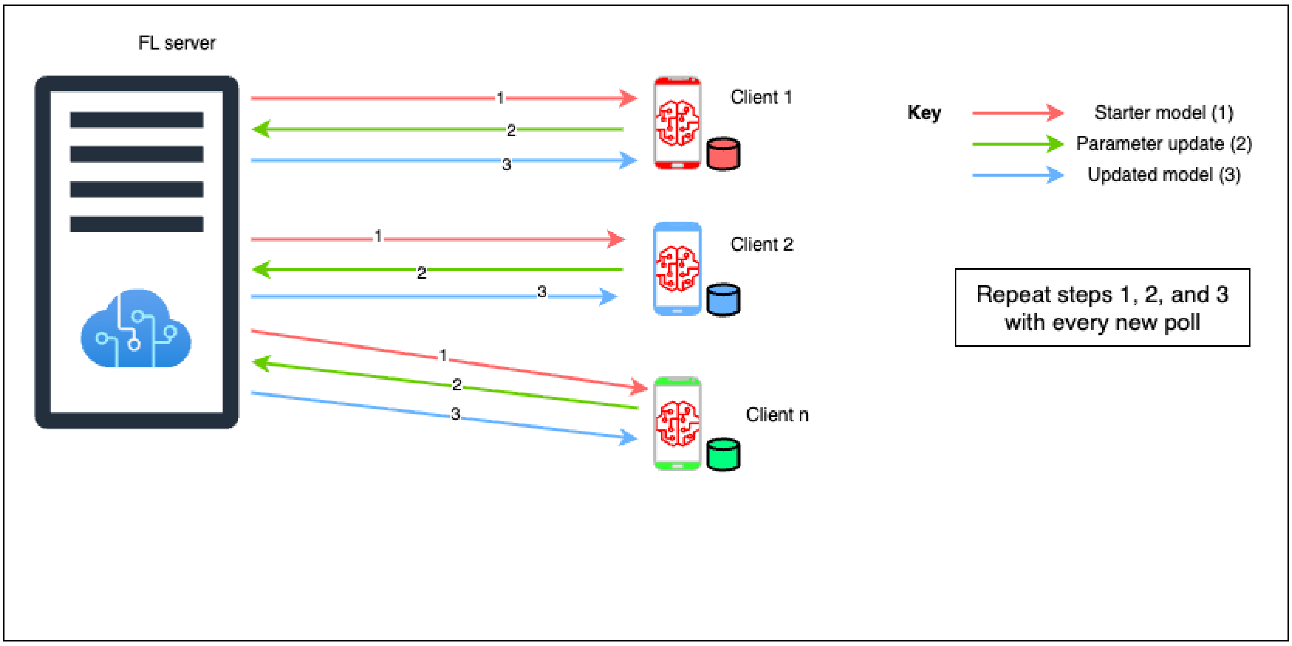
Background

Machine learning (ML) has become increasingly prevalent in critical sectors such as health care and security [1,2] driven by the need to process copious amounts of edge device data [3]. However, highly performant ML algorithms often operate as “black boxes” [4,5], creating a need for ML explainability to build trust. This has led to increased research in the field of explainable artificial intelligence (XAI) [2,4,6]. How a ML model works is important in building trust and reliability in its prediction or classification results, especially in critical areas. XAI approaches such as linear interpretable model-agnostic explanations (LIME) [7] and Shapley Additive Explanations (SHAP) [8] perform well with centralized models, although challenges remain [9]. Growing data privacy legislation such as the General Data Protection Regulation [10], HIPAA (Health Insurance Portability and Accountability Act) [11], and Kenya’s

Data Protection Act [12] have further complicated centralized ML development.

Federated learning (FL), introduced by McMahan et al [13] in 2016, enables privacy-preserving training on decentralized data stored on edge devices [13,14]. A central server distributes a global model to clients, who train it locally and send updates (learned parameters) back, ensuring that data never leave the device. The federated ML process is outlined in Figure 1. These updates are aggregated from selected clients (polling) typically using the federated average algorithm [13] to refine the global model. This process is repeated over several rounds, preserving privacy while improving model performance [15]. The federated averaging algorithm is outlined in Textbox 1.

Figure 1. Federated machine learning process showing global model distribution and update of the global model on the federated learning (FL) aggregation server.



Textbox 1. Federated averaging algorithm showing its mechanism.

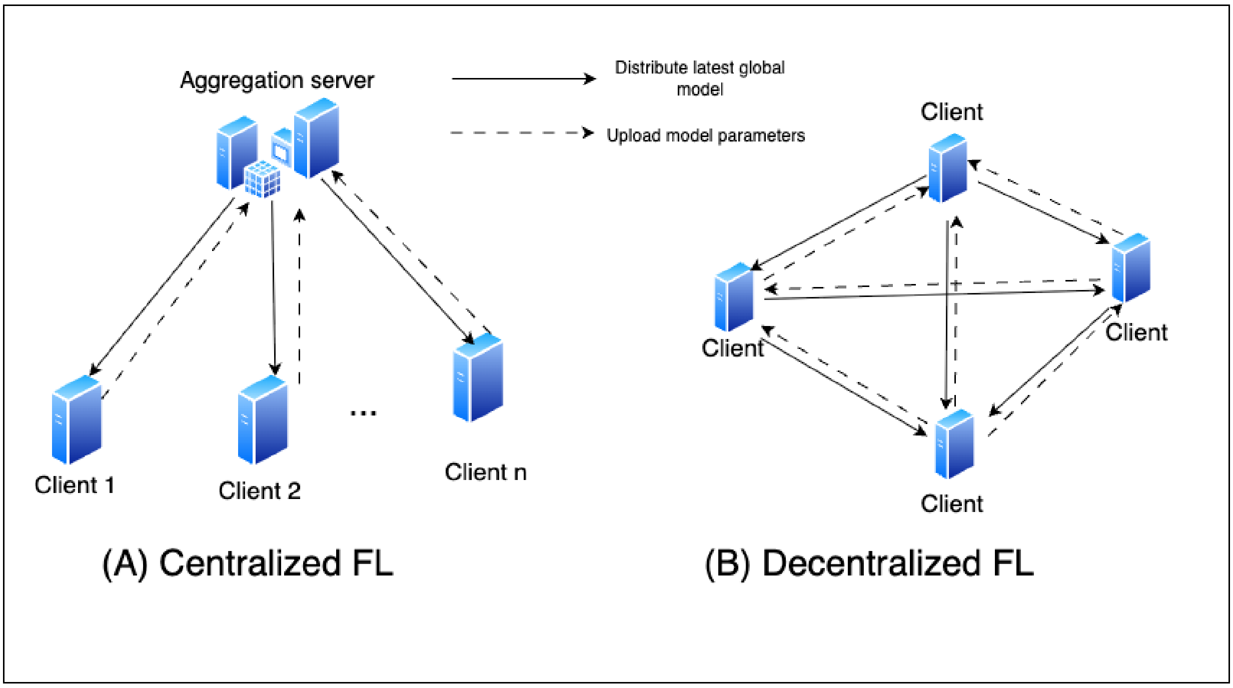
```
Instructions
Initialize global model weights $w_0$
For each communication round $t$= 1, 2,..., $T$ do
Server sends current model weights $w_t$ to a subset of clients
Each selected client $k$ trains on local data for $E$ epochs with learning rate η:
$w_{t+1}^k = w_t - \eta \nabla \ell(w_t; \xi)$ , where ξ is a batch of local data
Clients send updated weights ${w_{t+1}}^k$ back to the server
Server aggregates client updates:
$w_{t+1}^k = \Sigma \left(\frac{n_k}{n}\right) w_{t+1}^k$ (weighted by client data size)
End For
Return final global model weights $w$*
```

FL has demonstrated its potential as a privacy-preserving technique suitable for real-world applications despite its challenges [16,17]. However, its deployment in sensitive domains such as patient-embedded devices requires a high level of trust. This opens up significant research opportunities in integrating XAI techniques in FL environments. By enabling explanations on model generalizations at the data source while maintaining privacy, XAI can offer real-time benefits and enhance trust in artificial intelligence (AI)–driven embedded systems. FL can be categorized based on communication architecture or data partitioning. By communication architecture, FL models can be categorized as centralized or decentralized. By data partitioning, FL models can be categorized as horizontal, vertical, or transfer learning (TL) [18].

Centralized FL

In centralized FL (CFL), a global model is shared with various clients, who train it locally and send back the learned parameters. The server aggregates these updated parameters using algorithms such as federated averaging to improve the global model. Clients are selected through polling, and differential privacy can be applied by adding noise to the updates. CFL faces challenges such as client heterogeneity, limited communication and computing resources, fairness, security, and trust [19]. The structure of CFL is shown in Figure 2A.

Figure 2. Centralized and decentralized federated learning (FL) in action.



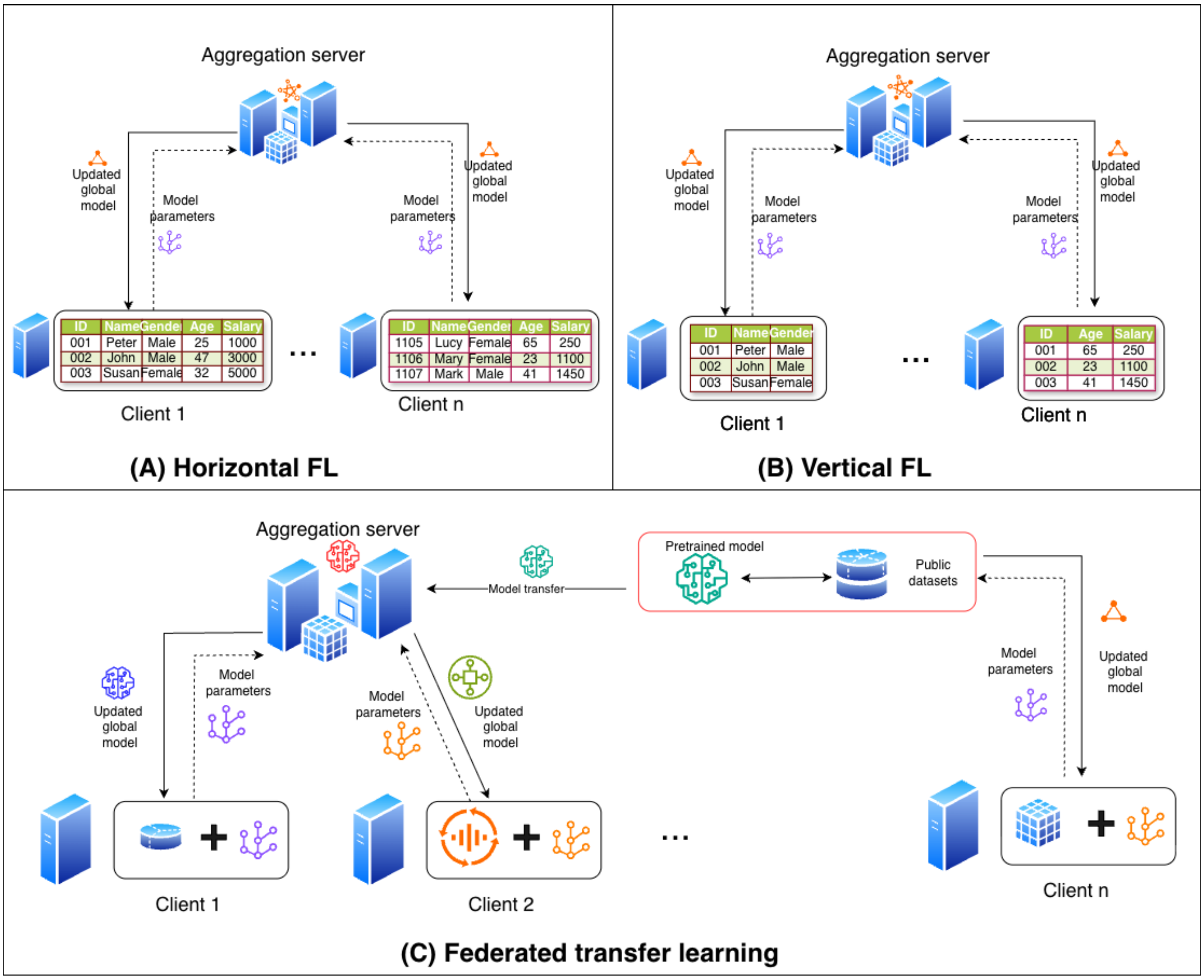
Decentralized FL

Decentralized FL—also known as distributed FL—eliminates the need for a central server. Each client trains a local model and shares the parameters with their peers using protocols such as pointing, gossip, and broadcast. Clients act as both learners and aggregators while refining their model based on peer updates. Therefore, the global model is developed from peer to peer [20,21]. The structure of decentralized FL is shown in Figure 2B.

Horizontal FL

Horizontal FL (HFL) involves clients that share the same data features but have different data samples. Each client holds instances with similar attributes (eg, name, gender, date of birth, and salary), but the individual records (samples and rows) differ. This setup is ideal when datasets have high feature overlap across clients but differ in the entities they contain [22]. Figure 3A depicts the structure of HFL.

Figure 3. Federated learning (FL) types, showing horizontal FL, vertical FL, and federated transfer learning.



ID	Name	Gender	Age	Salary
001	Peter	Male	25	1000
002	John	Male	47	3000
003	Susan	Female	32	5000

ID	Name	Gender	Age	Salary
1105	Lucy	Female	65	250
1106	Mary	Female	23	1100
1107	Mark	Male	41	1450

ID	Name	Gender
001	Peter	Male
002	John	Male
003	Susan	Female

ID	Age	Salary
001	65	250
002	23	1100
003	41	1450

Vertical FL

Vertical FL (VFL) is where clients share the same data samples but have different feature sets. Each client holds part of the information for the same users; for example, one client may have demographic data, whereas another may have financial data. VFL is ideal when full data sharing is not possible, such as in health care settings with multiple institutions holding complementary patient data [23]. Figure 3B shows the structure of VFL.

Federated TL

Federated TL (FTL) merges the concepts of FL and TL. In FTL, a pretrained model from a related task is distributed to all the clients. Each client fine-tunes (adapts) the pretrained model using their local data. FTL is useful when training data are limited or privacy sensitive, such as in health care, allowing clients to benefit from existing models while preserving data privacy. FTL structured is showcased in Figure 3C.

Contributions

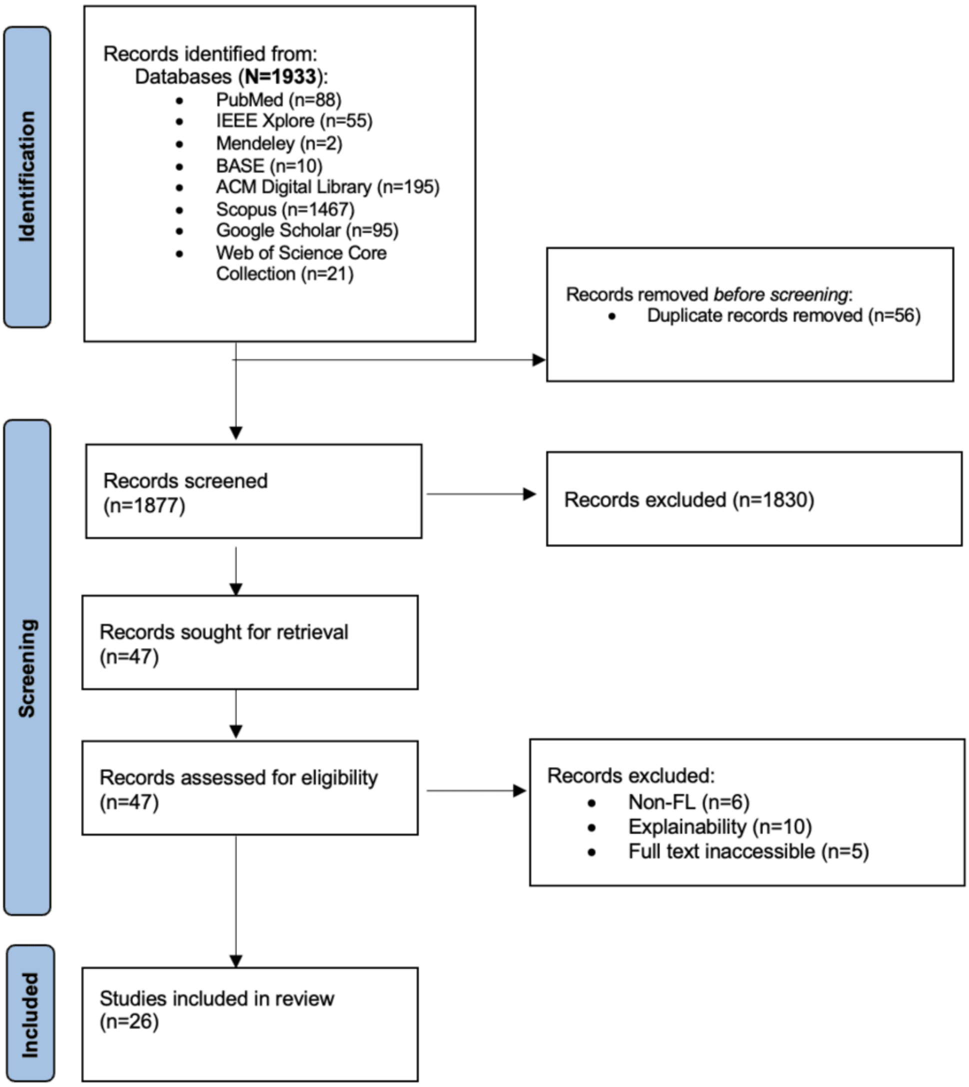
This study makes contributions to the field of explainable FL in the following ways: it offers original insights into the explainability of FL models, including the methods used to explain the models, whether novel or existing, and how they have been used. This study also delves into the deployment contexts for FL models, including the types of FL used. Unlike prior works such as the study by Singh et al [24], which broadly examines FL applications, and the study by Aggarwal et al [25], which explores general FL use cases, this study also focused on the application areas for explainable FL models and their associated challenges, as well as providing the direction of the trends.

Methods

Overview

This study followed established guidelines for systematic literature review studies [26] and adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting standards (Figure 4) [27]. Its main objective was to assess the development of XAI within FL. To achieve this, the following review questions were formulated.

Figure 4. PRISMA flowchart for selection of systematic review literature. FL: federated learning.


Identification
Records identified from:
Databases (N=1933):
PubMed (n=88)
IEEE Xplore (n=55)
Mendeley (n=2)
BASE (n=10)
ACM Digital Library (n=195)
Scopus (n=1467)
Google Scholar (n=95)
Web of Science Core Collection (n=21)
Records removed before screening:
Duplicate records removed (n=56)
Screening
Records screened (n=1877)
Records excluded (n=1830)
Records sought for retrieval (n=47)
Records assessed for eligibility (n=47)
Records excluded:
Non-FL (n=6)
Explainability (n=10)
Full text inaccessible (n=5)
Included
Studies included in review (n=26)

Research Questions

To understand the explainable approaches in FL, research questions (RQs) were raised and grouped under 1 of 3 categories.

RQ 1: Trends and Contributions

To understand the contributions of the existing literature, three questions were raised: (1) when were the explainable FL studies published? (2) In which countries or regions are the studies or study applications located, or which countries or regions are the authors of the studies affiliated with? (3) What are the main contributions of the studies identified?

RQ 2: Application Areas

The application areas for FL, coupled with the application areas for explainability, were explored based on the following questions: (1) what are the application areas of explainable FL models? (2) What types of FL have been applied in the studies? (3) Why was FL adopted in the studies?

RQ 3: Model Explainability

The XAI models and their categories were reviewed based on the following questions: (1) which XAI algorithms or models have been applied or used in the studies? (2) What category of XAI do the models or algorithms used in the studies fall under? (3) What data sources or datasets (if available) were used in the development of the models used in the studies?

Search Strategy

The reported results followed the population, intervention, comparison, and outcome guidelines [28]. The search string generation process is outlined in Multimedia Appendix 1. The generated search string was adapted to the 8 different databases, as outlined in Multimedia Appendix 2.

Eligibility Criteria

Of the 1933 initial search results, 26 (1.3%) peer-reviewed studies published between 2016 and 2024 were selected. Inclusion was based on relevance to XAI within any FL context. Exclusion criteria included non–English-language papers, non–peer-reviewed studies, and inaccessible full texts and gray literature as they are not easily retrievable [29].

Screening

Screening was conducted by 2 independent reviewers using the CADIMA software [30]. Initial screening was based on the titles and abstracts, followed by a blind full-text review. Conflicts were resolved through discussion, and a third party was involved when there was lack of consensus. A strong interrater reliability was achieved, with a κ value of 0.74.

Data Extraction and Synthesis

Key details from the selected studies, such as title, authorship, affiliation, publication year, data used, and answers to the RQs, were extracted and synthesized using Google Sheets. This process was undertaken by 2 reviewers to minimize bias. Multimedia Appendix 3 contains all the data used for analysis and synthesis.

Quality Assessment

Overview

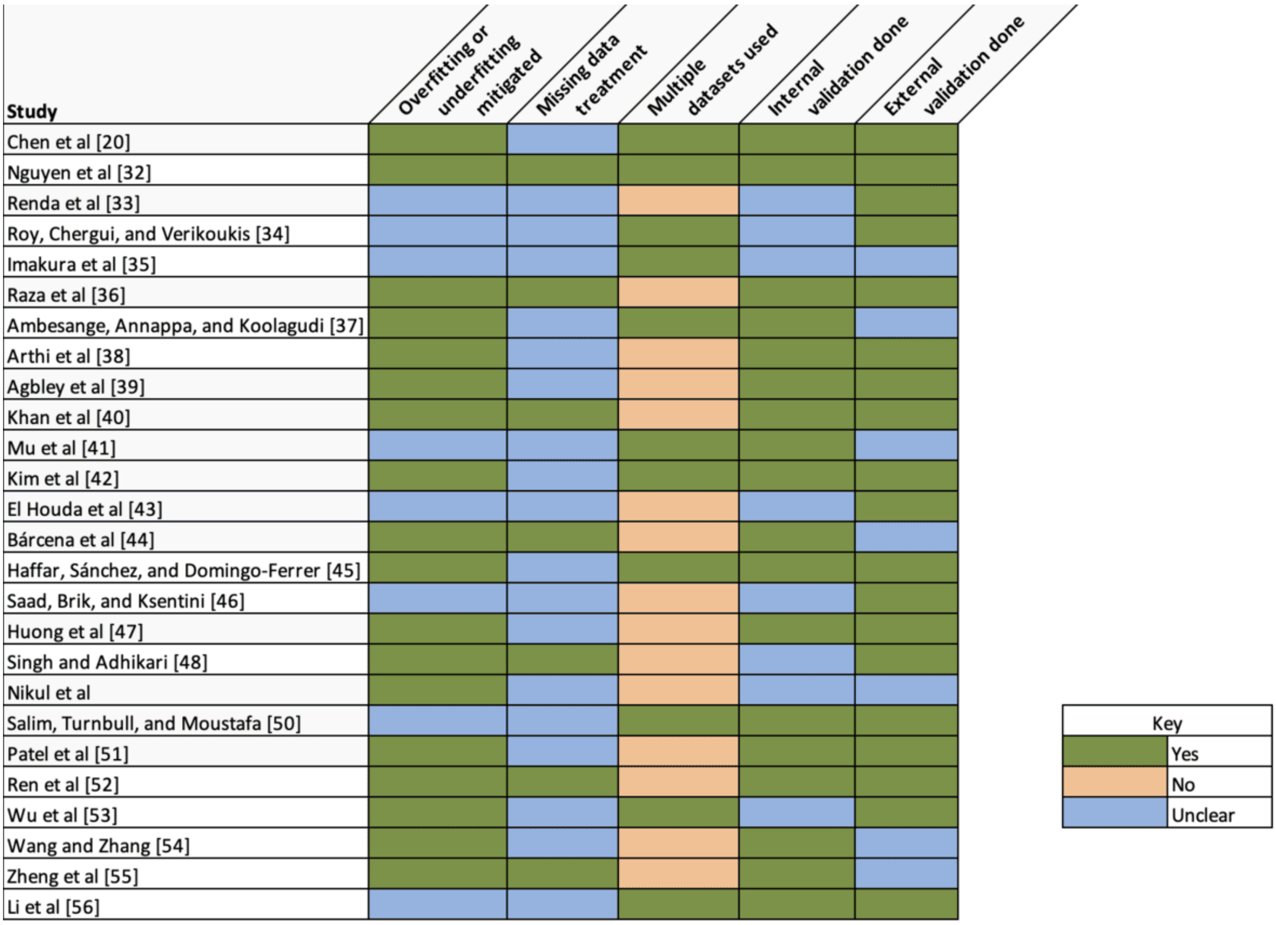
Quality assessment was undertaken by the 2 researchers (TT and BS) as recommended by Xiao and Watson [26]. The criteria used included handling of overfits, missing data, and use of multiple datasets and validation techniques. The evaluation was based on the PRISMA guidelines [27].

Risk-of-Bias Analysis: Individual Studies

The risk of bias of the individual studies focused on potential biases of data selection and model training. The criteria used included handling of overfit and underfit, missing data treatment, use of multiple datasets, and ML evaluation metrics. A total of 69% (18/26) of the studies reported clear mechanisms for mitigating against overfitting and underfitting. In total, 31 (8/26) of the studies lacked evidence of such mitigation. A total of 77% (20/26) of the studies did not address missing data treatment, increasing the risk of data and selection biases [31], especially as most of the studies used preexisting datasets.

Figure 5 [20,32-56] shows the risk of bias per study, highlighting how each implemented underfitting and overfitting, missing data treatment, use of multiple datasets, and internal and external validation. Missing data treatment was not clearly identified in most studies (19/26, 73%), with only 27% (7/26) reporting any treatment done. Internal and external validation was conducted in most of the studies (19/26, 73%).

Figure 5. Heat map showing risk mitigation by study for the selected studies.



Study	Overfitting or underfitting mitigated	Missing data treatment	Multiple datasets used	Internal validation done	External validation done
Chen et al [20]	🟩	🟦	🟩	🟩	🟩
Nguyen et al [32]	🟩	🟩	🟩	🟩	🟩
Renda et al [33]	🟦	🟦	🟧	🟦	🟩
Roy, Chergui, and Verikoukis [34]	🟦	🟦	🟩	🟦	🟩
Imakura et al [35]	🟦	🟦	🟩	🟦	🟦
Raza et al [36]	🟩	🟩	🟧	🟩	🟩
Ambesange, Annappa, and Koolagudi [37]	🟩	🟦	🟩	🟩	🟦
Arthi et al [38]	🟩	🟦	🟧	🟩	🟩
Agbley et al [39]	🟩	🟦	🟧	🟩	🟩
Khan et al [40]	🟩	🟩	🟧	🟩	🟩
Mu et al [41]	🟦	🟦	🟩	🟩	🟦
Kim et al [42]	🟩	🟦	🟩	🟩	🟩
El Houda et al [43]	🟦	🟦	🟧	🟦	🟩
Bárcena et al [44]	🟩	🟩	🟧	🟩	🟦
Haffar, Sánchez, and Domingo-Ferrer [45]	🟩	🟦	🟩	🟩	🟩
Saad, Brik, and Ksentini [46]	🟦	🟦	🟧	🟦	🟩
Huong et al [47]	🟩	🟦	🟧	🟩	🟩
Singh and Adhikari [48]	🟩	🟩	🟧	🟦	🟩
Nikul et al	🟩	🟦	🟧	🟦	🟦
Salim, Turnbull, and Moustafa [50]	🟦	🟦	🟩	🟩	🟩
Patel et al [51]	🟩	🟦	🟧	🟩	🟩
Ren et al [52]	🟩	🟩	🟧	🟩	🟩
Wu et al [53]	🟩	🟦	🟩	🟦	🟩
Wang and Zhang [54]	🟩	🟦	🟧	🟩	🟦
Zheng et al [55]	🟩	🟩	🟧	🟩	🟦
Li et al [56]	🟦	🟦	🟩	🟩	🟩

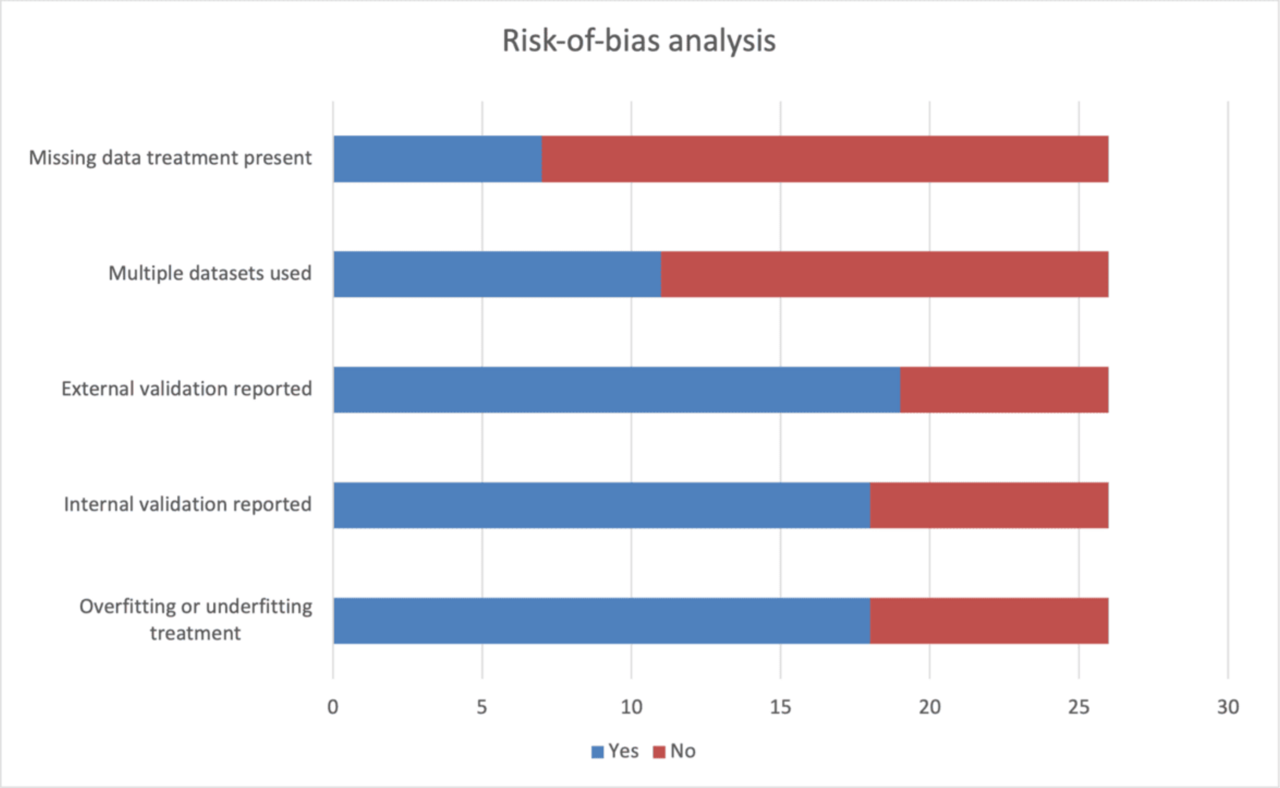
Key	
🟩	Yes
🟧	No
🟦	Unclear

All studies used ML evaluation techniques such as precision, recall, accuracy, F_1-score, mean squared error, mean absolute error, R^2, area under the receiver operating characteristic curve, and the Kolmogorov-Smirnov test. A total of 69% (18/26) of the studies used internal validation techniques (train-test validation split or k-fold cross-validation), with 31% (8/26) of the studies reporting no clear internal validation. Most of the studies (15/26, 58%) had a low risk of bias for their model training, although the lack of missing data training was a key concern.

Risk-of-Bias Analysis Across Studies

The risk of bias across studies was evaluated on the use of multiple datasets and the use of external ML validation techniques such as benchmarking against state-of-the-art models. A total of 73% (19/26) of the studies performed external validation. In total, 27% (7/26) of the studies lacked external validation. Only 42% (11/26) of the studies used multiple datasets, increasing the risk of bias (Figure 6).

Figure 6. Risk-of-bias analysis showing different bias evaluation methods.



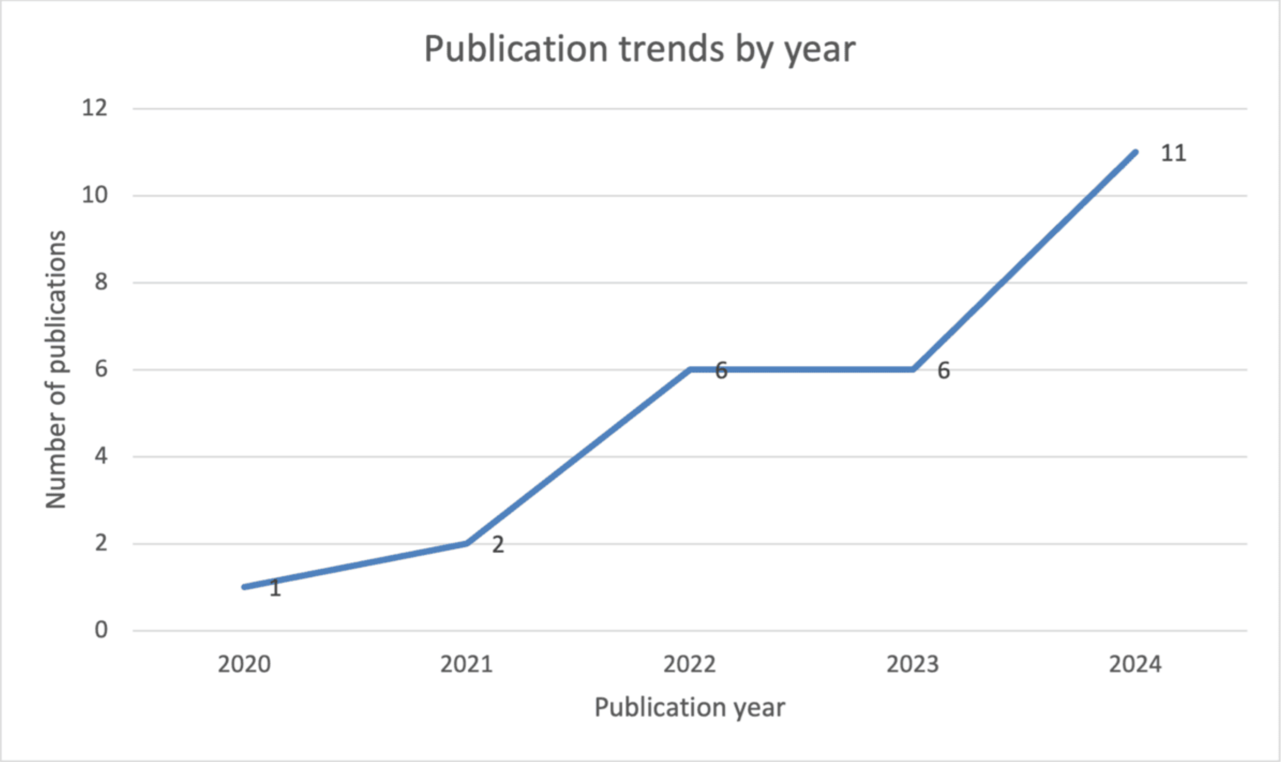
Results

The selection of the articles is illustrated in Figure 4. The results regarding the RQs are presented in the following sections (Multimedia Appendix 4).

RQ Category 1: Trends and Contributions

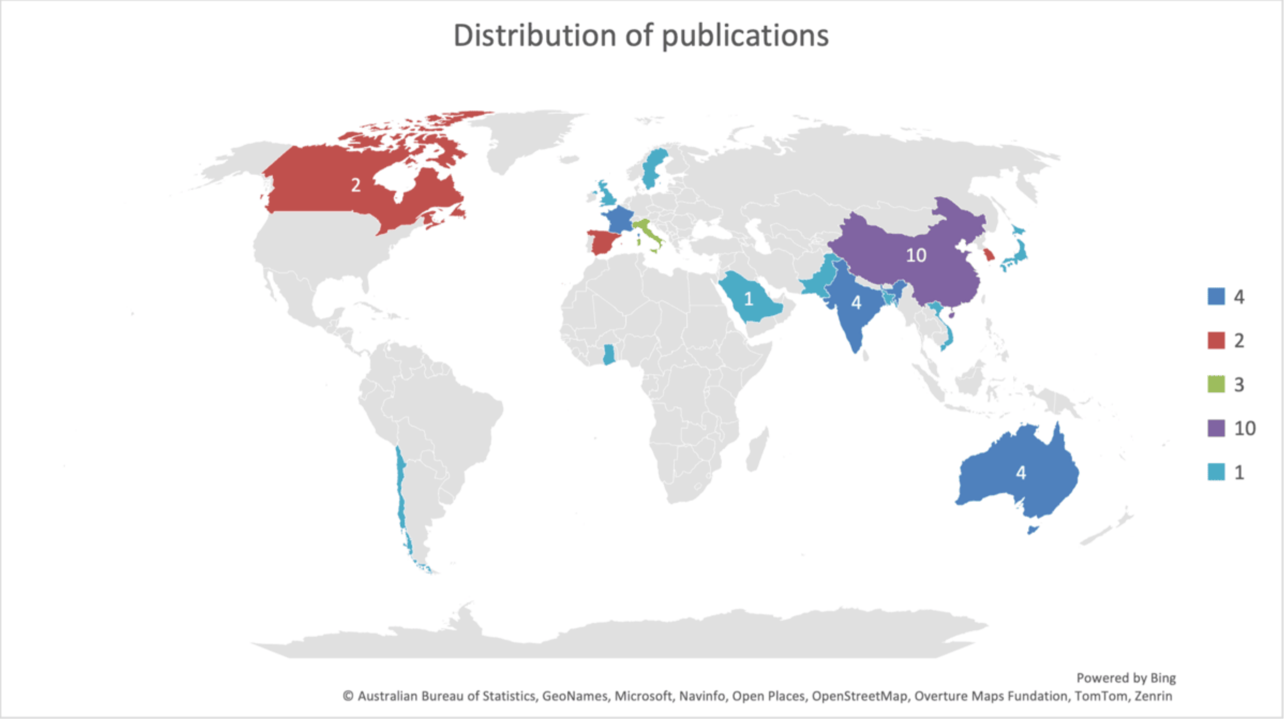
We analyzed the publication trends in explainable FL. While FL emerged in 2016, the first article on XAI for FL was published in 2020(1 publication). The number of articles showed consistent annual growth, culminating in 11 studies in 2024 (Figure 7), which represents the current peak and nearly half (11/26, 42%) of the included studies. The trajectory showed increased interest in this research area despite the low number of total publications (N=26 studies), indicating significant opportunities for future research.

Figure 7. Publication trends for the selected studies by year.



Our analysis of author affiliation revealed a pronounced geographical imbalance, with Asian and European institutions dominating. In contrast, African and South American institutions remained significantly underrepresented, a critical gap given Africa's potential to benefit from privacy-preserving ML solutions amidst resource constraints. Figure 8 shows the authors affiliation by continent were Asia (23), Europe (11), Australia (4), North America (1), South America (1) and Africa (1).

Figure 8. Author affiliation by country for the selected studies (created using the Bing Maps integration in Microsoft Excel [57], which is published under limited license per the Microsoft Bing Maps Terms of Use [58]).



Powered by Bing

© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, Open Places, OpenStreetMap, Overture Maps Fundation, TomTom, Zenrin

Despite the African continent having huge potential for rich, diverse, and high-volume data that can be used in ML research, collating and accessing the distributed data (stored in geographically sparse locations or in different institutions, and also in different formats) still poses a challenge. Lack of a computing backbone—including internet connectivity and cloud computing—further leads to data being sourced from high-income countries [59]. Moreover, data scarcity and the lack of proper infrastructure have been highlighted by Fabila et al [60] and Nieto-Mora et al [61] as limiting the research in data-rich diverse areas such as Africa.

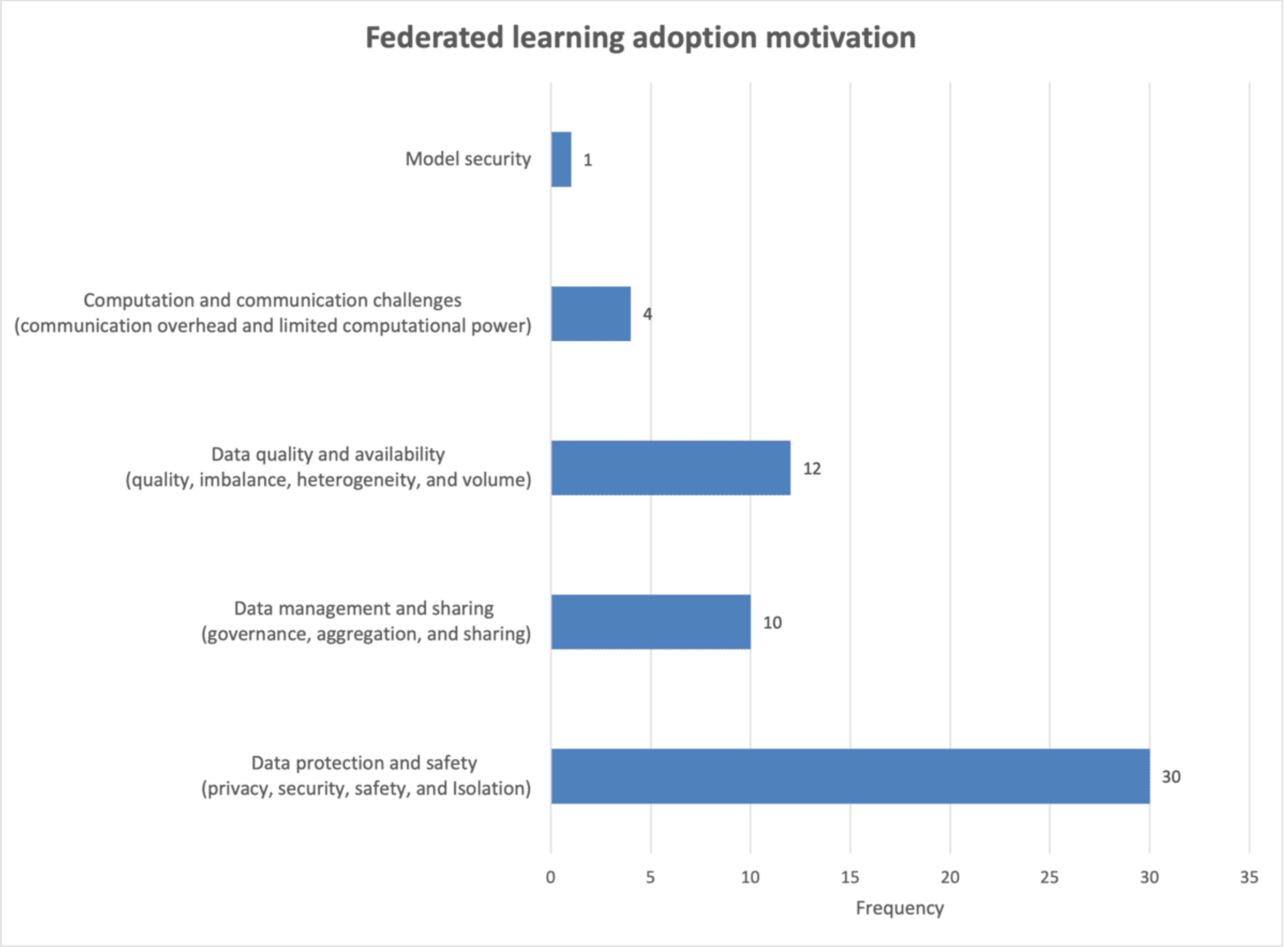
Two dominant approaches for achieving explainability in FL systems emerged: those that are intrinsically explainable (ante hoc) [20,32-35] and those that use a surrogate model for explainability (post hoc) [36-53]. In total, 8% (2/26) of the studies [54,55] could not be properly categorized and were classified as "Unspecified."

RQ Category 2: Application Areas

Overview

The motivations for adoption of FL were analyzed. They were categorized into model security, computation and communication challenges, data quality and availability, data management and sharing, and data protection and safety. The results are shown in Figure 9. The main motivation was data management and sharing, followed by data quality and availability.

Figure 9. Frequency of federated learning adoption motivations.



Application Area and Type of FL Used

The application area and type of FL applied were assessed, and the results are summarized in Table 1. The application area with the highest number of studies was health with 27% (7/26). Networking and finance followed closely with 23% (6/26) and 15% (4/26) of the studies, respectively. Fault detection encompassed 8% (2/26) of the studies, and agriculture, space exploration, urban planning, and social media encompassed 4% (1/26) of the studies each.

Table 1. Summary of the studies based on application area, type, and category of federated learning (FL).

Application area and type of FL	Centralized FL	Studies
Health		
Transfer learning	Yes	[36,37]
Horizontal FL	Yes	[32,38-40]
Vertical FL	Yes	[41]
Space exploration		
Horizontal FL	—[a]	[42]
Networking		
Horizontal FL	Yes	[33,43-46]
Vertical FL	Yes	[34]
Finance		
Vertical FL	Yes	[20,42,55]
Horizontal FL	Yes	[35]
Fault detection		
Horizontal FL	Yes	[47,54]
Agriculture		
Horizontal FL	Yes	[48]
Urban planning		
Vertical FL	Yes	[49]
Social media		
Horizontal FL	No	[50]
Manufacturing		
Horizontal FL	Yes	[51]
Energy		
Horizontal FL	Yes	[52]
Generic		
Vertical FL	Yes	[53,56]

[a] Not applicable.

HFL (17/26, 65% of the studies) was the major type of FL used, with VFL and TL reported in 31% (8/26) and 8% (2/26) of the studies, respectively.

RQ Category 3: Model Explainability

The selected studies were reviewed for their approach to model explainability, which is essential to building trust in predictions. In FL, understanding model outputs helps assess their reliability and identify the need for adjustments or improvements.

XAI Techniques

Overview

XAI, first introduced by the Defense Advanced Research Projects Agency in 2015, helps experts understand how ML models arrive at their decisions, thereby increasing trust in the outputs. XAI techniques can be categorized as either global or local depending on the level of explainability. Global XAI techniques offer a broad view of the model's behavior by highlighting important features. Local XAI techniques focus on explaining individual predictions.

XAI techniques also differ based on whether they are intrinsic to the model (ante hoc or white box), such as decision trees, or applied after training (post hoc), such as LIME [7], which uses simpler models to explain complex ones.

Additionally, some model explainers are model agnostic and can be applied to a wide group of ML models, whereas others are model specific and tailored to particular algorithms, offering deeper insights but requiring more expertise. We provide a brief overview of the techniques in the following sections.

LIME Technique

LIME [7] is a popular model-agnostic explainer that uses a simple surrogate model, typically a sparse linear model, trained on locally perturbed data to approximate and explain the individual predictions of a complex model. While it is widely adopted, LIME's effectiveness depends on the quality of the surrogate fit, and its sampling process introduces uncertainty, resulting in nondeterministic and potentially inconsistent explanations for the same input [62].

SHAP Technique

SHAP [8] is a local and global explainer that is based on game theory. SHAP explains a prediction of each instance by computing the contribution of each feature to the prediction. SHAP uses additive contribution to compute a fair value for each feature by computing the contribution of each feature to the final model outcome to understand the importance of each feature. The SHAP explanation is shown in the following equation, where *g* is the explanation model, *x'* is the coalition vector, *M* is the maximum coalition size, and is the feature attribution for feature *i*:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i$$

Gradient-Weighted Class Activation Mapping

Gradient-Weighted Class Activation Mapping [63] is an explainer that uses the spatial information naturally retained in the last convolutional layer. This is a model-agnostic post hoc explainer that works with different classes of convolutional neural networks. It is a visualization technique that generates heat maps that highlight the important regions of the image that contribute to the model's prediction.

RuleFit

The RuleFit algorithm is a method to generate a model that combines rules and linear regression. First posited by Friedman and Popescu [64] in 2008, RuleFit develops interpretable models that can predict an outcome based on various features. A set of rules is generated from a dataset and then fit into a model using the L1-regularized (least absolute shrinkage and selection operator) regression. The simpler linear models are interpretable like "normal" linear models [65].

Partial Dependence Plot

Partial dependence plot (PDP) [66] is an explainer that shows the marginal effect of 1 or 2 features on the predicted outcome of an ML model. It is a post hoc model-agnostic explainer. One or 2 features are selected, and their changes are mapped by changing the values to see their impact on the predicted outcome. The PDP highlights the relationship between the target and the feature as linear, monotonic, or more complex [65]. A newer variant of PDP is called incremental PDP [67], which expands the working of PDP by considering time-dependent effects in nonstationary learning environments. This newer approach considers how the model's reasoning changes over time while considering the effects of concept drift.

Integrated Gradients

Integrated gradients [68] is an axiomatic-based local explainer that attributes the importance value of each input feature of an ML model based on the gradients of the model outputs with reference to the input.

Causal Models

Causal models [69] use counterfactual reasoning to explain the cause-effect explanations of a particular model. A counterfactual explanation for a prediction is a description of the smallest change to an input feature that will alter the prediction to a predefined output [65]. Counterfactual explanations describe the causes in the form of "if X had not occurred, then Y would not be the result." The computation of counterfactual explanations is done by comparing the causal chain paths of the actions not taken by the model [62].

Anchors

Anchors [70] are a model-agnostic way of explaining the workings of complex (black-box) models through the use of high-precision rules. Anchors use perturbations to generate the local explanations, but instead of using surrogate models, the explanations are provided using if-then rules that are easy to understand. The if-then rules are called anchors. A rule "anchors" the prediction if changes in the other feature values do not alter the prediction made [65].

Deep Taylor Decomposition

Deep Taylor decomposition [71] is an approach for explaining neural networks by decomposing the output of a model into contributions from individual input features. It redistributes the output to the input variables layer by layer. The approach relies on Taylor expansion to determine the relative contributions of the layers. The final relevance scores at the input layer reveal which input features were the most influential in the prediction.

Layerwise Relevance Propagation

Layerwise relevance propagation (LRP) [72] is a technique for explaining predictions made by neural network models. LRP identifies the input features that contributed the most to the decision made by the model. LRP relies on deep Taylor decomposition and works by tracing the prediction backward through the network using backward propagation while assigning relevance scores to each input feature [62].

Prediction Difference Analysis

Prediction difference analysis [73] generates explanations for neural networks by comparing the model's prediction when a specific feature is present with the prediction of the model when that feature is absent. The comparison allows for measurement of the feature's impact on the final model's prediction. Each feature is removed (knocked out), and a relevance score is assigned to them based on their impact [62].

Testing With Concept Activation Vectors

Testing with concept activation vectors [74] is an approach to generate global explanations for neural networks based on the idea of concept activation vectors. It measures the importance of a concept to a prediction based on the directional sensitivity of a concept in the neural network layers. The concept can be anything from color and objects to ideas [65].

Explainable Graph Neural Networks

Explainable graph neural networks [75] are model-level explainers that show how graph neural networks make decisions. Explainable graph neural networks use reinforcement learning to build a new graph stepwise, which the original graph neural network can classify as a certain label, for example, "spam." The new (generated) graph acts as an example for what the model has learned.

Explainable FL

XAI can be applied to FL environments to explain the workings of ML models.

Explainable FL Techniques Used

This study aimed to explore the types of XAI models used in FL (first question in RQ category 3) and their classification (second question in RQ category 3). Most studies (19/26, 73%) applied existing XAI techniques, especially those originally developed for centralized ML such as LIME [7] and SHAP [8]. A few novel methods such as vertical decision tree ensembles [20] were specifically developed for federated settings. Most reviewed studies (23/26, 89%) used post hoc explainability methods, followed by intrinsically explainable models (5/26, 19%). In total, 8% (2/26) of the studies could not be categorized. Most of the techniques were model agnostic, highlighting the adaptability and widespread use of tools such as LIME in FL environments. Table 2 summarizes the various categorizations of XAI approaches as applied in FL.

Table 2. Summary of categorization of explainable artificial intelligence approaches in federated learning, application areas, and performance metrics used.

Approach and model or algorithm	Type (model agnostic or model specific)	Studies	Application area	Performance metrics
Post hoc				
Grad-CAM[a]	Model agnostic	[36,37,39,51,56]	Health care [36,37,39], manufacturing [51], and generic [56]	Accuracy (all studies), precision [36], recall [36,39], and F_1-score [36,39]
Falcon-INP[b]	Model agnostic	[53]	Generic	Accuracy, precision, and MSE[c]
RuleFit	Model agnostic	[43,46]	Networking	Accuracy, F_1-score [43], and PDP[d] and percentage of feature impact [46]
SHAP[e]	Model agnostic	[43,46-50,52,54]	Networking [50], fault detection [47], agriculture [48], urban planning [49], social media [50], and energy [52]	Accuracy [43,47,49,50,52,54], F_1-score [43,47,50], PDP [46], precision [47,50], recall [47,50], RMSE[f] [48], MAE[g] [48], and loss [49]
LIME[h]	Model agnostic	[38,40,46,49,51]	Health care [38,40], networking [46], urban planning [49], and manufacturing [51]	Accuracy [38,40,49,51], F_1-score [38,40], precision [38,40], recall [38,40], and PDP [46]
PDP	Model agnostic	[46]	Networking	—[i]
Causal models	Model agnostic	[41]	Health care	Accuracy
CPA[j] Net	Model specific	[42]	Space exploration	Maximum input sensitivity analysis
Random decision forest	Model agnostic	[45]	Networking	Accuracy
Rule based	Unspecified	[44]	Networking	MSE and R^2
Ante hoc				
Vertical decision tree ensembles	Model specific	[20]	Finance	AUC[k] and KS[l] curve analysis
Decision trees	Model specific	[33,35]	Networking [33] and finance [35]	MSE, MAE and R^2 [33], and accuracy [35]
Integrated gradients	Model agnostic	[32,34]	Health care [32] and networking [34]	AUROC[m] [32], AUPRC[n] [32], and MSE [34]
Unspecified				
Gradient-based method	Unspecified	[55]	Finance	ROC[o] and KS curve analysis
Interpretable adaptive sparse-depth networks	Unspecified	[54]	Fault detection	Accuracy

[a] Grad-CAM: Gradient-Weighted Class Activation Mapping.

[b] Falcon-INP: Falcon Interpretability Framework.

[c] MSE: mean squared error.

[d] PDP: partial dependence plot.

[e] SHAP: Shapley Additive Explanations.

[f] RMSE: root mean square error.

[g] MAE: mean absolute error.

[h] LIME: linear interpretable model-agnostic explanations.

[i] Not applicable.

[j] CPA: Cascading Pyramid Attention.

[k] AUC: area under the curve.

[l] KS: Kolmogorov-Smirnov.

[m] AUROC: area under the receiver operating characteristic curve.

[n] AUPRC: area under the precision-recall curve.

[o]ROC: receiver operating characteristic.

Challenges Faced in Explainable FL

Explaining ML models in an FL environment presents unique challenges typically not encountered in centralized setups, especially in real-world scenarios. The challenges include data heterogeneity, security and privacy, communication costs and resource constraints, and scalability.

Data Heterogeneity

In centralized ML, data from multiple sources are combined into a single dataset, allowing explainability models to analyze a unified, consistent data distribution. In contrast, FL involves data from different, often heterogenous sources that follow different distributions, resulting in non–independently and identically distributed (IID) data [76]. Non-IID data are common in FL and are characterized by skewed class distributions and varying data volumes across clients [76]. This variability challenges explainability as the explainer model must handle randomly polled clients with diverse and uneven data, complicating interpretation.

Security and Privacy

FL was developed to enable ML model training while preserving data privacy, addressing strict data protection regulations. Unlike centralized ML, where XAI techniques risk data leaks or reverse engineering by requiring access to training data, FL introduces new challenges such as vulnerability to model poisoning [77]. Moreover, applying explainability in federated environments can raise privacy concerns as explanation methods might inadvertently reveal some attributes of the client data.

Communication Costs and Resource Constraints

FL involves clients sharing model updates via either a centralized or decentralized approach, necessitating continuous and efficient communication. Additionally, the use of perturbation-based explainers such as SHAP adds overheads on client devices due to complex estimation of Shapley values as well as communication costs when sharing the learned perturbations to the central aggregator [78].

Scalability

In non-IID FL setups, randomly polling clients is often ineffective, necessitating smarter client selection strategies that prioritize clients with valuable data for improving the global model [79]. Moreover, increasing the number of clients can lead to communication bottlenecks and strain the aggregation server's resources due to the growing volume of model updates.

Discussion

Summary of Findings

This study aimed to understand the current situation in the XAI field and how it has been applied to the field of FL. This was done through a comprehensive review process of the existing openly accessible primary studies on XAI approaches in federated ML. The role of privacy in the choice of ML model was evident in the studies analyzed. FL has proven to be robust and useful in mitigating privacy concerns to comply with privacy legislation and ensure data integrity within the devices [22].

It is noteworthy that most of the studies (10/26, 39%) did not originate from highly sensitive fields such as health and security, which are arguably fields that could benefit most from explainable federated AI approaches. These fields are traditionally conservative, heavily regulated (eg, HIPAA) [11], and still suffer from trust issues due to the lack of explainability of the models. These fields are highly impactful as the problems defined require complex solutions, which necessitate the use of black-box models. Areas such as health, cybersecurity, finance, education, and autonomous vehicles could invariably benefit from explainable FL as they are heavily reliant on privacy and security. Federated XAI could also be applied in edge devices as this would bring the computation closer to the data source while at the same time enhancing privacy and security [80].

The FTL approach, which can help alleviate the challenge of limited training data [81]—the second reported reason for the use of FL—has also not been used fully. Despite the use of real-world datasets, the implementations assessed largely used the HFL approach, which did not fully account for data heterogeneity [82]. Real-world implementations of these approaches might suffer due to the data and environment not being representative. It would be important for more research to be conducted addressing these challenges.

Implications

There has been a steady increase in the number of studies in the field of FL and XAI. This increase can be mapped from 2016, when FL was first introduced. However, there is still a lot of room for more research to be conducted. The development of explainable FL models can help unlock great potential in the fields of health and security [2], but caution needs to be taken to ensure that the development is not concentrated in specific regions.

Model explainability using state-of-the-art techniques, whether post hoc or intrinsic in nature, has been proven to work well. Several novel explainability techniques that can work well in FL environments, such as those in the studies by Corcuera Bárcena et al [44] and Wang and Zhang [54], highlight the potential for improvement of existing explainability techniques and approaches and development of more robust novel techniques that can perform better in the federated environments. This also offers fertile research potential for experimentation with more real-world data and techniques such as TL.

More research needs to be conducted to mitigate the challenges faced by explainable FL. There is a need to develop models that are scalable and can operate in real-world FL settings where data are non-IID. There is also a need for robust systems that can operate more efficiently when generating the explanations to make them useful for personalized explainable FL. This would help unlock an even greater potential for trustworthy AI.

Limitations

This review was limited to 26 studies. The novelty of the 2 areas—XAI and FL—meant that a lot of studies (including most studies from the initial total of 1933 identified in the databases) were not eligible for review. Moreover, the strict requirement for primary research and not review papers, coupled with the need for accessible documents, meant that the papers reviewed were limited in nature.

Conclusions

This study attempted to analyze the existing landscape and provide an overview of the approaches that could be used in implementing XAI in FL. This review was conducted based on the RQs posited, and 26 studies that fit the criteria were assessed. One of the key findings was that, despite the need for explainability in critical areas, there is limited research that has been conducted. More research in these critical areas needs to be conducted to develop more novel approaches that mitigate the challenges. FL remains a useful approach to model development in cases in which privacy is important and limited data exist. This study highlights the potential areas that can be explored by future researchers.

Acknowledgments

The authors thank D Benyoucef for the help in providing access to his laboratory and support in providing a conducive work environment as part of the German Academic Exchange Service research stay at Furtwangen University, where this study was undertaken.

Funding

This research was supported by a research grant from the German Academic Exchange Service under the In-Country/In-Region scholarship program Kenya 2022 (57607744; category for personal reference number 91843031).

Authors' Contributions

TT contributed to the conceptualization of this systematic review (Introduction, Methods, Results, Discussion, and Conclusions sections). He also contributed to the original draft's preparation and validation. BS contributed to the review and editing of the manuscript, as well as supervision and validation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search string formulation.
[DOCX File , 37 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search string results.
[DOCX File , 39 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data extraction.
[DOCX File , 37 KB-Multimedia Appendix 3]

Multimedia Appendix 4

PRISMA checklist.
[PDF File (Adobe PDF File), 81 KB-Multimedia Appendix 4]

References

1. Nichols JA, Herbert Chan HW, Baker MA. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev. Feb 4, 2019;11(1):111-118. [FREE Full text] [doi: 10.1007/s12551-018-0449-9] [Medline: 30182201]
2. Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl Sci. Jan 27, 2022;12(3):1353. [doi: 10.3390/app12031353]
3. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160. [FREE Full text] [doi: 10.1007/s42979-021-00592-x] [Medline: 33778771]
4. Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Decis Anal J. Jun 2023;7:100230. [doi: 10.1016/j.dajour.2023.100230]
5. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl Based Syst. Mar 05, 2023;263:110273. [doi: 10.1016/j.knosys.2023.110273]

6. Vilone G, Longo L. Explainable artificial intelligence: a systematic review. arXiv. Preprint posted online on May 29, 2020. [FREE Full text] [doi: 10.48550/arXiv.2006.00093]
7. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. arXiv. Preprint posted online on February 16, 2016. [FREE Full text] [doi: 10.1145/2939672.2939778]
8. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [FREE Full text] [doi: 10.48550/arXiv.1705.07874]
9. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. arXiv. Preprint posted online on November 6, 2019. [FREE Full text] [doi: 10.1145/3375627.3375830]
10. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Union. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2021-10-20]
11. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html [accessed 2025-12-26]
12. Data Protection Act, 2019. National Council for Law Reporting with the Authority of the Attorney-General. 2019. URL: https://www.kentrade.go.ke/wp-content/uploads/2022/09/Data-Protection-Act-1.pdf [accessed 2026-01-08]
13. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on February 17, 2016. [FREE Full text] [doi: 10.48550/arXiv.1602.05629]
14. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. arXiv. Preprint posted online on February 4, 2019. [FREE Full text] [doi: 10.48550/arXiv.1902.01046]
15. Bhatia L, Samet S. A decentralized data evaluation framework in federated learning. Blockchain Res Appl. Dec 2023;4(4):100152. [doi: 10.1016/j.bcra.2023.100152]
16. Sun T, Li D, Wang B. Decentralized federated averaging. arXiv. Preprint posted online on April 23, 2021. [FREE Full text] [doi: 10.48550/arXiv.2104.11375]
17. Iqbal Z, Chan HY. Concepts, key challenges and open problems of federated learning. Int J Eng. Jul 2021;34(7):1667-1683. [FREE Full text] [doi: 10.5829/IJE.2021.34.07A.11]
18. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol. Jan 28, 2019;10(2):1-19. [doi: 10.1145/3298981]
19. Yuan L, Wang Z, Sun L, Yu PS, Brinton CG. Decentralized federated learning: a survey and perspective. arXiv. Preprint posted online on June 2, 2023. [FREE Full text] [doi: 10.48550/arXiv.2306.01603]
20. Chen X, Zhou S, Guan B, Yang K, Fao H, Wang H. Fed-EINI: an efficient and interpretable inference framework for decision tree ensembles in vertical federated learning. In: Proceedings of the 2021 IEEE International Conference on Big Data. 2021. Presented at: Big Data '21; December 15-18, 2021; Orlando, FL. URL: https://ieeexplore.ieee.org/document/9671749 [doi: 10.1109/bigdata52589.2021.9671749]
21. Martínez Beltrán ET, Pérez MQ, Sánchez PM, Bernal SL, Bovet G, Pérez MG, et al. Decentralized federated learning: fundamentals, state of the art, frameworks, trends, and challenges. IEEE Commun Surv Tutorials. 2023;25(4):2983-3013. [doi: 10.1109/comst.2023.3315746]
22. Wen J, Zhang Z, Lan Y, Cui Z, Cai J, Zhang W. A survey on federated learning: challenges and applications. Int J Mach Learn Cybern. Nov 11, 2022;14(2):513-535. [FREE Full text] [doi: 10.1007/s13042-022-01647-y] [Medline: 36407495]
23. Khan M, Glavin FG, Nickles M. Federated learning as a privacy solution - an overview. Procedia Comput Sci. 2023;217:316-325. [doi: 10.1016/j.procs.2022.12.227]
24. Singh J, Goyal SB, Kumar RK, Kumar N, Singh SS. Applied Data Science and Smart Systems. London, UK. CRC Press; 2024.
25. Aggarwal M, Khullar V, Rani S, Prola TA, Bhattacharjee SB, Shawon SM, et al. Federated learning on internet of things: extensive and systematic review. Comput Mater Con. May 15, 2024;79(2):1795-1834. [doi: 10.32604/cmc.2024.049846]
26. Xiao Y, Watson M. Guidance on conducting a systematic literature review. J Plan Educ Res. Aug 28, 2017;39(1):93-112. [doi: 10.1177/0739456x17723971]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Keele University and University of Durham. Jul 09, 2007. URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2025-12-26]
29. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. Syst Rev. Sep 29, 2016;5(1):164. [FREE Full text] [doi: 10.1186/s13643-016-0337-y] [Medline: 27686611]

30. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid. Mar 27, 2018;7:8. [FREE Full text] [doi: 10.1186/s13750-018-0124-4]
31. Wang Y, Singh L. Analyzing the impact of missing values and selection bias on fairness. Int J Data Sci Anal. May 31, 2021;12(2):101-119. [doi: 10.1007/s41060-021-00259-z]
32. Nguyen TN, Yang HJ, Kho BG, Kang SR, Kim SH. Explainable deep contrastive federated learning system for early prediction of clinical status in intensive care unit. IEEE Access. Aug 23, 2024;12:117176-117202. [doi: 10.1109/access.2024.3447759]
33. Renda A, Ducange P, Marcelloni F, Sabella D, Filippou MC, Nardini G, et al. Federated learning of explainable AI models in 6G systems: towards secure and automated vehicle networking. Information. Aug 20, 2022;13(8):395. [doi: 10.3390/info13080395]
34. Roy S, Chergui H, Verikoukis C. Toward bridging the FL performance-explainability tradeoff: a trustworthy 6G RAN slicing use-case. arXiv. Preprint posted online on September 19, 2024. [doi: 10.1109/tvt.2024.3364363]
35. Imakura A, Inaba H, Okada Y, Sakurai T. Interpretable collaborative data analysis on distributed data. Expert Syst Appl. Sep 1, 2021;177:114891. [doi: 10.1016/j.eswa.2021.114891]
36. Raza A, Tran KP, Koehl L, Li S. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. Knowl Based Syst. Jan 2022;236:107763. [doi: 10.1016/j.knosys.2021.107763]
37. Ambesange S, Annappa B, Koolagudi SG. Simulating federated transfer learning for lung segmentation using modified UNet model. Procedia Comput Sci. 2023;218:1485-1496. [FREE Full text] [doi: 10.1016/j.procs.2023.01.127] [Medline: 36743787]
38. Arthi NT, Mubin KE, Rahman J, Rafi GM, Sheja TT, Reza MT. Decentralized federated learning and deep learning leveraging XAI-based approach to classify colorectal cancer. In: Proceedings of the 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering. 2022. Presented at: CSDE '22; December 18-20, 2022; Gold Coast, Australia. [doi: 10.1109/csde56538.2022.10089344]
39. Agbley BL, Li JP, Haq AU, Bankas EK, Mawuli CB, Ahmad S, et al. Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things. IEEE J Biomed Health Inform. Jun 2024;28(6):3389-3400. [doi: 10.1109/JBHI.2023.3256974] [Medline: 37028353]
40. Khan IA, Razzak I, Pi D, Zia U, Kamal S, Hussain Y. A novel collaborative SRU network with dynamic behaviour aggregation, reduced communication overhead and explainable features. IEEE J Biomed Health Inform. Jun 2024;28(6):3228-3235. [doi: 10.1109/JBHI.2024.3352013]
41. Mu J, Kadoch M, Yuan T, Lv W, Liu Q, Li B. Explainable federated medical image analysis through causal learning and blockchain. IEEE J Biomed Health Inform. Jun 2024;28(6):3206-3218. [doi: 10.1109/JBHI.2024.3375894] [Medline: 38470597]
42. Kim T, Jeon M, Lee C, Kim J, Ko G, Kim JY, et al. Federated onboard-ground station computing with weakly supervised cascading pyramid attention network for satellite image analysis. IEEE Access. 2022;10:117315-117333. [doi: 10.1109/ACCESS.2022.3219879] [Medline: 38470597]
43. El Houda ZA, Moudoud H, Brik B, Khoukhi L. Securing federated learning through blockchain and explainable AI for robust intrusion detection in IoT networks. In: Proceedings of the IEEE INFOCOM 2023 Conference on Computer Communications Workshops. 2023. Presented at: INFOCOM WKSHPS '23; May 20, 2023; Hoboken, NJ. [doi: 10.1109/INFOCOMWKSHPS57453.2023.10225769]
44. Corcuera Bárcena JL, Ducange P, Marcelloni F, Nardini G, Noferi A, Renda A, et al. Enabling federated learning of explainable AI models within beyond-5G/6G networks. Comput Commun. Oct 2023;210:356-375. [doi: 10.1016/j.comcom.2023.07.039]
45. Haffar R, Sánchez D, Domingo-Ferrer J. Explaining predictions and attacks in federated learning via random forests. Appl Intell. Apr 13, 2022;53(1):169-185. [doi: 10.1007/s10489-022-03435-1]
46. Saad SB, Brik B, Ksentini A. A trust and explainable federated deep learning framework in zero touch B5G networks. In: Proceedings of the 2022 IEEE Global Communications Conference. 2022. Presented at: GLOBECOM '22; December 4-8, 2022; Rio de Janeiro, Brazil. URL: https://ieeexplore.ieee.org/document/10001371 [doi: 10.1109/globecom48099.2022.10001371]
47. Huong TT, Bac TP, Ha KN, Hoang NV, Hoang NX, Hung NT, et al. Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access. 2022;10:53854-53872. [doi: 10.1109/access.2022.3173288]
48. Singh N, Adhikari M. Real-time paddy field irrigation using feature extraction and federated learning strategy. IEEE Sensors J. Nov 1, 2024;24(21):36159-36166. [doi: 10.1109/jsen.2024.3462496]
49. Patel AN, Srivastava G, Reddy Maddikunta PK, Murugan R, Yenduri G, Reddy Gadekallu T. A trustable federated learning framework for rapid fire smoke detection at the edge in smart home environments. IEEE Internet Things J. Dec 1, 2024;11(23):37708-37717. [doi: 10.1109/jiot.2024.3439228]
50. Salim S, Turnbull B, Moustafa N. A blockchain-enabled explainable federated learning for securing internet-of-things-based social media 3.0 networks. IEEE Trans Comput Soc Syst. Aug 2024;11(4):4681-4697. [doi: 10.1109/tcss.2021.3134463]

51. Patel T, Murugan R, Yenduri G, Jhaveri RH, Snoussi H, Gaber T. Demystifying defects: federated learning and explainable AI for semiconductor fault detection. IEEE Access. 2024;12:116987-117007. [doi: 10.1109/access.2024.3425226]
52. Ren C, Dong ZY, Yu H, Xu M, Xiong Z, Niyato D. ESQFL: digital twin-driven explainable and secured quantum federated learning for voltage stability assessment in smart grids. IEEE J Sel Top Signal Process. Jul 2024;18(5):964-978. [doi: 10.1109/jstsp.2024.3485878]
53. Wu Y, Xing N, Chen G, Dinh TT, Luo Z, Ooi BC, et al. Falcon: a privacy-preserving and interpretable vertical federated learning system. Proc VLDB Endow. Aug 08, 2023;16(10):2471-2484. [doi: 10.14778/3603581.3603588]
54. Wang S, Zhang Y. Multi-level federated network based on interpretable indicators for ship rolling bearing fault diagnosis. JMSE. May 28, 2022;10(6):743. [doi: 10.3390/jmse10060743]
55. Zheng F, Erihe, Li K, Tian J, Xiang X. A vertical federated learning method for interpretable scorecard and its application in credit scoring. arXiv. Preprint posted online on September 14, 2020. [FREE Full text] [doi: 10.48550/arXiv.2009.06218]
56. Li Z, Chen H, Ni Z, Gao Y, Lou W. Towards adaptive privacy protection for interpretable federated learning. IEEE Trans Mobile Comput. Dec 2024;23(12):14471-14483. [doi: 10.1109/tmc.2024.3443862]
57. Bing Maps. Microsoft | Marketplace. URL: https://marketplace.microsoft.com/en-au/product/office/WA102957661?tab=Overview [accessed 2026-01-20]
58. Microsoft Bing Maps Platform APIs Terms Of Use. Bing maps | Dev Center. 2024. URL: https://www.bingmapsportal.com/terms [accessed 2026-01-20]
59. Ezugwu AE, Oyelade ON, Ikotun AM, Agushaka JO, Ho YS. Machine learning research trends in Africa: a 30 years overview with bibliometric analysis review. Arch Comput Methods Eng. Apr 29, 2023;30(7):1-31. [FREE Full text] [doi: 10.1007/s11831-023-09930-z] [Medline: 37359741]
60. Fabila J, Garrucho L, Campello VM, Martín-Isla C, Lekadir K. Federated learning in low-resource settings: a chest imaging study in Africa -- challenges and lessons learned. arXiv. Preprint posted online on May 20, 2025. [FREE Full text] [doi: 10.48550/arXiv.2505.14217]
61. Nieto-Mora DA, Rodríguez-Buritica S, Rodríguez-Marín P, Martínez-Vargaz JD, Isaza-Narváez C. Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. Heliyon. Oct 22, 2023;9(10):e20275. [FREE Full text] [doi: 10.1016/j.heliyon.2023.e20275] [Medline: 37790981]
62. Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W. xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Cham, Switzerland. Springer International Publishing; 2022.
63. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. Oct 11, 2019;128(2):336-359. [doi: 10.1007/s11263-019-01228-7]
64. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. Sep 2008;2(3):916-954. [FREE Full text] [doi: 10.1214/07-AOAS148]
65. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd edition. Munich, Germany. Shroff/Christoph Molnar; 2025.
66. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. Oct 1, 2001;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
67. Muschalik M, Fumagalli F, Jagtani R, Hammer B, Hüllermeier E. iPDP: on partial dependence plots in dynamic modeling scenarios. In: Proceedings of the First World Conference on Explainable Artificial Intelligence. 2023. Presented at: xAI '23; July 26-28, 2023; Lisbon, Portugal. URL: https://link.springer.com/chapter/10.1007/978-3-031-44064-9_11#citeas [doi: 10.1007/978-3-031-44064-9_11]
68. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv. Preprint posted online on March 4, 2017. [FREE Full text] [doi: 10.48550/arXiv.1703.01365]
69. Van Looveren A, Klaise J. Interpretable counterfactual explanations guided by prototypes. arXiv. Preprint posted online on July 13, 2019. [FREE Full text] [doi: 10.1007/978-3-030-86520-7_40]
70. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. 2018. Presented at: AAAI '18/IAAI '18/EAAI '18; February 2-7, 2018; New Orleans, LA. URL: https://dl.acm.org/doi/abs/10.5555/3504035.3504222 [doi: 10.1609/aaai.v32i1.11491]
71. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. May 2017;65:211-222. [doi: 10.1016/j.patcog.2016.11.008]
72. Binder A, Montavon G, Lapuschkin S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa AE, Masulli P, Pons Rivero AJ, editors. Artificial Neural Networks and Machine Learning – ICANN 2016. Cham, Switzerland. Springer International Publishing; 2016:63-71.
73. Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: prediction difference analysis. arXiv. Preprint posted online on February 15, 2017. [FREE Full text] [doi: 10.48550/arXiv.1702.04595]

74. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). arXiv. Preprint posted online on November 30, 2017. [FREE Full text] [doi: 10.48550/arXiv.1711.11279]
75. Yuan H, Tang J, Hu X, Ji S. XGNN: towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020. Presented at: KDD '20; July 6-10, 2020; Virtual Event. [doi: 10.1145/3394486.3403085]
76. Daole M, Ducange P, Marcelloni F, Renda A. Trustworthy AI in heterogeneous settings: federated learning of explainable classifiers. In: Proceedings of the 2024 IEEE International Conference on Fuzzy Systems. 2024. Presented at: FUZZ-IEEE '24; June 30-July 5, 2024; Yokohama, Japan. [doi: 10.1109/fuzz-ieee60900.2024.10612109]
77. Hulsen T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI. Aug 10, 2023;4(3):652-666. [doi: 10.3390/ai4030034]
78. Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of XAI models in healthcare: a case study on Parkinson’s disease. Cogn Comput. Aug 28, 2024;16:3051-3076. [doi: 10.1007/s12559-024-10332-x]
79. Chiarani M, Roy S, Verikoukis C, Granelli F. XAI-driven client selection for federated learning in scalable 6G network slicing. arXiv. Preprint posted online on March 16, 2025. [FREE Full text] [doi: 10.1109/icc52391.2025.11161532]
80. Corcuera Bárcena JL, Daole M, Ducange P, Marcelloni F, Marcelloni A, Schiavo A. Fed-XAI: federated learning of explainable artificial intelligence models. In: Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence. 2022. Presented at: XAI.it '22; November 28-December 2, 2022; Udine, Italy. URL: https://ceur-ws.org/Vol-3277/paper8.pdf [doi: 10.1016/j.softx.2023.101505]
81. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data. 2022;9(1):102. [FREE Full text] [doi: 10.1186/s40537-022-00652-w] [Medline: 36313477]
82. Huang W, Li T, Wang D, Du S, Zhang J, Huang T. Fairness and accuracy in horizontal federated learning. Inf Sci. Apr 2022;589:170-185. [doi: 10.1016/j.ins.2021.12.102]

Abbreviations

AI: artificial intelligence
CFL: centralized federated learning
FL: federated learning
FTL: federated transfer learning
HFL: horizontal federated learning
HIPAA: Health Insurance Portability and Accountability Act
IID: independently and identically distributed
LIME: linear interpretable model-agnostic explanations
LRP: layerwise relevance propagation
ML: machine learning
PDP: partial dependence plot
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
SHAP: Shapley Additive Explanations
TL: transfer learning
VFL: vertical federated learning
XAI: explainable artificial intelligence

Edited by J-L Raisaro; submitted 12.Dec.2024; peer-reviewed by N Goyal, Y Pang; comments to author 26.Jan.2025; revised version received 30.Apr.2025; accepted 11.Oct.2025; published 03.Feb.2026

Please cite as:
Tunduny T, Shibwabo B
Explainable AI Approaches in Federated Learning: Systematic Review
JMIR AI 2026;5:e69985
URL: https://ai.jmir.org/2026/1/e69985
doi: 10.2196/69985
PMID: