applied sciences

MDPI

Review

A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications

Prayitno [1,2], Chi-Ren Shyu [3,4], Karisma Trinanda Putra [1,5], Hsing-Chung Chen [1,6], Yuan-Yu Tsai [7], K. S. M. Tozammel Hossain [3,4], Wei Jiang [3,4] and Zon-Yin Shae [1,*]

1 Department of Computer Science and Information Engineering, Asia University, Taichung City 413, Taiwan; prayitno@polines.ac.id (P.); karisma@ft.umy.ac.id (K.T.P.); cdma2000@asia.edu.tw (H.-C.C.)
2 Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang 50275, Indonesia
3 Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA; shyuc@missouri.edu (C.-R.S.); hossaink@missouri.edu (K.S.M.T.H.); wjiang@missouri.edu (W.J.)
4 Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
5 Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Bantul 55183, Indonesia
6 Department of Medical Research, China Medical University Hospital, China Medical University, Taichung City 404, Taiwan
7 Department of M-Commerce and Multimedia Applications, Asia University, Taichung City 413, Taiwan; yytsai@asia.edu.tw
* Correspondence: zshae1@asia.edu.tw

Citation: Prayitno; Shyu, C.-R.; Putra, K.T.; Chen, H.-C.; Tsai, Y.-Y.; Hossain, K.S.M.T.; Jiang, W.; Shae, Z.-Y. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Appl. Sci.* **2021**, *11*, 11191. https://doi.org/10.3390/app112311191

Academic Editors: Stefano Silvestri and Francesco Gargiulo

Received: 12 November 2021
Accepted: 23 November 2021
Published: 25 November 2021

Abstract: Recent advances in deep learning have shown many successful stories in smart healthcare applications with data-driven insight into improving clinical institutions' quality of care. Excellent deep learning models are heavily data-driven. The more data trained, the more robust and more generalizable the performance of the deep learning model. However, pooling the medical data into centralized storage to train a robust deep learning model faces privacy, ownership, and strict regulation challenges. Federated learning resolves the previous challenges with a shared global deep learning model using a central aggregator server. At the same time, patient data remain with the local party, maintaining data anonymity and security. In this study, first, we provide a comprehensive, up-to-date review of research employing federated learning in healthcare applications. Second, we evaluate a set of recent challenges from a data-centric perspective in federated learning, such as data partitioning characteristics, data distributions, data protection mechanisms, and benchmark datasets. Finally, we point out several potential challenges and future research directions in healthcare applications.

Keywords: federated learning; deep learning; artificial intelligence; healthcare; data privacy-preserving

1. Introduction

Deep learning technology has shown promising results in smart healthcare applications to assist medical diagnosis and treatment based on clinical data. For instance, deep learning assists cancer diagnosis and prediction [1–3], brain tumor segmentation and classification from magnetic resonance image (MRI) [4–6], and text detection of medical laboratory reports [7,8]. Good performance of the deep learning model on smart healthcare applications highly depends on a diverse and vast amount of training data [9]. These training data were obtained from various clinical observations such as biomedical sensors, individual patients, clinical institutions, hospitals, pharmaceutical industries, and health insurance companies. However, acquiring the healthcare data required to develop a deep learning model may be challenging due to fewer patients and pathologies with a low incidence rate available in a single healthcare institution. Furthermore, Zech et al. [10] showed that deep learning models trained with single institutional data are vulnerable to institutional data bias, as shown in Figure 1a. This institutional data bias has been shown to have high accuracy when evaluated on the same

clinical institution's data. However, it does not work well when applied to data from a different institution or even across departments within the same institution. Simultaneously, training deep learning models in a centralized data lake [11], as depicted in Figure 1b, is infeasible because of patient privacy and government regulations related to clinical data. Thus, to increase both the diversity and quantity of training data is through the collaboration of several healthcare institution to create a single deep learning model while maintaining patient privacy and confidentially.

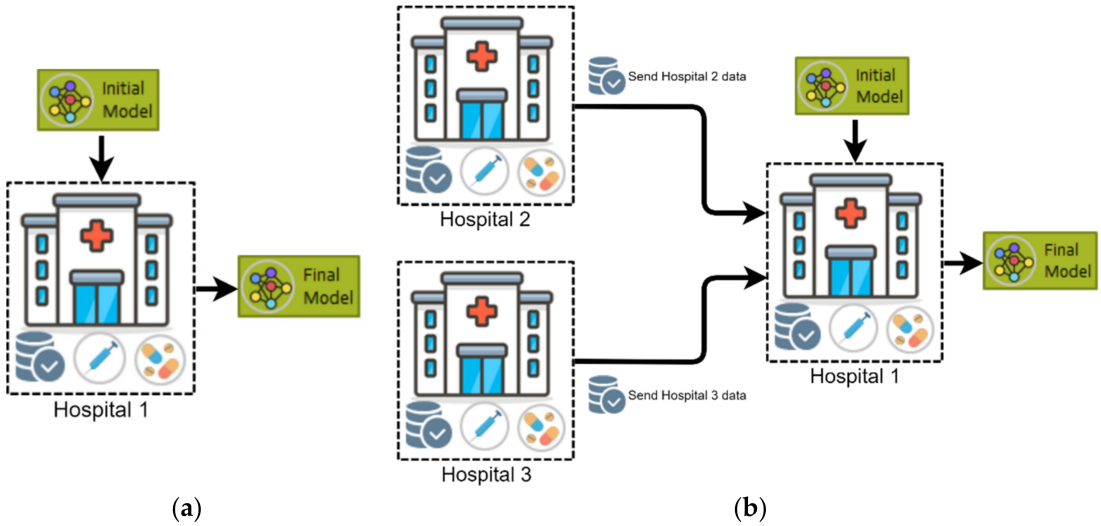


Figure 1. Single-institution and collaborative learning: (**a**) single-institution learning: machine learning model trained and validated with single institution dataset; (**b**) collaborative learning: machine learning model trained and validated with medical data collected from external institutions pooled in a central data lake.

Medical data are usually fragmented due to the complex nature of the medical system and processes. For instance, each medical institution may be able to access the medical data of their patients only. As protected health information (PHI), these medical data are only disclosed strictly regulated by law to third parties. The process of accessing and analyzing medical data is strictly regulated by laws and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) [12]. In addition, with an increasing number of data breaches at healthcare organizations, the prominence of data security and privacy protection has become a global consensus. For instance, in the American Medical Collection Agency (AMCA) recent healthcare data breach, the perpetrators have access to medical data, financial information, and payment details, affecting 11.9 million patients [13]. As a result, many countries around the globe are enacting stricter legislations to protect data security. For example, the General Data Protection Regulation (GDPR) went into effect in 2018 by the European Union to ensure users' privacy while protecting their data [14]. Under this GDPR, business entities must clearly explain why they need user data access and offer them the right to withdraw or delete their data. Business entities violating the regulation would face severe penalties. Many similar actions have taken place in the United States and Taiwan to protect individuals' privacy and security. For instance, Taiwan's Personal Data Protection Act (PDPA) and Cyber Security Management Act, enacted in 2018, prohibit online business entities from leaking or tampering with personal data details that they obtain [15]. This regulation enforces the business activities following the obligations of legal data protection. On the one hand, establishing these regulations will contribute to a more civil society's growth. On the other hand, these regulations introduce new challenges to data transaction and collaboration procedures for multi-institutional collaboration to train a deep learning model.

One recent approach to solving the problem of training a robust deep learning model from federated medical data while preserving patient privacy is federated learning

(FL) [16,17]. This method provides decentralized machine learning model training without transmitting medical data through a coordinated central aggregate server. Medical institutions, working as client nodes, train their deep learning models locally and then periodically forward them to the aggregate server. The central server coordinates and aggregates the local models from each node to create a global model, then distributes the global model to all the other nodes. It is worth noting that the training data are kept private to each node and never transmitted during the training process. Only the model's weight and parameters are transmitted, ensuring that medical data remain confidential. For these reasons, FL mitigates many security concerns because it retains sensitive and private data while enabling multiple medical institutions to work together. FL holds an excellent promise in healthcare applications to improve medical services for both institutions and patients—for instance, predict autism spectrum disorder [18], mortality and intensive care unit (ICU) stay-time prediction [19], wearable healthcare devices [20,21], and brain tumor segmentation [22]. However, FL algorithms face several challenges, mainly due to the properties of medical data, such as:

- **Data partitions:** FL technique aims to solve the limited sample size problem for training a secure collaborative machine learning model by aggregating a group of clients' data. However, choosing a data partition (horizontal or vertical) for FL is essential to solve the limited sample size, limited sample features, or both.
- **Data distribution (statistical challenge):** In developing a machine learning model in a centralized manner, the training data are centrally stored and balanced during training. However, with federated learning, each client generated the training data locally, remained decentralized, and cannot access the other clients' data. Thus, data distribution at one client can differ significantly from others, i.e., nonindependent and identically distributed (non-IID), impacting the performance of the federated learning model [23,24].
- **Privacy and security**: Data privacy and security are critical issues in medical applications. It is impossible to assume all of the clients in FL are reliable because the number of clients expected to participate is potentially thousands or millions. Thus, privacy-preserving mechanisms are needed to protect medical data from untrusted clients or third-party attackers.
- **Benchmark medical dataset:** Medical dataset quantity and quality have often limited the development of a robust solution to the FL algorithm. For various research purposes, the dataset used in FL experiments could vary significantly. For instance, some datasets focus on medical image classification and segmentation performance while others focus on network communication performance. However, the benchmark datasets have not already been compiled, specifically for medical datasets. Thus, a trusted benchmark is necessary to evaluate the performance of the FL that uses multiple medical data sources. Finally, we provide a comprehensive list of relevant medical datasets for future research on this topic.

Due to the ever-changing development in FL, several valuable studies on FL have been published in reputable publications from 2018 to 2021. Therefore, this paper aims to provide a recent review of federated learning in the medical domain. Specifically, this study describes the existing FL techniques related to solving the challenges inherent in medical data together with future research direction on FL for healthcare applications.

This study differs from existing reviews. General descriptions of FL are given in [16,17], while detailed discussions of recent challenges are presented in [25,26], security analysis [27], and personalization techniques [28]. Resumes of FL applications in edge computing [29], wireless networks [30], and healthcare [31,32] also have been published. However, none of the existing studies have explored the impact of medical data properties on the performance of FL in great detail. Moreover, it is necessary to provide a comprehensive overview related to benchmarking the FL in medical data. To fill the gap, this review presents a survey of FL from the perspective of data properties including data partitions, data distribution, data privacy, benchmarking, and its promising applications.

After a brief introduction of FL in this study, the rest of this paper is structured as follows. Section 2 describes the research method to conduct this study. Furthermore, in Section 3, we provide the search results from existing publications. Section 4 discusses our findings in data partition, data distribution properties, data privacy threats and protections, benchmark medical dataset, and open challenges applied in federated learning for medical applications. Finally, we have our paper's conclusion in Section 5.

2. Research Method

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [33] was the research method to guide this study. PRISMA technique is a widely accepted standard for reporting evidence in systematic reviews that health-related organizations and journals have adopted [34]. PRISMA approaches provide several advantages, such as showcasing the review's quality, allowing readers to assess the review's strengths and flaws, replicating review processes, and structuring and formatting the review using PRISMA headings [33]. However, doing a systematic review and thoroughly publishing it may take time. Additionally, it can soon become out of date, thus it must be updated regularly to incorporate all newly published primary material since the project began.

2.1. Formulate Research Questions

We divide the research question into the following research questions.

- **RQ1**: What are the state-of-the-art FL methods in the healthcare area?
- **RQ2**: What are the FL methods proposed by scholars to solve challenging medical applications from a data properties perspective?
- **RQ3**: What are the research gaps and potential future research directions of FL related to medical applications?

The first research question (RQ1) aims to provide a comprehensive and systematic overview of all articles related to FL. Furthermore, RQ1 aims to provide evidence that the healthcare area can benefit by incorporating FL. Additionally, the second research question's (RQ2) motivation is to answer FL medical data settings challenges in FL such as data partition, statistic heterogeneity, and security. Finally, the third research question (RQ3) provides future directions for a researcher in the FL field primarily related to medical data challenges.

2.2. Data Eligibility and Analysis of the Literature

The article selection procedure uses the PRISMA flow diagram [33], as shown in Figure 2, which outlines papers' search, inclusion, and exclusion. There are three steps in the PRISMA flow diagram: identification, screening, and included. Firstly, in the identification step, we performed a comprehensive literature review between 1 January 2018 and 31 June 2021, using PubMed, Web of Science (WoS), Association of Computing Machinery Digital Library (ACM DL), Science Direct, and IEEEXplore digital libraries. We start from 2018 because we are interested in further implementation in the medical area one year after federated learning was proposed in 2017 [16]. The following search phrases were used in general are "Federate learning," and "Healthcare," and "data privacy protection." Because each publication database has its own set of filters for search queries, the specific query terms are specified in Appendix A Table A1. The initial result from digital libraries showed 197 articles satisfying the search criteria. Then, 28 articles were removed due to duplications, ending with 169 articles in the identification step.

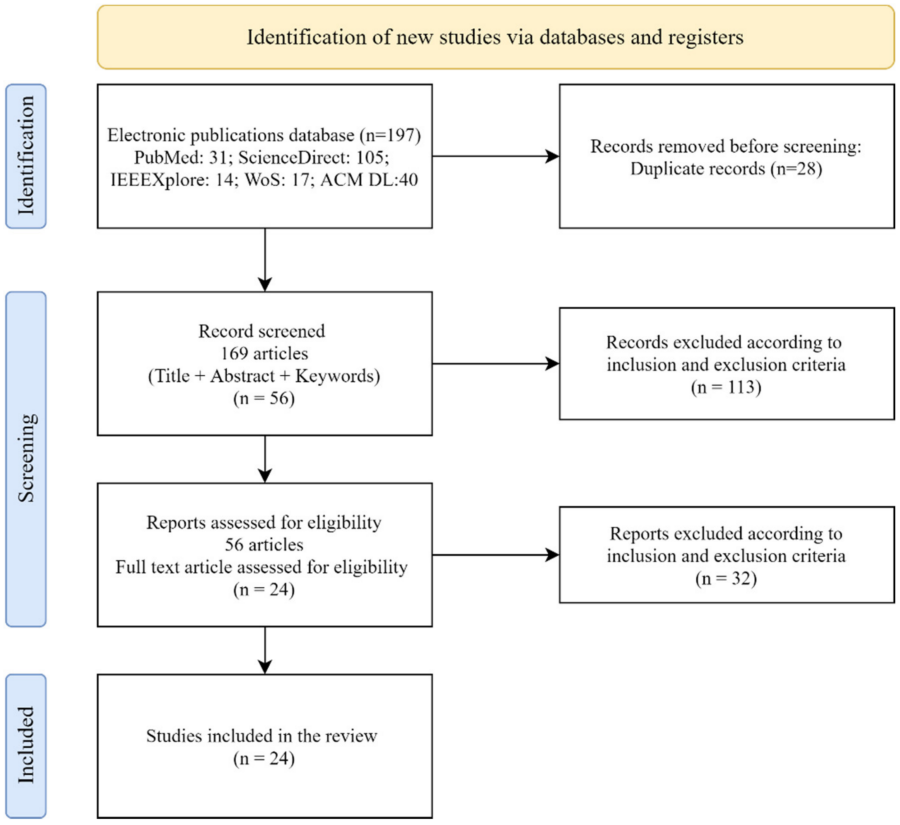


Figure 2. Study selection using PRISMA flow diagram method consisting of identification step, screening step, and included step.

While systematic reviews offer various advantages, they are prone to biases that obscure the study's objective results and should be evaluated cautiously [35]. Several approaches were used to eliminate bias and ambiguity in the research selection process, such as (i) conducting a dual review, (ii) defining clear and transparent inclusion and exclusion criteria, and (iii) tracing the resulting flow diagram using the PRISMA flow diagram. Firstly, two researchers independently analyzed the data and resolved inconsistencies through group discussion (P. and K.T.P). Then, the abstracts and complete texts of all relevant articles were carefully studied, and only those that fit the inclusion and exclusion criteria were chosen. Researchers then confirmed the selected papers and resolved any conflicts; if any disagreements persisted, third researchers were invited to discuss the matter, and the findings were appraised (Z.-Y.S., C.-R.S., and W.J.). There was no dispute over the papers included in this review.

This study should propose a good overview of FL for the healthcare sector and more in-depth about establishing FL's secure medical data mechanism. Thus, in the screening step, we define the inclusion and exclusion criteria. We included publications that (i) use FL to develop a model on a medical dataset, (ii) are published in well-known journals, and (iii) are published in English. Exclusion criteria were used to exclude the published studies that were not related, based on the following criteria: (i) articles that are not related to FL, (ii) FL for nonmedical application or not using medical dataset in the experiment, (iii) non-English language, (iv) review article, (v) proceeding or conference papers, (vi) arXiv preprints, and (vii) book, book chapter, book section.

Numerous considerations exist against the inclusion of conference papers in this study [36]. Firstly, conference proceedings usually contain various topics and much larger set of publications such that identifying suitable conferences, accessing their abstracts, and sifting through the frequent thousands of abstracts can be time-consuming and resource-consuming. Secondly, conference proceedings may lack sufficient information for systematic reviewers to evaluate the methods, risk of bias, and outcomes of the studies submitted

at the conference due to their brevity. Finally, the reliability of the results is also in question especially in the healthcare area, partly because they are frequently preliminary or based on limited investigations undertaken in a position to meet conference deadlines. Thus, we do not include the conference papers in the inclusion criteria.

After applying inclusion and exclusion criteria from each study's title, abstract, and keywords, 56 articles were identified in the screening step. Next, 32 articles were excluded in the reports assessed for eligibility step due to exclusion criteria from full text in the article, ending with 24 articles. Finally, in the included step, 24 articles using FL in the healthcare application were selected for further analysis, and their results are discussed in this study. All of the 24 selected FL studies in the healthcare domain are listed in Table A2.

To provide a numerical description of the literature review, we gathered information from each article as follows: (i) paper information, such as author, title, year, and keywords; (ii) proposed methods, such as FL training algorithms and deep learning/machine learning models; (iii) data properties, such as medical datasets, data distribution techniques and challenges, data partition techniques, privacy attacks, and privacy mechanisms; and (iv) experiment results and discussion.

3. Results

We compiled the data properties in FL for healthcare applications from 24 published articles, as shown in Figure 3. The data scheme settings consisted of four layers: (i) data partitions such as horizontal federated learning (HFL), vertical federated learning (VFL), and federated transfer learning (FTL) (as discussed in Section 4.2); (ii) data distribution characteristics (non-IID) such as quantity skew, label distribution skew, feature distribution skew, and concept shift skew (as discussed in Section 4.3); (iii) possible data privacy attacks such as model inversion and membership inference attacks (as discussed in Section 4.4.1); (iv) additional data privacy protections such as differential privacy and homomorphic encryption (as discussed in Section 4.4.2). Above the medical data properties is the application task, where the task can be a classification or segmentation (as discussed in Section 4.6).

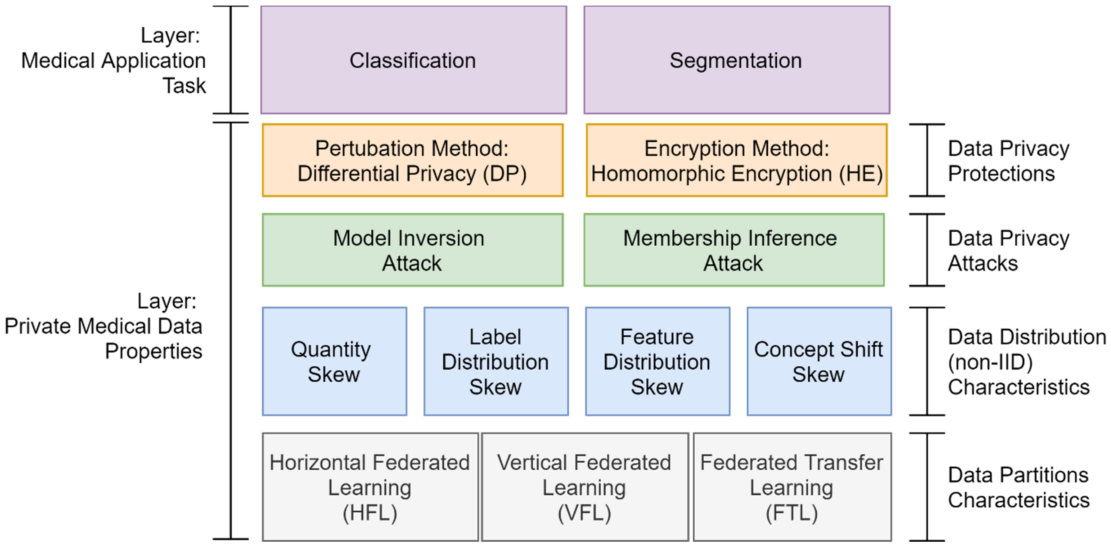


Figure 3. Medical data properties in federated learning for medical applications, consisting of data partitions, data distribution (i.e., non-IID) characteristics, possible data privacy attacks, and data privacy protections.

Numerical description. The following observations were made based on numerical analysis of the 24 included studies between 2018 and June 2021. Firstly, Figure 4a depicts the number of FL studies published in the medical application by year of publication. Since 2020, the number of articles published on FL has been continuously increasing. The number of papers published in 2021 should continue to increase linearly throughout the year. Secondly, Figure 4b shows the number of studies with data partition characteristics employed in FL.

According to the figure, most published FL studies use horizontal federated learning (HFL) as a medical data partition. Thirdly, Figure 4c shows the number of studies with various defense methods to protect from data privacy attacks. We can see that differential privacy is the most often employed type of data privacy protection. All of the possible data privacy protection methods will be discussed in Section 4.4. Based on Figure 4d, quantity skew is typical when dealing with multi-institutional medical data from FL experiments.

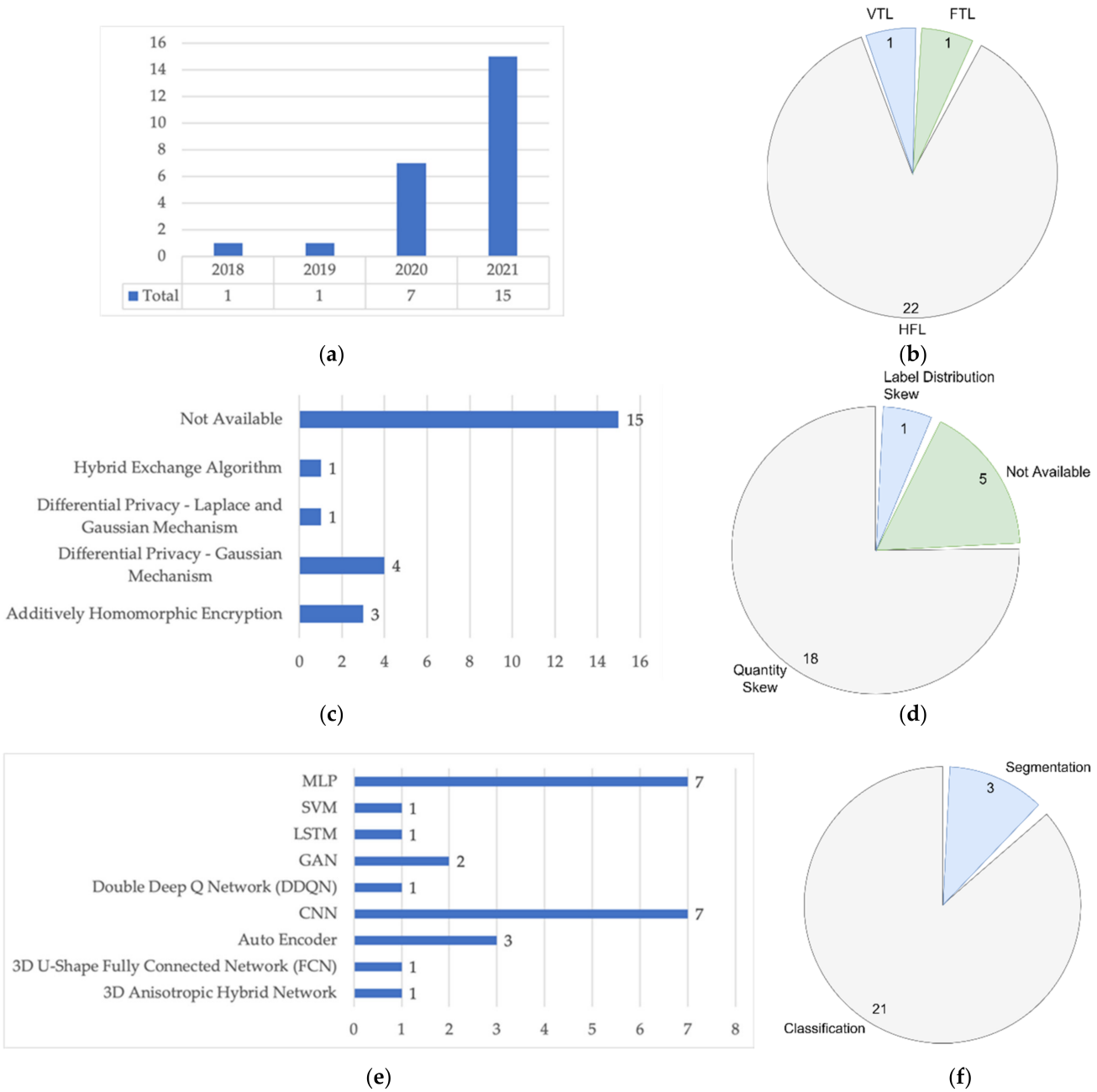


Figure 4. Numerical description of published articles in federated learning for medical applications. (**a**) The number of FL studies published in medical application included in the review, 2018–2021; (**b**) number of data partition characteristics employed in FL; (**c**) various data privacy algorithms employed in federated learning for the healthcare area; (**d**) number of non-IID characteristics discussed in FL studies published in the medical domain; (**e**) various machine learning models employed in federated learning for the healthcare area; (**f**) number of FL studies published in medical application included in the systematic review.

Machine learning algorithms. Additionally, we want to outline the machine learning models employed in the studies and evaluate their proposed FL algorithms. The outlined result of the machine learning model is shown in Figure 4e, where multilayer perceptron (MLP) is the most commonly used model when predicting with tabular medical datasets such as mortality prediction. Meanwhile, convolutional neural network (CNN) is the frequent model architecture used for medical image datasets. Other models include support vector machine (SVM) and autoencoder (AE) models. Additionally, we compile the machine learning task based on the 24 published articles, as shown in Figure 4f. There were 21 studies on classification tasks and three studies on segmentation tasks. Finally, we summarized in Table 1 the strengths and weaknesses of machine learning algorithms performing on federated learning.

Table 1. Summary of machine learning algorithms performing on federated learning, along with strengths and weaknesses.

ML Algorithms	Strength	Weakness	FL Study
AE	AE is mainly designed for dimensional feature reduction and denoising medical datasets via an unsupervised learning method. AE aims to recreate effective compact and effective feature representation.	An autoencoder may exclude essential information from a medical dataset's characteristics.	[19,21,37]
CNN	Performs well on medical image classification tasks such as prediction of COVID-19 using X-ray images	The training process of CNN that contains multiple layers will be time-consuming if the client in the FL environment does not have powerful computation resources.	[18,20,38–42]
GAN	Generate a synthetic sample of medical data for limited quantity in experiments datasets.	Training GAN is challenging due to the unstable training process, no standard metric evaluation, and numerous trial-and-error experiments required for effective outcomes.	[43,44]
LSTM	Performs well on time series or sequential medical datasets, for instance, detection of human activity recognition.	Due to the vanishing and exploding gradient challenges, training LSTM is difficult.	[45]
MLP	Good generalization performance on tabular medical datasets such as mortality prediction based on drug data	MLP is limited to learning elementary problems. Additionally, it is feature-scaling sensitive and involves setting numerous hyperparameters such as the number of hidden neurons and layers.	[46–51]
SVM	SVM is capable of modeling nonlinear decision boundaries and a variety of kernels are available. Additionally, it is highly resistant to overfitting, particularly in high-dimensional space.	SVM is memory-consuming, more difficult to modify because of critically selecting the appropriate kernel, and does not scale well to more extensive datasets.	[52]
U-Net	Achieve accurate results when performing segmentation tasks on medical image datasets, for example, when segmenting brain tumors disease using brain magnetic resonance medical images.	U-Net model development is time-consuming because the network must be operated independently for each patch, and redundancy due to overlapping patches. Additionally, a tradeoff exists between the precision of localization and the utilization of context.	[38,53]

AE: autoencoder; CNN: convolutional neural network; GA.: generative adversarial network; LSTM: long short-term memory; MLP: multilayer perceptron; SVM: support vector machine.

4. Discussion

RQ1: What are the state-of-the-art FL methods in the healthcare area?

4.1. Federated Learning Overview

FL is a technique to develop a robust quality shared global model with a central aggregate server from isolated data among many different clients. In a healthcare application scenario, assume there are K nodes where each node k holds its respective data $\mathcal{D}_k$ with n_k total number of samples. These nodes could be a healthcare wearable device, an internet of health things (IoHT) sensor, or a medical institution data warehouse. The FL objective is to minimize loss function given total data $n = \sum_{k=1}^{K} n_k$ and trainable machine learning weight vectors with d parameters $w \in R^d$ using Equation (1):

$$\min_{w \in R^d} F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \text{ where } F_k(w) = \frac{1}{n_k} \sum_{x_i \in \mathcal{D}_k} f_i(w) \tag{1}$$

where $f_i(w) = \updownarrow(x_i, y_i; w)$ denotes the loss of the machine learning model made with parameter w. For instance, Huang et al. [19] used the categorical cross-entropy loss function to update the model parameters on the binary classification of patient mortality. In addition, Yang et al. [53] used the soft dice loss function for the COVID-19 region segmentation application.

In 2016, the basic concept of data parallelism in FL namely federated averaging (FedAvg) algorithm, was introduced by McMahan et al. [16]. As stated in the FedAvg algorithm, every communication round t consists of four phases. Firstly, the aggregate server initializes a global model with initial weights w_t^g, then shared with a group of clients S_t (medical nodes in our case), which was picked randomly with a fraction of $C \in \{0, 1\}$. Secondly, each client $k \in S_t$, after received a global model w_t^g from the server, the client conducts local training steps with epoch E on minibatch $b \in B$ of n_k private data points. The local model parameters are updated with local learning rate η and optimized by minimizing loss function $\mathcal{L}(.)$. Thirdly, once client training is completed, the client k sends back its local model w_{t+1}^k to the server. Finally, after receiving the local model w_{t+1}^k from all selected groups of clients S_t, the aggregate server updates the global model w_{t+1}^g by averaging of local model parameters using Equation (2):

$$w_{t+1}^g \leftarrow \sum_{k=1}^{K} \alpha_k \times w_{t+1}^k \tag{2}$$

where α_k is a weighting coefficient to indicate the relative influence of each node k on the updating function in the global model, and K is the total nodes that participated in the training process. Choosing the proper weighting coefficient α_k in the averaging function can help improve the global model's performance (as discussed in Section 4.3.2 non-IID mitigation methods). The entire FL procedure is described in Algorithm 1.

Algorithm 1 FL with Federated Averaging (FedAvg) algorithm [16]

Input: T global round, C number of fractions for each training round, K number of clients, η learning rate at a local client, E number of epochs at a local client, B local minibatch at a local client.

Initialize global model $w_{t=0}^{g}$
for each round $t = 1, 2, \ldots, T$ **do**
  $m \leftarrow \max(C \times K, 1)$
  $S_t \leftarrow (m \text{ clients in a random order})$
  for each client $k \in S_t$ **do**
    $w_{t+1}^{k} \leftarrow ClientUpdate\left(k, w_t^g\right)$
    $w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \alpha_k \times w_{t+1}^{k}$

ClientUpdate$\left(k, w_t^g\right)$:
  $w_k \leftarrow \mathrm{w}_{\mathrm{t}}^{\mathrm{g}}$
  for each local epoch $e = 1, 2, \ldots E$ **do**
    for each local batch $b \in B$ **do**
      $w_k \leftarrow w_k - \eta \nabla \mathcal{L}(b; \mathrm{w_k})$
  return local model w_k

Output: w_{t+1}^{g} a global model at round $t+1$

FL has differentiated from the standard collaborative learning in the following properties: (1) training is carried out across a vast number of many client nodes, and communication speed between the client nodes and the aggregate server is slow; (2) the central aggregate server does not have a control to individual nodes or devices, and full participation of all nodes is unrealistic because there are inactivate devices that do not respond to the server; (3) in real-world case scenario, data distribution is nonindependent and identically distributed (non-IID). Non-IID data distribution means that each node has a different distribution pattern from the other node. These properties are shown when the first proposed of FL algorithm is applied for mobile keyboard prediction [16,17]. However, these properties are different when FL is implemented in the healthcare area. First, the FL training is carried out across a limited number of healthcare nodes from 2 to 100 as listed in Table 2, and communication speed between healthcare participants and the aggregate server is usually reliable. Second, the aggregate server coordinates the participant nodes in the FL training scheme without exposing the participant's local data to the network; thus, data privacy and security can be guaranteed.

FL is divided into two categories based on the aggregation schema: (a) centralized FL and (b) decentralized FL. As shown in Figure 5a, for centralized FL, the central server selects a subset of nodes at the beginning of training and aggregates the model updates received from client nodes. As nodes, the medical institutions periodically communicate the local updates w_{t-1}^{k} with a central server to learn a global model w_t^g. The central server aggregates the updates and sends back the parameters of the updated global model. However, if the centralized server fails, the whole FL environment will collapse. This failure is one of the reasons that the decentralized FL was proposed. Specifically, all nodes coordinate themselves and work together from node to node to develop a global model in decentralized FL, as shown in Figure 5b.

RQ2: What are the FL methods proposed by scholars to solve challenging medical applications from a data properties perspective?

4.2. Data Partition Characteristics

This section discusses FL based on the healthcare data partition characteristics. Since FL uses data kept in various medical institutions, it is frequently presented in a feature matrix. Let matrix $\mathcal{D}_k$ denote medical data held by the medical institution k. Notably, a row in the matrix represents a patient index denoted by $\mathcal{I}$, a column represents a patient features diagnosis denoted by $\mathcal{X}$, and some data may contain a label data $\mathcal{Y}$. The complete training medical dataset $\mathcal{D}_k$ in a medical institution k is denoted by $(\mathcal{I}_k, \mathcal{X}_k, \mathcal{Y}_k)$. Thus, data

partition in FL can be divided into horizontal federated learning (HFL), vertical federated learning (VFL), and federated transfer learning (FTL) [26].

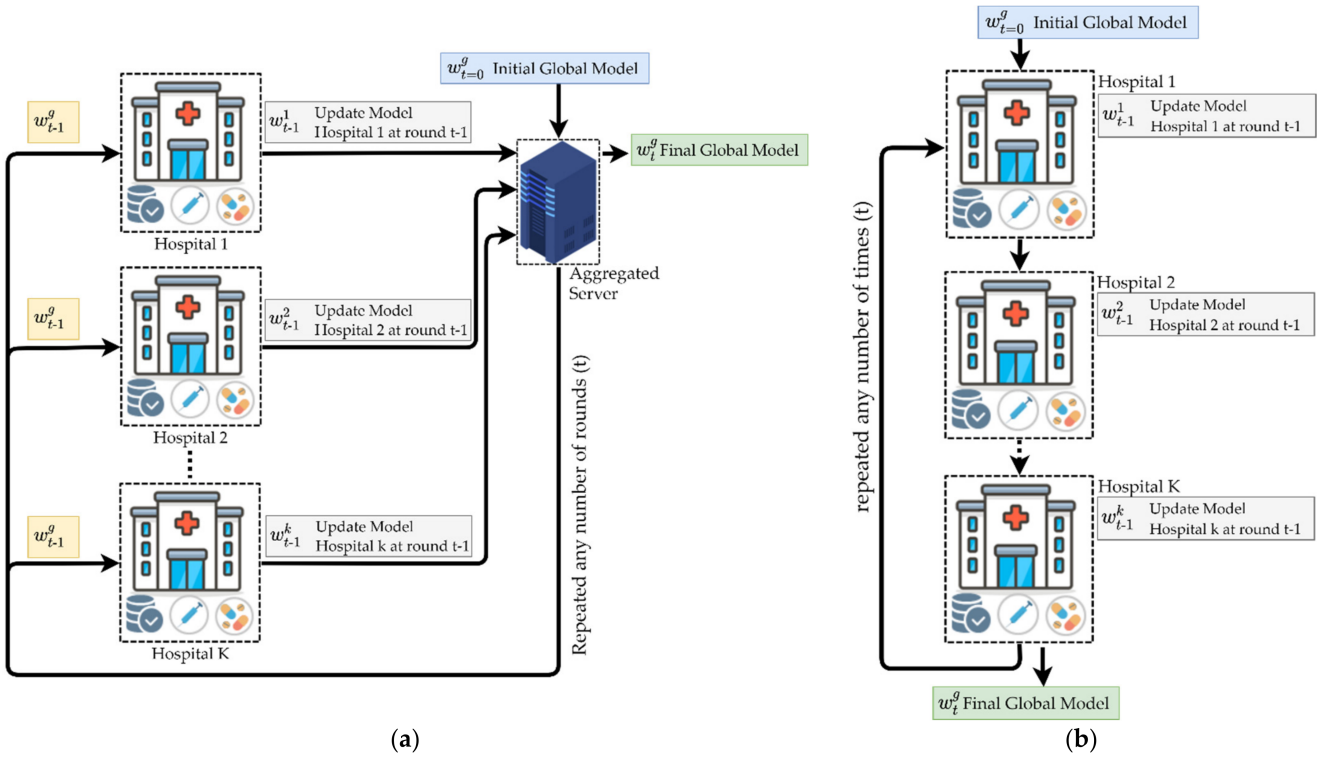


Figure 5. Federated learning framework for healthcare application based on aggregation schema. (**a**) Centralized FL: the central server selects the nodes, aggregates the updates, and sends back the updated global model parameters; (**b**) decentralized FL: to develop a global model, there is no central server to orchestrate all nodes.

4.2.1. Horizontal Federated Learning (HFL)

The horizontal federated learning (HFL) data partition, shown in Figure 6, is recommended in the case of limited sample size variability when developing a model. In this data partition setting, the nodes could be different health institutions or health data application providers. The HFL aims to develop a global model by integrating patients' sample data from different institutions without affecting patient privacy. Each node shares different patients' index $\mathcal{I}$ but has the same features $\mathcal{X}$ and labels $\mathcal{Y}$ information [26]. HFL is denoted as:

$$\mathcal{X}_j = \mathcal{X}_k, \ \ \mathcal{Y}_j = \mathcal{Y}_k, \ \ \mathcal{I}_j \neq \mathcal{I}_k, \ \ \forall \mathcal{D}_j, \mathcal{D}_k, \ \ j \neq k \tag{3}$$

where D_i represents the dataset held by client i. For instance, two healthcare providers of the same business located in different countries would like to develop an AI model. User features of these two healthcare providers will mostly be the same because both operate the same business. However, the patient samples held by the two healthcare providers are different due to geographic locations. In this regard, we can use HFL to increase the total training sample by aggregating both of the healthcare providers' user samples in a privacy-preserving manner to enhance the model's performance. Therefore, the HFL data partition resolves the lack of sample size in data training because it combines all healthcare institutions' sample data.

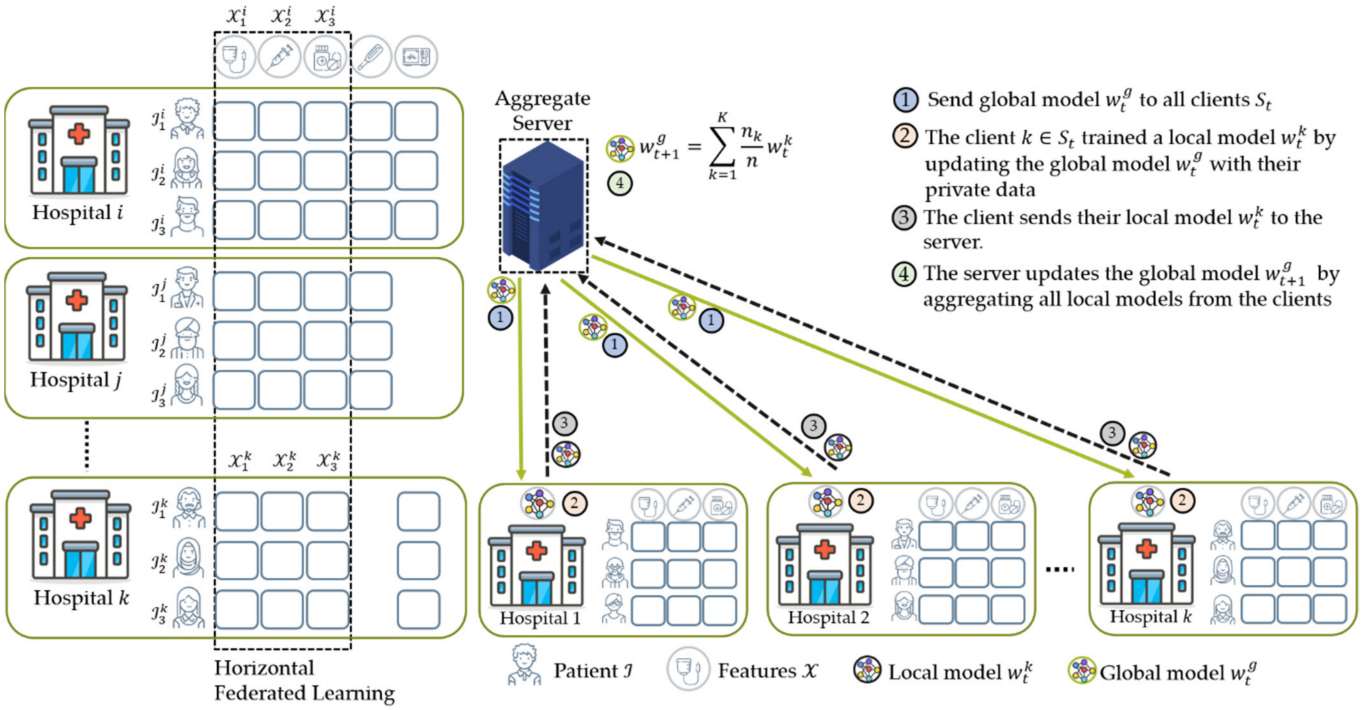


Figure 6. The typical medical data partitions scenario for horizontal federated learning (HFL). Each node is a medical institution data silo or wearable medical device. They share the same feature of medical diagnosis $\mathcal{X}_j = \mathcal{X}_k$ but have different patients index $\mathcal{I}_j \neq \mathcal{I}_k$.

HFL data partition is quite common in FL applied for medical applications. More than half of FL studies on medical applications implemented horizontal medical data partition in their experiment [18,19,21,37,39–49,51,52,54,55]. Unlike FL applied for nonmedical applications where training is carried out across many nodes, FL studies in medical applications only handle limited nodes from 2 to 100, as listed in Table 2. For instance, Li et al. [18] experimented with four medical institutions in different places for the autism spectrum disorder (ASD) prediction scenario. Each medical party shares the same user features generated by medical equipment and combines all patient samples from four medical nodes.

4.2.2. Vertical Federated Learning (VFL)

Data partition in vertical federated learning (VFL) is depicted in Figure 7. In this data partition setting, two nodes shared the same users' profile but different features information. The nodes could be different health institutions or health data application providers. VFL aims to develop a global model by integrating patient features from different institutions without directly sharing patient data. Each node shares different patients' features $\mathcal{X}$ and labels $\mathcal{Y}$ information but has the same sample data $\mathcal{I}$ [26]. VFL can be denoted as:

$$\mathcal{X}_j \neq \mathcal{X}_k, \ \mathcal{Y}_j \neq \mathcal{Y}_k, \ \mathcal{I}_j = \mathcal{I}_k, \ \forall \mathcal{D}_j, \mathcal{D}_k, \ j \neq k \tag{4}$$

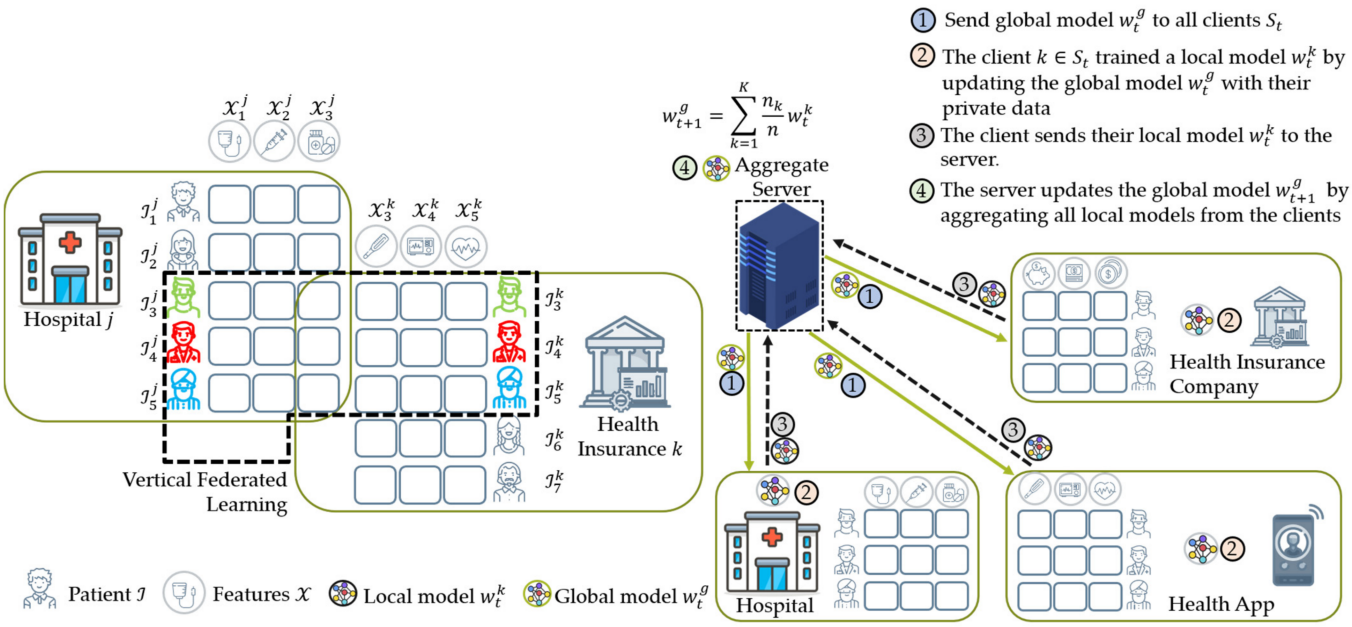


Figure 7. The typical medical data partitions scenario for vertical federated learning (VFL). Each node can be a different medical institution and application. They share the same patients' index $\mathcal{I}_j = \mathcal{I}_k$ but have different features of medical diagnosis $\mathcal{X}_j \neq \mathcal{X}_k$.

For example, two distinct healthcare organizations exist in the same region: one hospital and one health insurance company. Users of these two healthcare organizations may mostly be the same because they are the region's residents. However, the user features may not have anything in common because healthcare insurance records users' income and medical reimbursement, while hospitals keep users' medical treatment histories. VFL data partition securely combines different features sets to enhance the performance of the model. Thus, the VFL data partition increases feature dimension in data training.

In contrast to HFL, there are a few published VFL-based studies applied in medical applications. One such an example was proposed by Cha et al. [56]. The authors developed an autoencoder federated learning model for the vertically partitioned medical data. An autoencoder model is used for transforming user features in each client into a latent dimension. The proposed method does not share any raw medical data but latent dimensions as secure perturbed data. After receiving the clients' latent dimensions, the aggregate server concatenates all latent dimensions for training the global model. However, this approach is prone to reverse-engineering, which could discover the original medical data from the latent dimensions. In addition, the proposed method needs all the clients to perform data alignment, which means the user data has the same row indices in all data silos (first row data on clients k must be the same as client j).

4.2.3. Federated Transfer Learning (FTL)

Unlike the data configurations in HFL and VFL, data partition in federated transfer learning (FTL) considers the situation of multiple nodes shared neither the same users' profile nor features information, as shown in Figure 8. The main issue in this data partition configuration is that one node lacks labeled data. The nodes could be different health institutions or health data application providers located in different regions. Furthermore, each node shared different patients' features $\mathcal{X}$, labels $\mathcal{Y}$, and sample data index $\mathcal{I}$ [26]. FTL can be denoted as:

$$\mathcal{X}_j \neq \mathcal{X}_k, \ \ \mathcal{Y}_j \neq \mathcal{Y}_k, \ \ \mathcal{I}_j \neq \mathcal{I}_k, \ \ \forall \mathcal{D}_j, \ \mathcal{D}_k, \ \ j \neq k \tag{5}$$

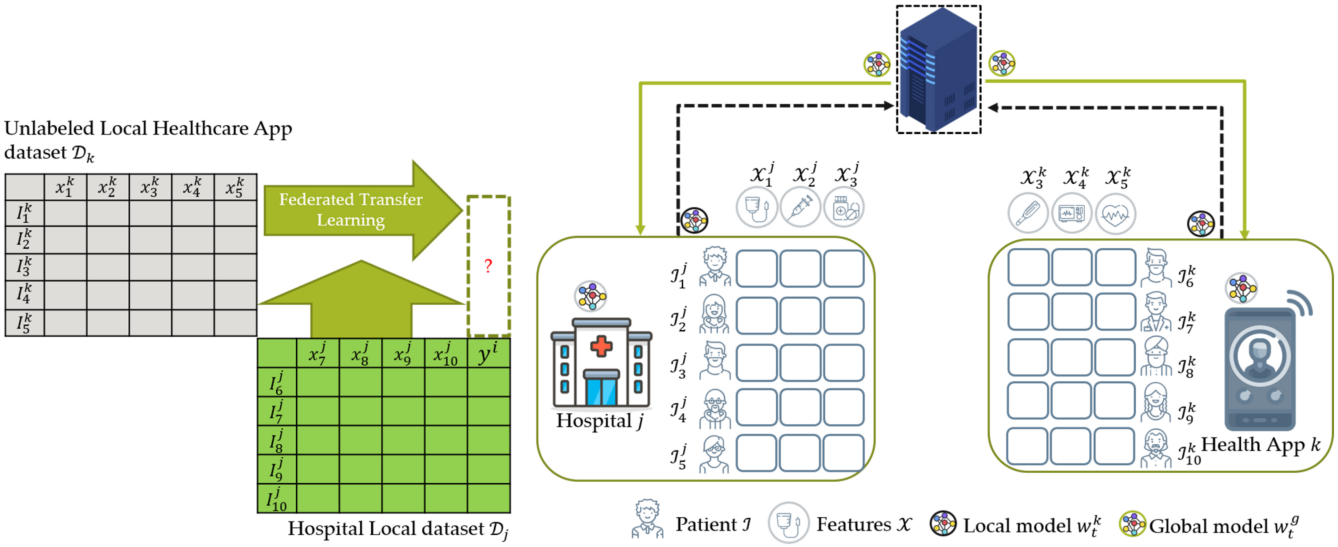


Figure 8. The typical medical data partitions scenario for federated transfer learning (FTL). One party is a medical institution, while the other is a healthcare application located in a different region. They share neither the patients' index $\mathcal{I}_j \neq \mathcal{I}_k$ nor features of medical diagnosis $\mathcal{X}_j \neq \mathcal{X}_k$.

For example, there are two distinct healthcare entities: one is a hospital in Taiwan, while the other is in the United States. Due to the geographical limitations, the two healthcare entities' user groups have little overlap, and the data features of the two entities datasets may slightly overlap. FTL addresses limited data sets and label samples in this scenario, thus increasing the model's performance while protecting user privacy.

The research in FTL is still in the early stages, and there is plenty of room for improvement. Chen et al. [20] proposed FedHealth assuming FTL data partition. FedHealth method collects data from several users/organizations using FL then offers a personalized model for each user/organization using transfer learning. First, the model learns to classify human activity and then extends the task to Parkinson's disease classification with transfer learning. In this case, FTL developed a global model for disease prediction in one task and then could be transferred to another task.

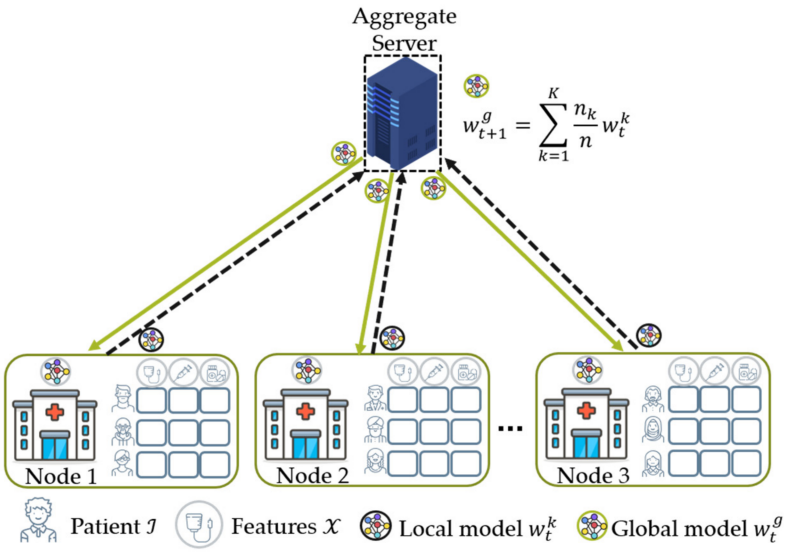
4.3. Data Distribution (Statistical Data Heterogeneity) Challenge

FL can solve the limited data quantity issue by combining data from each client without directly sharing each client's private data. However, FL also faces statistical data heterogeneity challenges due to data distribution at each client. The data distributions at each client are likely to be different, leading to poor global model performance [23,24]. Zhao et al. [23] demonstrated that the data distributions might considerably decrease FL model performance due to weight divergence induced by different population distributions. Within an FL environment, data distribution is frequently classified into IID and non-IID. Non-IID can result from an imbalance in the amount of data quantity, features, or labels. Non-IID is a common occurrence in the medical domain. Various medical tools manufacturers, different calibrated techniques, and different medical data acquisition techniques are the main reasons why each medical institution generates nonidentical data distribution. For instance, Li et al. [18] described how each medical institution uses various brain scanner manufacturers and instructions for each patient when taking autism brain imaging data. Specifically, during data acquisition, one medical site instructs patients to keep their eyes open while others instruct them to close their eyes during scanning. In the following subsection, we describe the non-IID characteristics and mitigation methods.

4.3.1. Non-IID Characteristics

The non-IID characteristics among healthcare nodes in the FL environment can take on four different forms such as (1) quantity distribution skew, (2) label distribution skew, (3) feature distribution skew, and (4) concept shift skew [24,25]. The non-IID characteristic summarized from 24 published FL studies applied for medical application is listed in Table 2.

Quantity skew (imbalance data) characteristic. Quantity skew characteristic in non-IID occurs when the class distribution of data instances $\mathcal{I}$ is not equal or far from equal across nodes in the FL scheme. An illustration of quantity skew is shown in Figure 9. In the IID scenario, the amount ratio of positive and negative instances is almost equal. For instance, in node two, the negative and positive amount ratios are 45% and 55%, respectively. In the non-IID case, the ratio of positive and negative instances is far from equal. For example, in node one, positive instances are around 5%, while negative ones are 95%. Krawczyk et al. [57] divided imbalance data categories into slight imbalance and severe imbalance. A slight imbalance is when the majority class is uneven by a small amount in the training dataset, and the ratio ranges from 1:4 up to 1:100. Severe imbalance data distribution is when the data distribution of the majority class is uneven by a vast amount in the training dataset, the ratio is more than 1:100. For example, the ratio of imbalance data in fraud detection tasks is up to 1:1000.



	Features $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$				Label
	x_1	x_2	x_3	x_4	y
Node 1	✓	✓	✓	✓	✓ 40 % instances ✗ 60 % instances balance data
Node 2	✓	✓	✓	✓	✓ 45 % instances ✗ 55 % instances balance data
Node 3	✓	✓	✓	✓	✓ 50 % instances ✗ 50 % instances balance data

Label ✗ (Dead) ✓ (Alive)

(a) Identically Distributed–Quantity distribution

	Features $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$				Label
	x_1	x_2	x_3	x_4	y
Node 1	✓	✓	✓	✓	✓ 5 % instances ✗ 95 % instances imbalance data
Node 2	✓	✓	✓	✓	✓ 10 % instances ✗ 90 % instances imbalance data
Node 3	✓	✓	✓	✓	✓ 50 % instances ✗ 50 % instances balance data

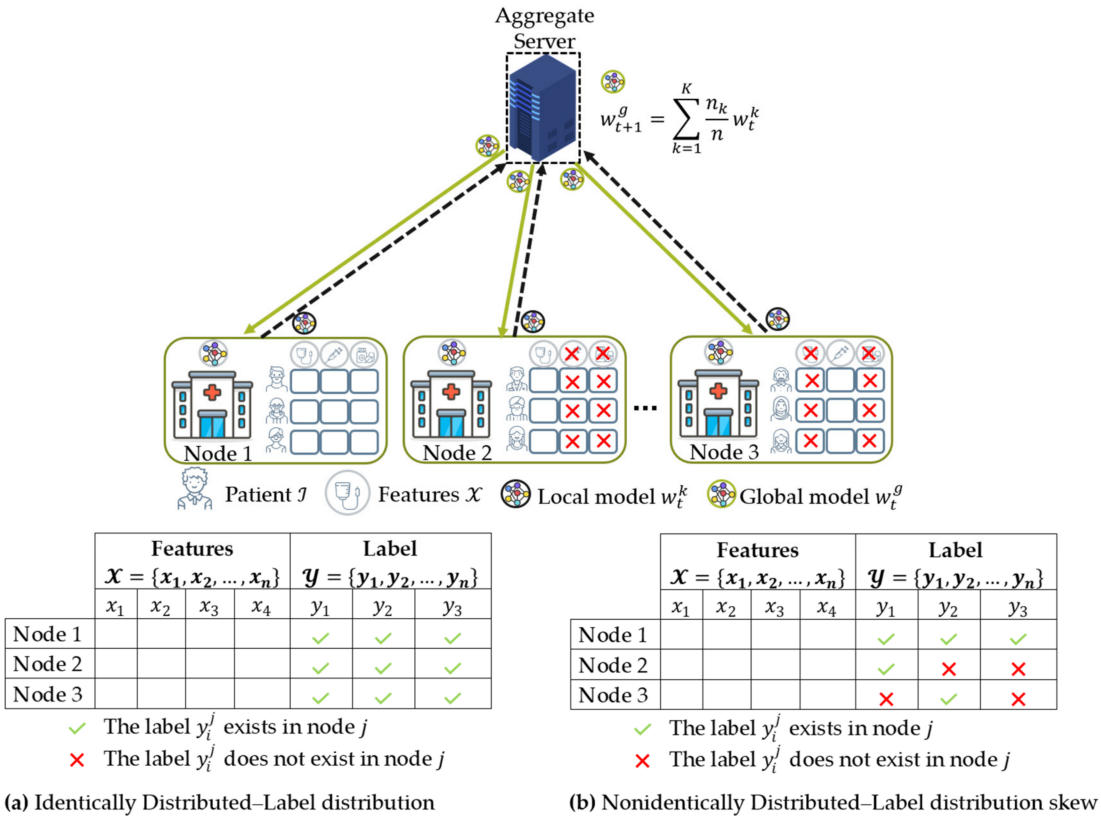
Label ✗ (Dead) ✓ (Alive)

(b) Nonidentically Distributed–Quantity distribution skew

Figure 9. Non-IID from quantity skew (i.e., imbalanced dataset) characteristic. (**a**) IID: the amount ratio of positive and negative instances is equal or slightly equal; (**b**) non-IID: the ratio of positive and negative instances is far from equal. For example, the positive and negative instances ratio is 5% and 95% in node one, respectively.

Quantity skew characteristic exists in FL for medical application experiment datasets such as [18,19,46,52,53]. Quantity skew (i.e., imbalanced dataset) is common in the medical dataset since it is acquired from multiple healthcare institutions, and the number of instances in a class is not equally distributed for each institution. For instance, larger hospitals have more patient records than small clinics in rural areas. Huang et al. [19] tried to resolve this challenge by developing an imbalanced eICU dataset to predict patient mortality where the ratio is 5% and 95% for death and alive categories, respectively.

Label distribution skew characteristic. For label distribution skew, the distribution of labels $P(y_i)$ varies between different nodes. In the medical case, larger hospitals generally have more disease-related records than small clinics in rural areas. An illustration for label distribution skew characteristic is shown in Figure 10. In the IID setting, the distribution of labels $\mathcal{Y}$ is the same across all nodes. However, in the non-IID setting, the distribution of labels $\mathcal{Y}$ varies between each node. Specifically, there is a label y_i that only exists in one or several nodes in the FL environment. This label distribution skew characteristic was initially demonstrated in FedAvg's experiment [16]. Data samples with the same label are divided into subsets, and each client is assigned to no more than two subsets with distinct labels. Following FedAvg, this configuration is employed in published FL studies for medical applications [38].



	Features $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$				Label $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$		
	x_1	x_2	x_3	x_4	y_1	y_2	y_3
Node 1					✓	✓	✓
Node 2					✓	✓	✓
Node 3					✓	✓	✓

✓ The label y_i^j exists in node j
× The label y_i^j does not exist in node j

(a) Identically Distributed–Label distribution

	Features $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$				Label $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$		
	x_1	x_2	x_3	x_4	y_1	y_2	y_3
Node 1					✓	✓	✓
Node 2					✓	×	×
Node 3					×	✓	×

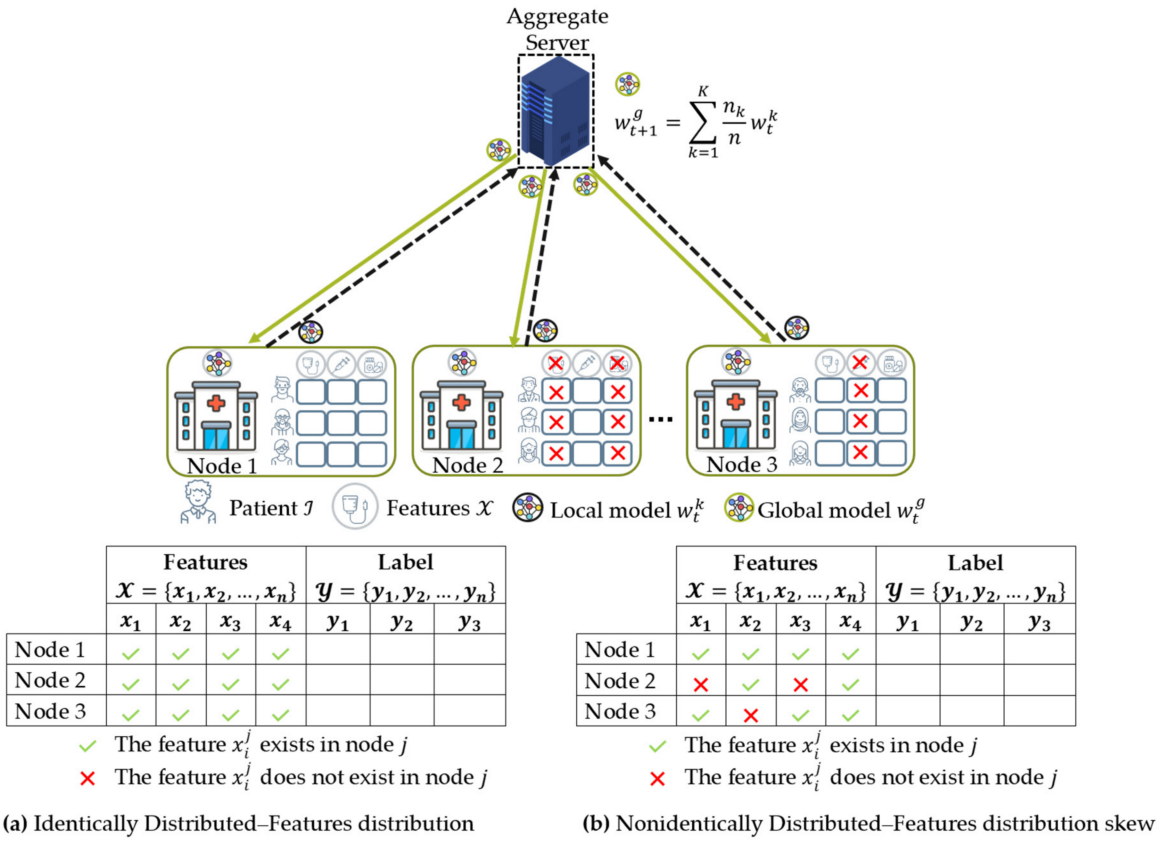
✓ The label y_i^j exists in node j
× The label y_i^j does not exist in node j

(b) Nonidentically Distributed–Label distribution skew

Figure 10. Non-IID from label distribution skew (prior probability shift) characteristic. (**a**) IID: The distribution of labels $\mathcal{Y}$ exists in all nodes; (**b**) non-IID: the distribution of labels $\mathcal{Y}$ varies between different nodes. For instance, node two does not have the labels y_2 and y_3 while node one has all labels.

Feature distribution skew characteristic. In the feature distribution skew characteristic, the distribution of features $P(x_i)$ varies between different nodes. An illustration of features distribution skew is shown in Figure 11. In the case of IID, the distribution of features $\mathcal{X}$ is the same across all nodes, while in the non-IID case, the distribution of features $\mathcal{X}$ varies between each node. Specifically, there is a feature x_i that only exists in one or several nodes in the FL environment. For instance, node two does not have the x_1

and x_2 features while other nodes have those features. Missing features or missing data is a common occurrence in medical datasets. For instance, missing features can be caused by failures of measurement on medical images. Measurement in medical image acquisition requires the images to be in focus. Medical images that are not in focus or blur can cause missing pixel values. The absence of some features in one or several nodes in the features distribution skew can be a problem in the FL training process. Data imputation techniques such as probability principal component analysis (PPCA) and multiple imputations using chained equations (MICE) can be employed to mitigate the problem [58].



(a) Identically Distributed–Features distribution **(b)** Nonidentically Distributed–Features distribution skew

Figure 11. Non-IID from feature distribution skew characteristic. (**a**) IID case: the distribution of features $\mathcal{X}$ exists in all nodes; (**b**) non-IID case: the distribution of features $\mathcal{X}$ varies between each node. For instance, node two does not have the features x_1 and x_3 while the other nodes have those features.

Concept Shift Skew. There are two forms in the concept shift skew: the same label but different features $P(x|y)$ and the same features but different label $P(y|x)$. An illustration of concept shift skew is depicted in Figure 12. The same label but different features in non-IID characteristic is related to vertical federated learning data partition where each node shares the sample index $\mathcal{I}$ but have different features $\mathcal{X}$, while in the case of the same features but the different label in non-IID characteristics is not applicable in most FL studies.

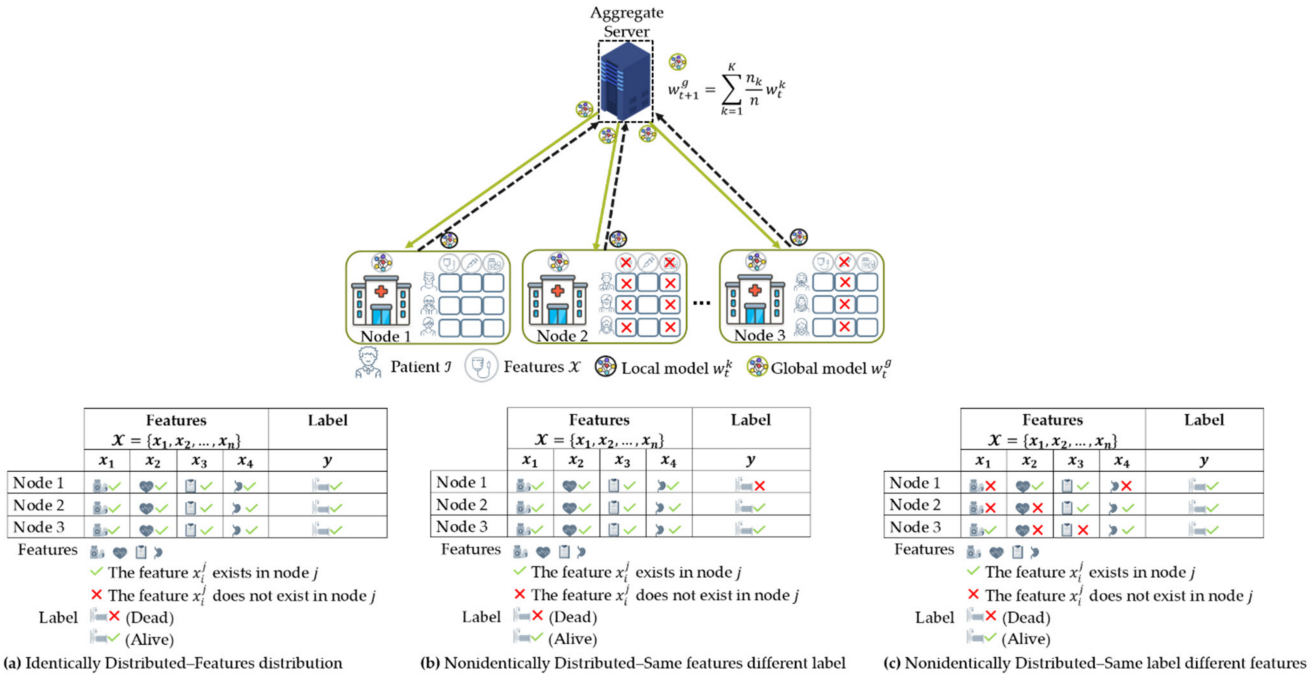


Figure 12. Non-IID from concept shift skew characteristic. (**a**) IID case; (**b**) non-IID same features but different label case; (**c**) non-IID same label but different features case.

4.3.2. Non-IID Mitigation Methods

Different non-IID characteristics may need different mitigating measures. There are three methods in the published FL for medical applications to improve the model performance with the non-IID dataset: (1) balancing the training dataset, (2) tuning the model hyperparameter in the FL algorithm, and (3) domain adaptation.

Balance the training dataset method. When dealing with quantity skew in non-IID characteristics, researchers balance the quantity of minority and majority classes in the training dataset with the synthetic data augmentation technique. It is important to note that the balancing method in the FL environment should keep the data secure and private. There are two methods to generate synthetic data augmentation in the FL environment: (1) local data augmentation and (2) server data sharing.

(1) The healthcare node generates a synthetic sample to balance the training dataset in the local data augmentation method. The synthetic minority oversampling technique (SMOTE) [21,49], generative adversarial method (GAN) [44], or geometric transformation [40,48,53] is employed to generate a synthetic sample in an FL environment. The SMOTE algorithm is an oversampling technique where the synthetic data are generated for the minority class. For instance, Wu et al. [21] and Rajendran et al. [49] employ SMOTE to balance the heavy imbalance in a fall detection and lung cancer training dataset, respectively. Zhang et al. [44] proposed secure synthetic COVID-19 data by combining the GAN and differential privacy method. Feki et al. [40], Duo et al. [48], and Yang et al. [53] applied geometric transformations such as random flipping, random rotation, and random translation to balance the quantity of minority class in their training dataset for the data augmentation method.

(2) The aggregate server securely shares a small portion of data to the healthcare node in the server data sharing method. For instance, Zhao et al. [23] proposed a global shared dataset partition to train non-IID data. The author demonstrated that by simply sharing 5% of data, they could get a 30% boost accuracy score. However, it raises model communication costs and is prone to data privacy attacks during the data sharing process.

Adaptive Hyperparameters Method. The adaptive hyperparameters method tries to find the proper FL hyperparameters values for each node during the training process. Each node can have different values of the FL hyperparameters, such as learning rate, loss score, and weighting coefficient. There are two published adaptive hyperparameters methods in the published FL studies for medical application: (1) weighting coefficient [16,19,20,45], and (2) adaptive loss function [46].

(1) The weighting coefficient α_k is a variable that indicates the relative influence of each node k on the aggregation equation in Equation (2) to update the global model. Initially, McMahan et al. [16] proposed FedAvg that the weighting coefficient is $\alpha_k = \frac{n_k}{n}$ as shown in Equation (6), where n_k and n are the private data points hold by node k and the total data from all nodes that participated during training, respectively. In this case, a node with significant data points has a considerable effect on the global model. This method worked well when dealing with label distribution skew characteristics experimented in their studies [16,20].

$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \times w_{t+1}^{k} \tag{6}$$

In comparison, Chen et al. [20] proposed that the weighting coefficient is $\alpha_k = \frac{1}{K}$, where K is the total nodes participating in FL as shown in Equation (7). In this scenario, the author considered that each node would contribute equally to the aggregation function.

$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \frac{1}{K} \times w_{t+1}^{k} \tag{7}$$

Huang et al. [19] proposed that the weighting coefficient is $\alpha_k = \frac{m_k^c}{\sum_{c=1}^{C} m_k^c}$, as shown in Equation (8), where m_k^c and $\sum_{c=1}^{C} m_k^c$ are denoted as the clusters size in medical node k and the total number of clusters in community-based federated learning, respectively. In their method, the algorithm considers the weighted average from the cluster patient community.

$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \frac{m_k^c}{\sum_{c=1}^{C} m_k^c} \times w_{t+1}^{k} \tag{8}$$

Finally, Chen et al. [45] proposed that the weighting coefficient is $\alpha_k = \frac{n_k}{n} \times \left(\frac{e}{2}\right)^{-(t-timestamp^k)}$, as shown in Equation (9), where e is the natural logarithm number to denote the time effect and $timestamp^k$ is the round in the newest updated local model. Their proposed weighting coefficient considers not only the data samples held by node k shown by the portion of data $\frac{n_k}{n}$ but also the time required to update the global model in the local node.

$$w_{t+1}^{g} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \times \left(\frac{e}{2}\right)^{-(t-timestamp^k)} \times w_{t+1}^{k} \tag{9}$$

(2) In addition, the adaptive loss function has the ability to change conditions based on the loss score function. The loss function was used to measure the model performance. The lower the loss score, the better a model was trained. Specifically, Huang et al. [46] proposed the LoAdaBoost method based on loss function in the FL environment for patient mortality prediction. In their proposed method, the adaptive loss function boosts the training process adaptively from the weak learners node. On each training step, the local node will send both the local model and training loss. If the training loss score is more than the loss threshold, it will be retraining again. Otherwise, it will send to the aggregate server.

Domain Adaptation Method. Domain adaptation (DA) is a subset of transfer learning in which a model developed in one or more "source domains" is applied to a new (but

related) "target domain." DA is used when the source and target domains share the same feature space but different data representations and distribution [59]. In comparison, transfer learning is used when the target domain's feature is different from the source domain's feature. The goal of DA is to minimize discrepancies in data distributions. Li et al. [18] incorporated domain adaptation in their FL algorithm. The fundamental assumption is that DA approaches can increase the overall performance of multiple nodes in the FL environment with non-IID. Specifically, the author implemented a mixture of expert (MoE) and adversarial domain adaptation methods. The MoE implements adaptation near the model output layer, whereas the adversarial domain alignment implements adaptation on the data knowledge representation level.

Table 2. Summary of different data partition methods, number of nodes, non-IID characteristics, and non-IID mitigation employed in the published federated learning for healthcare applications.

Data Partition	Purpose	Number of Nodes	Non-IID Characteristics	Non-IID Mitigation	Studies/Year
HFL	Combining all samples from a group of selected nodes S_t to increase the sample size	10	Quantity Skew	Balancing the training dataset	Brismi et al., 2018 [52]
		50	Quantity Skew	Not Available	Huang et al., 2019 [19]
		20	Quantity Skew	Balancing the training dataset	Chen et al., 2020 [45]
		90	Quantity Skew	Adaptive Hyperparameters: Adaptive Loss Function	Huang et al., 2020 [46]
		4	Quantity Skew	Domain Adaptation: Mixture of Expert and Domain Adversarial	Li et al., 2020 [18]
		5	Quantity Skew	Not Available	Shao et al., 2020 [47]
		10	Quantity Skew	Not Available	Sheller et al. [38]
		5	Quantity Skew	Balancing the training dataset: SMOTE Algorithm	Wu et al., 2020 [21]
		10	Not Available	Not Available	Abdul Salam et al., 2021 [54]
		4	Quantity Skew	Balancing the training dataset: Geometric Transformation	Chhikara et al., 2021 [37]
		8	Quantity Skew	Not Available	Cui et al., 2021 [39]
		3	Quantity Skew	Balancing the training dataset: Geometric Transformation	Dou et al., 2021 [48]
		4	Quantity Skew	Balancing the training dataset: Geometric Transformation	Feki et al., 2021 [40]
		6	Quantity Skew	Not Available	Lee et al., 2021 [41]
		10	Quantity Skew	Balancing the training dataset	Liu et al., 2021 [42]
		2	Quantity Skew	Balancing the training dataset: SMOTE Algorithm	Rajendran et al. [49]
		3	Not Available	Not Available	Sarma et al. [50]
		5	Quantity Skew	Not Available	Vaid et al., 2021 [55]
		8	Not Available	Not Available	Xue et al., 2021 [51]
		8	Not Available	Not Available	Yan et al., 2021 [43]
		3	Quantity Skew	Balancing the training dataset: Geometric Transformation	Yang et al., 2021 [53]
		100	Label Distribution Skew	Balancing the training dataset: Generative Adversarial Network (GAN)	Zhang et al., 2021 [44]

Table 2. *Cont.*

Data Partition	Purpose	Number of Nodes	Non-IID Characteristics	Non-IID Mitigation	Studies/Year
VFL	Combining all features from a group of selected nodes S_t to increase features dimension	7	Not Available	Not Available	Cha et al., 2021 [56]
FTL	Improve the model performance with small data size and unlabeled samples	7	Quantity Skew	Balancing the training dataset	Chen et al., 2020 [20]

HFL: horizontal federated learning; VFL: vertical federated learning; FTL: federated transfer learning.

4.4. Data Privacy Attacks and Protections

Data security and privacy are critical issues in medical applications. In FL, it is usual for all nodes to calculate and upload their local model weights and parameters to an aggregate server. The steps of uploading and processing the weights and parameters may leak sensitive patient information contained in the medical data. The possible attacks include model inversion and membership inference attacks, which may leak patient data to an attacker. The common solutions for data privacy protection include differential privacy and homomorphic encryption [21] based techniques, which can guarantee the security of transferring the local weights and parameters in federated learning. In the following subsection, we describe the possible data privacy attacks and protections in FL.

4.4.1. Data Privacy Attacks on Federated Learning

There are two types of possible data privacy attacks on federated learning. The first attack is trying to recreate the input data, such as model inversion attack, and the second attack is to discover the training data such as membership inference attack.

Model Inversion (MI) Attack. The model inversion attack is an attack method for recreating data on which a machine learning model was trained [60]. In the case of federated learning for healthcare applications, this can leak the sensitive patient data used in the model's training process. Fredrikson et al. [60] demonstrated the MI attack that, given the machine learning model and several demographic data about a patient, an attacker could generate the patient's genetic markers. Specifically, the attack exploits the predicted output probability confidence score from the machine learning model when predicting the class given the features data. Given a machine model learning model as a function $\hat{y} = f(w; x_1, \ldots, x_n)$ where $\hat{y}$, w, and $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ are predicted probability class, machine learning parameters, and features vector as an input, respectively. The model inversion attack aims to exploit a sensitive feature, for instance feature x_1, given some information about the other features $x_2, \ldots, x_n$ and the predicted output probability $\hat{y}$. One solution to overcome this threat is to use differential privacy mechanism which can be incorporated into the learning process to protect the data from inversion attacks, such as inferring model weights (discussed in Section 4.4.2).

Membership inference attack. Given a machine learning model $f(w; x_1, \ldots, x_n)$ and some sample instances, the membership inference attack task tries to discover whether the instance exists or not in the training dataset [61]. Membership inference attack poses a significant privacy issue as the membership can expose a person's private information. For instance, determining a person's presence in a hospital's clinical trial training dataset indicates that this patient was once a patient at the hospital. The patient and the hospital are the two key parties interested in defending against such membership inference attacks. The patients consider their memberships confidential and do not wish for their sensitive information to be made public. At the same time, the hospital does not want

to be prosecuted for leaking patient data. Almadhoun et al. [62] demonstrated the first membership inference attack in the medical area that infers the personal information of the participants in a genomic dataset. Truex et al. [63] showed the threats of membership inference attack when the attacker is a member in the FL environment. The member could be the aggregation service or one of the client nodes. Their FL configuration is different from the one discussed above. Instead of pooling the weights to construct a new global model, each node trains their local model and contributes just the prediction probability when inferring a new instance. The process of membership inference attack consists of three steps [61]. Firstly, the attackers aim to develop a shadow dataset $\mathcal{D}'$ that mimics the target model training dataset $\mathcal{D}$. Secondly, the attacker create a shadow model using the shadow dataset $\mathcal{D}'$ which mimics the target model behavior. In this step, the attacker observed the shadow model behavior in response to instances known to have been provided during training against those that were not. This behavior is utilized to create an attack dataset that captures the different instances in the training data and data that have not been seen previously. Finally, this attack dataset is used to construct a binary classifier that predicts whether an instance was previously used in the target model output.

4.4.2. Data Privacy Protections for Federated Learning

There are two methods to protect data privacy from data leakage and attacker in the FL environment: perturbation and encryption. The perturbation method preserves private data and model privacy by adding a controlled random noise to the training data or the machine learning model parameters during the training process. For instance, differential privacy [18,43,44,55] and hybrid exchange parameters [39] algorithms are the perturbations techniques implemented in the FL studies published in medical applications. In comparison, the encryption method preserves private data and model privacy by encrypting the parameters exchanged and the gradients in the aggregation process in the FL environment, such as the homomorphic encryption algorithm [20,21,51].

Data Privacy Protections with Differential Privacy (DP) Method. Combining a deep learning model with privacy protection is an emerging research focus. For instance, many researchers use differential privacy (DP) methods to secure the deep learning model. Inspired by the successfully implemented DP in centralized learning, several researchers implemented DP in distributed training, especially in FL studies for medical application [18,43,44,55]. Dwork et al. [64] introduced differential privacy as a notion of privacy, ensuring that data analytics do not compromise privacy. It ensures that the effect of an individual's data on the model output is restricted. In other words, differential privacy aims for an algorithm's result to be nearly identical whether or not the dataset contains data about a specific individual. This technique can prevent the membership inference attack where the attacker tries to find if a specific individual is in the training dataset. Differential privacy is achieved by adding controlled statistical noise to the machine learning model's input or output. Whereas the addition of noise ensures that specific individual data contributions are hidden, it also provides insights into the entire population without compromising privacy. The quantity of added noise is called the privacy budget denoted by epsilon (ϵ). Gaussian and Laplace are two controlled random noise mechanisms implemented in differential privacy for the FL studies in medical applications. Differential privacy with Gaussian noise mechanism is a common technique used in FL studies [18,43,44,55]. For instance, in their training dataset, Li et al. [18] and Vaid et al. [55] incorporated the Gaussian noise in the model learning process to protect from model inversion attacks. In addition, Zhang et al. [44] and Yan et al. [43] proposed a differential privacy technique with a generative adversarial network (DPGAN) to generate private data samples at a medical node in a federated environment. Specifically, Zhang et al. [44] implemented controlled noise to the gradient value in the discriminator part of their generated adversarial network (GAN) for image sampling in federated learning, interfering with original data distribution. Their experiments showed that this method could address the lack of data availability and the non-IID issue in FL while keeping patient

data private. In addition, Zhang et al. [44] evaluated that the smaller the Gaussian noise as part of DP will improve the model performance. Besides the Gaussian noise mechanism, differential privacy with the Laplace noise mechanism is implemented by Li et al. [18] in their studies. Li et al. [18] showed when the Laplace noise level was too high the deep learning model performance failed to classification task.

Data Privacy Protection with Homomorphic Encryption Method. Homomorphic encryption (HE) was used to ensure data privacy by encrypting the parameter exchanged in the gradient aggregation process. There are many recent FL studies for healthcare application that implemented HE during FL training [20,21,51]. Homomorphic encryption was categorized into fully homomorphic encryption (FHE) and additively homomorphic encryption (AHE) [65]. An FHE scheme is an encryption method that allows analytical functions to be run directly on the encrypted data while producing the same encrypted output as if the functions were executed in plaintext. In other words, if we perform an add or multiply operation on the ciphertext, the decryption result is the same as the actual result obtained by performing the same operation on the plaintext. In comparison, the AHE is an encryption method that allows only one type of operation to be run directly on the encrypted data and produces the same encrypted output as if the functions were executed in plaintext. In other words, the AHE scheme is intended for use with specific applications that require simple addition or multiplication operations. Formally, an encryption method is called homomorphic over an operation "+" if it supports Equation (10):

$$E\langle w_1\rangle + E\langle w_1\rangle = E\langle w_1 + w_2\rangle \ \forall w_1, w_2 \in W \tag{10}$$

where $E\langle . \rangle$ is the encryption method and W is the machine learning model parameters. For instance, in the AHE scheme, for parameters w_1 and w_2, one can obtain $E\langle w_1 + w_2\rangle$ by using $E\langle w_1\rangle$ and $E\langle w_2\rangle$ without knowing w_1 and w_2 explicitly. Most of the FL studies for healthcare applications leverage the AHE rather than the FHE since FHE is computationally more expensive than AHE. For example, Chen et al. [20] and Wu et al. [21] incorporated the AHE in their local model parameters sharing and gradient aggregation between healthcare nodes and the aggregate server. Xue et al. [51] adopted two AHE schemes for a lightweight privacy module to prevent the patient EMRs' privacy leakage in the medical edge devices.

4.5. Benchmark Medical Dataset for Federated Learning

The dataset utilized in FL studies can vary depending on the task. For instance, some datasets concentrate on the performance of classification tasks, while others concentrate on segmentation tasks. There are datasets such as LEAF [66] and FedVision [67] for FL algorithm benchmarking. However, there is no specific open public medical dataset for FL algorithm benchmarking due to limited quantity, patient security, and privacy. Therefore, a comprehensive list of relevant medical datasets is compiled from published FL papers for future research on this topic. From 24 published FL papers in the healthcare area, 16 publications used the public dataset listed in Table 3. We exclude eight publications from the list because these papers use their institution/private dataset.

Besides benchmark medical datasets for federated learning, numerous scientific research communities and industries have developed various tools to accelerate the growth of federated learning. We summarized in Table 4 the federated learning tools based on data configuration challenges.

Table 3. Summary of public medical datasets in recent FL studies applied for a medical area for algorithm benchmarking.

	Dataset Type	Dataset Name	Description	FL Study
Healthcare dataset	Medical Image Classification	Autism Brain Imaging Data Exchange (ABIDE) I [68]	The ABIDE I is a consortium dataset openly sharing 1112 functional magnetic resonance imaging (fMRI) dataset from 539 patients with autism spectrum disorders.	[18]
		Public COVID-19 Image Data Collection [69]	The dataset consists of 108 healthy chest X-ray images and 108 confirmed with COVID-19 chest X-ray images taken from 76 patients.	[40,44,54]
		Facial Emotion Recognition (FER) 2013 [70]	The FER2013 dataset consists of 35,887 human facial emotion images. The dataset is labeled into seven emotions: neutral, anger, disgust, sadness, happiness, surprise, and fear.	[37]
	Medical Image Segmentation	Brain Tumor Image Segmentation Benchmark (BraTS) 2017 and 2018 [71]	The BraTs 2017 were collected from 13 institutions and consisted of 359 patients' brain tumor scans.	[38]
		SPIE-AAPM PROSTATEx dataset [72]	The PROSTATEx dataset consists of 343 MRI prostate image cancer from Siemens 3T MR scanners, the MAGNETOM Trio, and Skyra.	[43,50]
	Electronic Health Record	MobiAct [73]	The MobiAct dataset is human activity dataset taken from 57 volunteers (42 men and 15 women).	[21]
		Human Activity Recognition (HAR) [74]	The HAR dataset was collected from 30 volunteers. Each subject performed different activities such as walking, sitting, standing, and laying. There are 10,299 with 561 time-series features.	[20,45]
		WESAD (Wearable Stress and Affect Detection) [75]	The WESAD is a dataset for wearable effect and stress detection. Taken from 15 participants, the WESAD consists of 12 features with 63,000,000 time-series samples.	[42]
		Medical Information Mart for Intensive Care (MIMIC) III [76]	The MIMIC III dataset was collected from 40,000 patients during stayed in the ICU at Beth Israel Deaconess Medical Center between 2001 and 2012.	[46]
		The eICU collaborative research database. [77]	Critical care datasets consist of 200,859 patients data from 208 hospitals in the United States.	[19,39,56]
Nonhealthcare dataset	Image classification, sentiment analysis	LEAF Dataset [66]	The LEAF Dataset Benchmarking framework consists of images and text datasets such as EMNIST, Celeba, Shakespeare, and Synthetic datasets.	[66]
	Image Classification	FedVision—Real World image dataset for FL [67]	The FedVision dataset contains more than 900 real-world images generated from 26 street cameras. Precisely, it consists of 7 classes with a detailed bounding box. This dataset has non-IID properties reflecting a real-world data distribution.	[67]

ABIDE: autism brain imaging data exchange; BraTS: brain tumor image segmentation benchmark; eICU: electronic intensive care unit; FER: facial emotion recognition; FL: federated learning; fMRI: functional magnetic resonance imaging; HAR: human activity recognition; MIMIC: medical information mart for intensive care; MR: magnetic resonance; IID: independent and identical data distribution; WESAD: wearable stress and affect detection.

Table 4. Federated learning tools.

Framework Name	Creator	Supported Techniques				URL
		Data Partition	Data Distribution	Data Privacy Attack Simulation	Data Privacy Protection Methods	
PySyft	Open Mined	✓ HFL, VTL	✓ IID, non-IID	×	✓ DP, HE	https://github.com/OpenMined/PySyft (accessed on 7 July 2021)
TFF	Google	✓ HFL	×	×	×	https://www.tensorflow.org/federated (accessed on 7 July 2021)
FATE	Tencent	✓ HFL, VFL, FTL	×	×	✓ HE	https://github.com/FederatedAI/FATE (accessed on 21 July 2021)
Sherpa.ai	Sherpa.ai	✓ HFL	✓ IID, non-IID	✓ Data Poison	✓ DP	https://developers.sherpa.ai/privacy-technology/ (accessed on 27 August 2021)
LEAF	Sebastian Caldas	✓ HFL	×	×	×	https://leaf.cmu.edu/ (accessed on 21 July 2021)

HFL: horizontal federated learning; VTL: vertical transfer learning; FTL: federated transfer learning; IID: independent and identically data distribution; DP: differential privacy; HE: homomorphic encryption.

4.6. FL Studies for Healthcare Applications

Published FL studies in medical applications mostly come with two tasks: classification and segmentation, as summarized in Table 5. In our selected papers, there are 24 studies. Out of these studies, 21 studies are on classification tasks, and three are on segmentation tasks. The following subsections describe the existing studies on FL for healthcare applications, organized by the application task type.

4.6.1. Classification Task in FL for Healthcare Applications

Classification is a common task tackled in the published FL applications in the medical domain. In machine learning, classification algorithms learn how to classify or annotate a given set of instances with classes or labels. There are several classification tasks that are studied in federated learning setting in healthcare, e.g., autism spectrum disorder (ASD) [18], cancer diagnosis [41,43,49], COVID-19 detection [40,44,48,54], human activity and emotion recognition [20,21,37,42,45], patient hospitalization prediction [52], patient mortality prediction [19,39,46,47,55,56], and sepsis disease diagnosis [51]. The summary of classification tasks in FL studies for medical application is listed in Table 5.

Cancer diagnosis. Recent studies show that researchers are employing FL technology to develop machine learning models for cancer diagnostic applications [41,43,49]. For instance, Lee et al. [27] proposed a CNN-based model to classify whether thyroid nodules were benign or malign. The training data were 8457 ultrasound images collected from six institutions. The results show that the performance of the FL-based method was comparable with centralized learning with accuracy, sensitivity, and specificity of 97%, 98%, and 95%, respectively. Similarly, Rajendran et al. [49] implemented FL with an MLP model for lung cancer classification using two independent cloud providers. The model initialized, trained, and transferred from one node to another node using a cloud repository. The model achieved 92.8% accuracy to classify cancer. Another study by Yan et al. [43] transformed all nodes' raw medical image data onto a common space via image-to-image

translation without violating FL's privacy settings. The image-to-image translation was done using a cycle generative adversarial network (CycleGAN) model. The performance of the proposed method trained with eight medical nodes achieved 98% accuracy and 99% area under the curve (AUC) to classify prostate cancer.

Table 5. Summary of FL publications applied in medical applications.

ML Task	Clinical Tasks	Medical Input Data	Model Architecture	FL Study
Classification	Autism spectrum disorders (ASD) or Healthy control (HC)	fMRI	CNN	[18]
	Cancer diagnosis:			
	- Prostate cancer - Thyroid cancer - Lung cancer	- MRI - Ultrasound images - Tobacco and radon data	- GAN - CNN - MLP	[43] [41] [49]
	COVID-19 detection	X-ray images	CNN	[40,44,48,54]
	Human activity	Wearable device	LSTM	[20,21,45]
	Human emotion	Wearable device	CNN	[37,42]
	Patient hospitalization	Patient EHR	SVM	[52]
	Patient mortality	Critical care data	MLP	[19,39,46,47,55,56]
	Sepsis disease	Patient EHR	Double Deep Q Network	[51]
Segmentation	Brain tumor	MRI	U-Net	[38]
	COVID-19 region	3D Chest CT	3D U-Net	[53]
	Prostate cancer	MRI	3D Anisotropic Hybrid Network	[50]

CNN: convolutional neural network; CT: computed tomography; EHR: electronic health record; fMRI: functional magnetic resonance imaging; GAN: generative adversarial network; MLP: multilayer perceptron; MRI: magnetic resonance imaging; SVM: support vector machine.

COVID-19 detection. For COVID-19 detection applications [40,44,48,54], FL is a potential approach for connecting medical images data from medical institutions, enabling them to develop a model while maintaining patient privacy. In this case, the model's performance is considerably enhanced from diverse medical datasets from several institutions. For instance, Abdul Salam et al. [54] experimented with different federated learning architectures for binary COVID-19 classification. Their results showed that the federated learning model with GAN architecture and stochastic gradient descent (SGD) optimizer had a higher accuracy while keeping the loss score lower than the centralized machine learning model. The model performance achieved accuracy and AUC of 98.30% and 9.63%, respectively. Similarly, Dou et al. [48] showed the efficacy of a federated learning system for detecting COVID-19-related CT anomalies using patients' medical data from one country hospital as training data, then validating the model with medical data from other countries. Specifically, the authors trained an MLP-model using 132 patients from three hospitals in Hong Kong and validated the model generalizability performance using the medical dataset from China and Germany. The system achieved 83.12% in terms of AUC. Feki et al. [40] showed that increasing the number of medical nodes will decrease the training round for the model to converge and increase the model performance in CT–X-ray COVID-19 prediction. The authors proposed the CNN-based model architecture and achieved a performance of 95.27% AUC score. Similar results were obtained by Zhang et al. [44], who proposed an FL framework that enables medical nodes to generate high-quality training data samples with a privacy-protection approach. Specifically, the proposed method solves

the challenge of lacking COVID-19 medical training data in a federated environment. The GAN-based architecture was employed in the proposed system and achieved a comparable performance of 94.11% accuracy.

Human activity and emotion recognition. With increasing research on wearable technology and the internet of health things (IoHT), FL technology is one of the solutions to keep users' privacy while collaborating to develop a model for human activity and emotion recognition [20,21,37,42,45]. For example, Chen et al. [20] developed a deep learning model for human activity classification such as walking, sitting, standing, and laying. Then the author elaborates the trained CNN-based model with federated transfer learning to achieve a personalization model for each edge device. The system achieved 99.4% accuracy in classifying human activities. Similarly, Wu et al. [21] developed a cloud-edge federated learning infrastructure to create a patient privacy-aware deep learning model for in-home monitoring applications. The authors developed an autoencoder (AE) model architecture then deployed the model into five different healthcare nodes. The FL system achieved an accuracy of 95.41%. Chhikara et al. [37] combined the speech signal and facial expression to create an emotion index for monitoring the patient's mental health. Using the facial emotion recognition (FER) dataset collected from several data silos, the author employed a federated learning technique and AE-based architecture to create a secure machine learning model to classify a human emotion. The FL algorithm showed an AUC of 88%.

Patient mortality prediction. Similarly, FL enables early predictive modeling based on several sources, which can help to assist clinicians with extra information into the risks and benefits of treating patients earlier [19,46,47,51,52,55,56]. Huang et al. [19] used drug features to forecast critical care patients' mortality, and ICU stays time. Their algorithm based on AE architecture also addresses non-IID ICU patient data by grouping patients into clinically significant communities with shared diagnoses and geographical regions, then training one model per community. The proposed FL algorithm showed an AUC of 69.13%. In a similar study, Brismi et al. [52] proposed a method to forecast future patient hospitalizations with heart-related disorders by solving the L1-regularized sparse support vector machine (SVM) classifier in a federated learning environment. The proposed FL model performed an AUC of 77.47%. Shao et al. [47] proposed an MLP-based model framework to predict in-hospital mortality among patients admitted to the intensive care unit. Their findings indicate that training the model in a federated learning framework produces outcomes comparable to those obtained in a centralized learning environment with an AUC of 97.76%. Vaid et al. [55] demonstrated federated learning with an MLP-based model architecture to predict patient mortality with COVID-19 disease within seven days. Their experiment showed that the federated learning algorithm successfully produces a robust predictive model while preserving the patient's confidential information with an 82.9% AUC score.

Other healthcare areas. Besides the healthcare areas mentioned above, FL also applied for sepsis disease [51] and autism spectrum disorder classification [18]. Xue et al. [51] developed a fully decentralized federated framework (FDFF) that integrates a neural network model across edge devices to extract knowledge from internet-of-things for healthcare applications. The edge devices using FDFF can create a double deep Q-network (DDQN) that gives suggestions for sepsis treatment. In addition, Li et al. [18] proposed FL for multisite autism spectrum disorder (ASD) fMRI analysis.

4.6.2. Segmentation Task in FL for Healthcare Applications

Segmentation tasks with medical images have become an essential clinical task in healthcare applications. The medical image segmentation task is the process of identifying and selecting a region of interest within a medical image. Medical images can be in several forms, such as MRI or CT image scan. There are several published FL studies in medical image segmentation, namely brain tumor disease [38], COVID-19 region [53], and prostate cancer region [50]. The summary of published FL studies on segmentation tasks is listed in Table 5. Specifically, in brain tumor segmentation using brain MRI medical images, Sheller et al. [38]

applied the FL algorithm with CNN-based architecture for multi-institutional collaboration in brain tumor segmentation tasks while preserving the patient data. Compared to existing collaborative learning approaches, FL achieved the highest dice score of 85% and scaled more effectively as the number of collaborating institutions increases. Using multinational three-dimensional chest CT images from three countries, Yang et al. [53] applied federated semi-supervised learning with 3D u-shape fully connected layer model architecture to segment the COVID-19 disease region. Federated semi-supervised learning can assure good training performance even when some healthcare sites have a limited number of annotated data compared to unannotated data. Additionally, the semi-supervised environment may alleviate some of the strain associated with expert annotation, which is critical given the present pandemic crisis. Similarly, Sarma et al. [50] performed prostate segmentation with a 3D anisotropic hybrid network (3D AH-Net) model on MRI with collaboration from industry, public universities, and the federal institution. The proposed FL algorithm experimented with three medical nodes showed a dice score of 88.9%.

RQ3: What are the research gaps and potential future research directions of FL related to medical data?

4.7. Open Challenges

In this survey, we review the current progress on federated learning in the healthcare field. We highlight the comprehensive solutions to federated learning issues related to medical data configurations to provide a valuable resource for researchers. In what follows, we list some potential research directions or open questions when federated learning is applied in the healthcare area.

FL with Medical Data Stream. Medical data streams are collections of medical data that increase constantly and rapidly over time, generated during the treatment and monitoring of patients. For instance, in telemedicine or patient monitoring, the medical monitoring devices generate a large amount of time-sensitive data when monitoring patient vital signs such as temperature, heart rate, and blood pressure. This medical data is a stream of medical signals displayed for interpretation by physicians. Certain pieces of these data could be used in real-time to alert physicians about changes in patient circumstances. Medical data streams arrive periodically, and we would like to develop an analytic model that extracts meaningful patterns or risk factors in real-time. Federated learning incorporated with the medical data stream could improve training tasks and security performance, as inconsistencies in evolving medical datasets and the data transmission between the FL coordinator and participant nodes can be highly decreased [25]. However, the medical data streams are usually fast, large, and we must handle them in real-time. In addition, the medical data streams are dynamic, so our FL algorithm has to respond to these changes. Thus, it is essential to design an efficient federated learning algorithm to achieve good accuracy, low total memory, and minimum time in medical data streams.

FL with Hybrid Medical Data Partition. In the HFL data partition, the nodes share the same features $\mathcal{X}$ and label $\mathcal{Y}$ but have different data samples $\mathcal{I}$. Thus, the HFL aims to solve limited sample size variability by combining data samples from all nodes when developing a model, while for the VFL data partition the nodes share the same data samples $\mathcal{I}$ but have different features $\mathcal{X}$ and labels $\mathcal{Y}$. Therefore, the VFL aims to enrich the features by combining features from all nodes when developing a model. However, we need to simultaneously solve a limited sample size variability and enrich the features when developing a model in practice. For instance, a healthcare node may possess either partial features or data samples in healthcare insurance, which serves only a fraction of users and only has partial records. Incorporating both the HFL and the VFL data partition will result in a hybrid data partition. Compared to the HFL and the VFL, a hybrid FL data partition has its challenges. In HFL, each node shares neither its local data nor labels. In contrast, in VFL, the node shares the user's index to the server or is securely stored in one node as a key for aligning the features [56]. A hybrid FL data partition needs to deal with both types of nodes, so the FL training algorithm can run without requiring the aggregate server to

access any data, including the users' index. New architecture and training algorithms in FL will be required to utilize the benefits of the hybrid data partition effectively.

FL with Incentive Mechanism for Good Data Contributor. The internet of health things (IoHT) uses internet of things (IoT) devices on e-health applications that enable the connection between healthcare resources and patients. The IoHT devices such as smartwatches and healthcare wearable trackers can record heart rate, body temperature, and blood pressure. These rich healthcare data are excellent for personal smartphone healthcare apps that can run on device federated learning. However, the IoHT nodes are burdened by significant computation and communication costs during the federated model training process. Without a proper incentive mechanism design, those IoHT nodes will be reluctant to participate in federated learning. In addition, a suitable incentive mechanism can have rewards and punishments. A good quality personal healthcare data contributor can obtain a good incentive, while harmful data contributors can receive a punishment. Thus, an effective and efficient incentive mechanism can attract good data contributors to join federated learning.

Limitation and future perspective. There are two limitations to the present study. The first limitation is that existing FL experiments focus exclusively on one of the non-IID properties, such as data imbalance or label skew. However, there are no comprehensive experiments in the medical dataset that examine multiple properties of non-IID. The future perspective will find additional algorithms for addressing the issues associated with hybrid non-IID features. The second limitation is the hyperparameter framework search for FL. Hyperparameter tuning is a critical yet time-consuming step in the machine learning workflow. Optimization of hyperparameters becomes considerably more difficult in federated learning, in which models are trained across a dispersed network of heterogeneous data silos. Thus, an automatic tool or framework to select the optimal hyperparameters in the FL model is critically needed in the future research.

5. Conclusions

We presented the advancement of federated learning growth in the context of healthcare applications over the last four years in terms of data properties such as data partition, data distribution, data privacy attack and protection, and benchmark datasets. We hope that this study stimulates additional research into FL in healthcare applications and eventually becomes a guideline for handling sensitive medical data. Several open challenges remain, including FL for the medical data stream, FL with medical data hybrid partitions, and incentive mechanisms for good medical data contributors. We envision the increased popularity of FL for medical purposes in the near future, resulting in more advanced protocols with security and privacy guarantees and the actual deployment of FL technology for solving real-world problems in the healthcare domain.

Author Contributions: Conceptualization, P. and Z.-Y.S.; methodology, P., Z.-Y.S. and C.-R.S.; software and validation, Y.-Y.T., K.T.P. and H.-C.C.; formal analysis, W.J. and K.S.M.T.H.; investigation, P. and K.T.P.; resources, P. and K.T.P.; data curation, P., Z.-Y.S. and C.-R.S.; writing—original draft preparation, P. and Z.-Y.S.; writing—review and editing, Z.-Y.S., W.J., Y.-Y.T., K.S.M.T.H. and C.-R.S.; visualization, P.; supervision, Z.-Y.S.; project administration, Z.-Y.S. and Y.-Y.T.; funding acquisition, Z.-Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of Taiwan under the grants MOST 110-2321-B-468-001 and MOST 110-2511-H-468-005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This study was supported by the Ministry of Science and Technology of Taiwan under the project grants MOST 110-2321-B-468-001 and MOST 110-2511-H-468-005. Furthermore, the

authors express their gratitude to the anonymous reviewers for their comments and recommendations, which significantly improved the original work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AI	Artificial intelligence
AMCA	American medical collection agency
AUROC	Area under the receiver operating characteristic curve
CNN	Convolutional neural network
CT	Computerized tomography image
DNN	Deep neural network
FCN	Fully connected network
FL	Federated learning
fMRI	Functional magnetic resonance image
GDPR	General data protection right
HIPAA	Health insurance portability and accountability act
IID	Independent and identical data distribution
LSTM	Long short-term memory
ML	Machine learning
MLP	Multilayer perceptron
MRI	Magnetic resonance image
PDPA	Personal data protection act
PHI	Protected health information

Appendix A

Table A1. Full query term in publication databases.

Scientific Database	Query	Studies Results #
PubMed	((federated learning AND ((fft[Filter]) AND (english[Filter]) AND (2018:2021[pdat]))) AND (healthcare OR hospital OR clinic AND ((fft[Filter]) AND (english[Filter]) AND (2018:2021[pdat]))) AND (“data quality” OR privacy protection OR non iid AND ((fft[Filter]) AND (english[Filter]) AND (2018:2021[pdat]))) AND ((fft[Filter]) AND (english[Filter])) AND ((fft[Filter]) AND (english[Filter]))	21
IEEE Xplore	(“All Metadata”:federated learning) AND (“All Metadata”:healthcare OR “All Metadata”:hospital OR “All Metadata”:clinic) AND (“All Metadata”:data quality OR “All Metadata”:privacy protection OR “All Metadata”:non iid)	14
Web of Science	”Healthcare OR Hospital OR Clinic” AND ”federated learning” AND ”Data Quality OR Privacy Protection OR non iid”	17
Science Direct	(“federated learning”) AND (healthcare OR hospital OR clinic) AND (“data quality” OR “privacy protection” OR “non iid”)	105
ACM Digital Library	[All: “federated learning”] AND [[All: healthcare] OR [All: clinic] OR [All: hospital]] AND [[All: “data quality”] OR [All: “privacy protection”] OR [All: “non iid”]] AND [Publication Date: (1 January 2018 TO 30 June 2021)]	40

Table A2. Federated learning studies for medical applications.

Authors	Year	Title	Journal	FL Studies
Brismi et al.	2018	Federated learning of predictive models from federated electronic health records	*International Journal of Medical Informatics*	[52]
Huang et al.	2019	Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records	*Journal of Biomedical Informatics*	[19]
Chen et al.	2020	FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare	*IEEE Intelligent Systems*	[20]
Chen et al.	2020	Communication-Efficient Federated Deep Learning With Layerwise Asynchronous Model Update and Temporally Weighted Aggregation	*IEEE Transactions on Neural Networks and Learning Systems*	[45]
Huang et al.	2020	LoAdaBoost: Loss-based AdaBoost federated machine learning with reduced computational complexity on IID and non-IID intensive care data	*PLOS ONE*	[46]
Li et al.	2020	Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results	*Medical Image Analysis*	[18]
Shao et al.	2020	Stochastic Channel-Based Federated Learning With Neural Network Pruning for Medical Data Privacy Preservation: Model Development and Experimental Validation	*JMIR Formative Research*	[47]
Sheller et al.	2020	Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data	*Scientific Reports*	[38]
Wu et al.	2020	FedHome: Cloud-Edge based Personalized Federated Learning for In-Home Health Monitoring	*IEEE Transactions on Mobile Computing*	[21]
Abdul Salam et al.	2021	COVID-19 detection using federated machine learning	*PLOS ONE*	[54]
Cha et al.	2021	Implementing Vertical Federated Learning Using Autoencoders: Practical Application, Generalizability, and Utility Study	*JMIR Medical Informatics*	[56]
Chhikara et al.	2021	Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications	*IEEE Internet of Things Journal*	[37]
Cui et al.	2021	FeARH: Federated machine learning with anonymous random hybridization on electronic medical records	*Journal of Biomedical Informatics*	[39]
Dou et al.	2021	Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study	*npj Digital Medicine*	[48]
Feki et al.	2021	Federated learning for COVID-19 screening from chest X-ray images	*Applied Soft Computing*	[40]
Lee et al.	2021	Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment	*JMIR Medical Informatics*	[41]
Liu et al.	2021	Learning From Others Without Sacrificing Privacy: Simulation Comparing Centralized and Federated Machine Learning on Mobile Health Data	*JMIR mHealth and uHealth*	[42]
Rajendran et al.	2021	Cloud-Based Federated Learning Implementation Across Medical Centers	*JCO Clinical Cancer Informatics*	[49]

Table A2. *Cont.*

Authors	Year	Title	Journal	FL Studies
Sarma et al.	2021	Federated learning improves site performance in multicenter deep learning without data sharing	*Journal of the American Medical Informatics Association*	[50]
Vaid et al.	2021	Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach	*JMIR Medical Informatics*	[55]
Xue et al.	2021	A Resource-Constrained and Privacy-Preserving Edge-Computing-Enabled Clinical Decision System: A Federated Reinforcement Learning Approach	*IEEE Internet of Things Journal*	[51]
Yan et al.	2021	Variation-Aware Federated Learning with Multi-Source Decentralized Medical Image Data	*IEEE Journal of Biomedical and Health Informatics*	[43]
Yang et al.	2021	Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan	*Medical Image Analysis*	[53]
Zhang et al.	2021	FedDPGAN: Federated Differentially Private Generative Adversarial Networks Framework for the Detection of COVID-19 Pneumonia	*Information Systems Frontiers*	[44]

References

1. Feng, Y.; Zhang, L.; Mo, J. Deep manifold preserving autoencoder for classifying breast cancer histopathological images. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 91–101. [CrossRef] [PubMed]
2. McWilliams, A.; Beigi, P.; Srinidhi, A.; Lam, S.; MacAulay, C.E. Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 2044–2054. [CrossRef]
3. Chen, S.; Yang, H.; Fu, J.; Mei, W.; Ren, S.; Liu, Y.; Zhu, Z.; Liu, L.; Li, H.; Chen, H. U-Net Plus: Deep semantic segmentation for esophagus and esophageal cancer in computed tomography images. *IEEE Access* **2019**, *7*, 82867–82877. [CrossRef]
4. Ge, C.; Gu, I.Y.; Jakola, A.S.; Yang, J. Enlarged training dataset by pairwise GANs for molecular-based brain tumor classification. *IEEE Access* **2020**, *8*, 22560–22570. [CrossRef]
5. Sultan, H.H.; Salem, N.M.; Al-Atabany, W. Multi-classification of brain tumor images using deep neural network. *IEEE Access* **2019**, *7*, 69215–69225. [CrossRef]
6. Noreen, N.; Palaniappan, S.; Qayyum, A.; Ahmad, I.; Imran, M.; Shoaib, M. A deep learning model based on concatenation approach for the diagnosis of brain tumor. *IEEE Access* **2020**, *8*, 55135–55144. [CrossRef]
7. Xue, W.; Li, Q.; Xue, Q. Text detection and recognition for images of medical laboratory reports with a deep learning approach. *IEEE Access* **2020**, *8*, 407–416. [CrossRef]
8. Harerimana, G.; Kim, J.W.; Yoo, H.; Jang, B. Deep learning for electronic health records analytics. *IEEE Access* **2019**, *7*, 101245–101259. [CrossRef]
9. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 843–852.
10. Zech, J.R.; Badgeley, M.A.; Liu, M.; Costa, A.B.; Titano, J.J.; Oermann, E.K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **2018**, *15*, e1002683. [CrossRef]
11. O'Leary, D.E. Embedding AI and crowdsourcing in the big data lake. *IEEE Intell. Syst.* **2014**, *29*, 70–73. [CrossRef]
12. Moore, W.; Frye, S. Review of HIPAA, part 1: History, protected health information, and privacy and security rules. *J. Nucl. Med. Technol.* **2019**, *47*, 269–272. [CrossRef]
13. Mark Allen Group, Data breach at major healthcare firms. *Comput. Fraud. Secur.* **2019**, *2019*, 3–19. [CrossRef]
14. Voigt, P.; von dem Bussche, A. *The EU General Data Protection Regulation (GDPR)*; Springer: Cham, Switzerland, 2017.
15. Laws and Regulations Database of the Republic of China. Personal Data Protection Act. 2015. Available online: https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=I0050021 (accessed on 7 July 2021).
16. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics Conference, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
17. Hard, A.; Rao, K.; Mathews, R.; Ramaswamy, S.; Beaufays, F.; Augenstein, S.; Eichner, H.; Kiddon, C.; Ramage, D. Federated learning for mobile keyboard prediction. *arXiv* **2019**, arXiv:1811.03604.

18. Li, X.; Gu, Y.; Dvornek, N.; Staib, L.H.; Ventola, P.; Duncan, J.S. Multi-site FMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **2020**, *65*, 101765. [CrossRef] [PubMed]
19. Huang, L.; Shea, A.L.; Qian, H.; Masurkar, A.; Deng, H.; Liu, D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* **2019**, *99*, 103291. [CrossRef]
20. Chen, Y.; Qin, X.; Wang, J.; Yu, C.; Gao, W. FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.* **2020**, *35*, 83–93. [CrossRef]
21. Wu, Q.; Chen, X.; Zhou, Z.; Zhang, J. FedHome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans. Mobile Comput.* **2020**. [CrossRef]
22. Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M.J.; et al. Privacy-preserving federated brain tumour segmentation. In *Machine Learning in Medical Imaging*; Suk, H.-I., Liu, M., Yan, P., Lian, C., Eds.; Springer: Cham, Switzerland, 2019; Volume 11861, pp. 133–141.
23. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated learning with non-IID data. *arXiv* **2018**, arXiv:1806.00582. [CrossRef]
24. Hsieh, K.; Phanishayee, A.; Mutlu, O.; Gibbons, P.B. The non-IID data quagmire of decentralized machine learning. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020), Virtual Event, 13–18 July 2020.
25. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *arXiv* **2019**, arXiv:1912.04977.
26. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
27. Mothukuri, V.; Parizi, R.M.; Pouriyeh, S.; Huang, Y.; Dehghantanha, A.; Srivastava, G. A survey on security and privacy of federated learning. *Future Gener. Comput. Syst.* **2021**, *115*, 619–640. [CrossRef]
28. Wu, Q.; He, K.; Chen, X. Personalized federated learning for intelligent IoT applications: A cloud-edge based framework. *IEEE Open J. Comput. Soc.* **2020**, *1*, 35–44. [CrossRef] [PubMed]
29. Du, Z.; Wu, C.; Yoshinaga, T.; Yau, K.-L.A.; Ji, Y.; Li, J. Federated learning for vehicular internet of things: Recent advances and open issues. *IEEE Open J. Comput. Soc.* **2020**, *1*, 45–61. [CrossRef]
30. Putra, K.T.; Chen, H.-C.; Prayitno; Ogiela, M.R.; Chou, C.-L.; Weng, C.-E.; Shae, Z.-Y. Federated compressed learning edge computing framework with ensuring data privacy for PM2.5 prediction in smart city sensing applications. *Sensors* **2021**, *21*, 4586. [CrossRef]
31. Xu, J.; Glicksberg, B.S.; Su, C.; Walker, P.; Bian, J.; Wang, F. Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **2020**, *5*, 1–19. [CrossRef] [PubMed]
32. Pfitzner, B.; Steckhan, N.; Arnrich, B. Federated learning in a medical context: A systematic literature review. *ACM Trans. Internet Technol.* **2021**, *21*, 1–31. [CrossRef]
33. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
34. PRISMA. PRISMA Endorsers. Available online: http://www.prisma-statement.org/Endorsement/PRISMAEndorsers (accessed on 21 November 2021).
35. McDonagh, M.; Peterson, K.; Raina, P.; Chang, S.; Shekelle, P. Avoiding bias in selecting studies. In *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*; Agency for Healthcare Research and Quality: Rockville, MD, USA, 2008.
36. Scherer, R.W.; Saldanha, I.J. How should systematic reviewers handle conference abstracts? A view from the trenches. *Syst. Rev.* **2019**, *8*, 264. [CrossRef]
37. Chhikara, P.; Singh, P.; Tekchandani, R.; Kumar, N.; Guizani, M. Federated learning meets human emotions: A decentralized framework for human–computer interaction for IoT applications. *IEEE Internet Things J.* **2021**, *8*, 6949–6962. [CrossRef]
38. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R.; et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 12598. [CrossRef]
39. Cui, J.; Zhu, H.; Deng, H.; Chen, Z.; Liu, D. FeARH: Federated machine learning with anonymous random hybridization on electronic medical records. *J. Biomed. Inform.* **2021**, *117*, 103735. [CrossRef] [PubMed]
40. Feki, I.; Ammar, S.; Kessentini, Y.; Muhammad, K. Federated learning for COVID-19 screening from chest X-ray images. *Appl. Soft Comput.* **2021**, *106*, 107330. [CrossRef] [PubMed]
41. Lee, H.; Chai, Y.J.; Joo, H.; Lee, K.; Hwang, J.Y.; Kim, S.-M.; Kim, K.; Nam, I.-C.; Choi, J.Y.; Yu, H.W.; et al. Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment. *JMIR Med. Inform.* **2021**, *9*, e25869. [CrossRef]
42. Liu, J.C.; Goetz, J.; Sen, S.; Tewari, A. Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. *JMIR mHealth uHealth* **2021**, *9*, e23728. [CrossRef]
43. Yan, Z.; Wicaksana, J.; Wang, Z.; Yang, X.; Cheng, K.-T. Variation-aware federated learning with multi-source decentralized medical image data. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2615–2628. [CrossRef]

44. Zhang, L.; Shen, B.; Barnawi, A.; Xi, S.; Kumar, N.; Wu, Y. FedDPGAN: Federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. *Inf. Syst. Front.* **2021**. [CrossRef]
45. Chen, Y.; Sun, X.; Jin, Y. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 4229–4238. [CrossRef]
46. Huang, L.; Yin, Y.; Fu, Z.; Zhang, S.; Deng, H.; Liu, D. LoAdaBoost: Loss-based AdaBoost federated machine learning with reduced computational complexity on IID and non-IID intensive care data. *PLoS ONE* **2020**, *15*, e0230706. [CrossRef]
47. Shao, R.; He, H.; Chen, Z.; Liu, H.; Liu, D. Stochastic channel-based federated learning with neural network pruning for medical data privacy preservation: Model development and experimental validation. *JMIR Form. Res.* **2020**, *4*, e17265. [CrossRef] [PubMed]
48. Dou, Q.; So, T.Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H.H.C.; Yu, K.; et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study. *NPJ Digit. Med.* **2021**, *4*, 60. [CrossRef]
49. Rajendran, S.; Obeid, J.S.; Binol, H.; D'Agostino, R.; Foley, K.; Zhang, W.; Austin, P.; Brakefield, J.; Gurcan, M.N.; Topaloglu, U. Cloud-based federated learning implementation across medical centers. *JCO Clin. Cancer Inform.* **2021**, *5*, 1–11. [CrossRef]
50. Sarma, K.V.; Harmon, S.; Sanford, T.; Roth, H.R.; Xu, Z.; Tetreault, J.; Xu, D.; Flores, M.G.; Raman, A.G.; Kulkarni, R.; et al. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1259–1264. [CrossRef]
51. Xue, Z.; Zhou, P.; Xu, Z.; Wang, X.; Xie, Y.; Ding, X.; Wen, S. A resource-constrained and privacy-preserving edge-computing-enabled clinical decision system: A federated reinforcement learning approach. *IEEE Internet Things J.* **2021**, *8*, 9122–9138. [CrossRef]
52. Brisimi, T.S.; Chen, R.; Mela, T.; Olshevsky, A.; Paschalidis, I.C.; Shi, W. Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* **2018**, *112*, 59–67. [CrossRef] [PubMed]
53. Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H.R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Med. Image Anal.* **2021**, *70*, 101992. [CrossRef]
54. Abdul Salam, M.; Taha, S.; Ramadan, M. COVID-19 detection using federated machine learning. *PLoS ONE* **2021**, *16*, e0252573. [CrossRef]
55. Vaid, A.; Jaladanki, S.K.; Xu, J.; Teng, S.; Kumar, A.; Lee, S.; Somani, S.; Paranjpe, I.; De Freitas, J.K.; Wanyan, T.; et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. *JMIR Med. Inform.* **2021**, *9*, e24207. [CrossRef]
56. Cha, D.; Sung, M.; Park, Y.-R. Implementing vertical federated learning using autoencoders: Practical application, generalizability, and utility study. *JMIR Med. Inform.* **2021**, *9*, e26598. [CrossRef]
57. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
58. Hegde, H.; Shimpi, N.; Panny, A.; Glurich, I.; Christie, P.; Acharya, A. MICE vs. PPCA: Missing data imputation in healthcare. *Inform. Med. Unlocked* **2019**, *17*, 100275. [CrossRef]
59. Tran, K.; Bøtker, J.P.; Aframian, A.; Memarzadeh, K. Artificial intelligence for medical imaging. In *Artificial Intelligence in Healthcare*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 143–162.
60. Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; Ristenpart, T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. *Proc. USENIX Secur. Symp.* **2014**, *2014*, 17–32.
61. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 2017; pp. 3–18.
62. Almadhoun, N.; Ayday, E.; Ulusoy, Ö. Inference attacks against differentially private query results from genomic datasets including dependent tuples. *Bioinformatics* **2020**, *36*, i136–i145. [CrossRef] [PubMed]
63. Truex, S.; Liu, L.; Gursoy, M.E.; Yu, L.; Wei, W. Demystifying membership inference attacks in machine learning as a service. *IEEE Trans. Serv. Comput.* **2019**. [CrossRef]
64. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*; Springer: Heidelberg, Germany, 2006; pp. 265–284.
65. Acar, A.; Aksu, H.; Uluagac, A.S.; Conti, M. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.* **2018**, *51*, 1–35. [CrossRef]
66. Caldas, S.; Meher Karthik Duddu, S.; Wu, P.; Li, T.; Konečný, J.; McMahan, H.B.; Smith, V.; Talwalkar, A. LEAF: A benchmark for federated settings. *arXiv* **2018**, arXiv:1812.01097.
67. Luo, J.; Wu, X.; Luo, Y.; Huang, A.; Huang, Y.; Liu, Y.; Yang, Q. Real-world image datasets for federated learning. *arXiv* **2021**, arXiv:1910.11089.
68. Di Martino, A.; Yan, C.-G.; Li, Q.; Denio, E.; Castellanos, F.X.; Alaerts, K.; Anderson, J.S.; Assaf, M.; Bookheimer, S.Y.; Dapretto, M.; et al. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **2014**, *19*, 659–667. [CrossRef]
69. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597.

70. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing*; Springer: Heidelberg, Germany, 2013; pp. 117–124.
71. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef]
72. Litjens, G.; Debats, O.; Barentsz, J.; Karssemeijer, N.; Huisman, H. SPIE-AAPM PROSTATEx challenge data. *Cancer Imaging Arch.* **2017**. [CrossRef]
73. Vavoulas, G.; Chatzaki, C.; Malliotakis, T.; Pediaditis, M.; Tsiknakis, M. The MobiAct dataset: Recognition of activities of daily living using smartphones. In Proceedings of the International Conference on Information and Communication Technologies for Ageing Well and e-Health, Rome, Italy, 21–22 April 2016; pp. 143–151.
74. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 24–26 April 2013.
75. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 400–408.
76. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
77. Pollard, T.J.; Johnson, A.E.W.; Raffa, J.D.; Celi, L.A.; Mark, R.G.; Badawi, O. The EICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* **2018**, *5*, 180178. [CrossRef] [PubMed]