# Open problems in medical federated learning

Joo Hun Yoo
*Department of Artificial Intelligence, College of Computing and Informatics, Sungkyunkwan University, Suwon, Republic of Korea, and*

Hyejun Jeong, Jaehyeok Lee and Tai-Myoung Chung
*Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon, Republic of Korea*

Received 15 April 2022
Revised 14 June 2022
Accepted 14 June 2022

## Abstract

**Purpose** – This study aims to summarize the critical issues in medical federated learning and applicable solutions. Also, detailed explanations of how federated learning techniques can be applied to the medical field are presented. About 80 reference studies described in the field were reviewed, and the federated learning framework currently being developed by the research team is provided. This paper will help researchers to build an actual medical federated learning environment.

**Design/methodology/approach** – Since machine learning techniques emerged, more efficient analysis was possible with a large amount of data. However, data regulations have been tightened worldwide, and the usage of centralized machine learning methods has become almost infeasible. Federated learning techniques have been introduced as a solution. Even with its powerful structural advantages, there still exist unsolved challenges in federated learning in a real medical data environment. This paper aims to summarize those by category and presents possible solutions.

**Findings** – This paper provides four critical categorized issues to be aware of when applying the federated learning technique to the actual medical data environment, then provides general guidelines for building a federated learning environment as a solution.

**Originality/value** – Existing studies have dealt with issues such as heterogeneity problems in the federated learning environment itself, but those were lacking on how these issues incur problems in actual working tasks. Therefore, this paper helps researchers understand the federated learning issues through examples of actual medical machine learning environments.

**Keywords** Heterogeneity, Data security, Data privacy, Federated learning, Incentive mechanism, Medical application

**Paper type** Research paper

This research is supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, AI Graduate School Support Program(Sungkyunkwan University)) and funded by Institute of Information & Communications Technology Planning & Evaluation (IITP) grand funded by the Korea government(MSIT) (No.2020-0-00990, Platform Development and Proof High Trust & Low Latency Processing for Heterogeneous·Atypical·Large Scaled Data in 5G-IoT Environment).

International Journal of Web Information Systems
Vol. 18 No. 2/3, 2022
pp. 77-99
Emerald Publishing Limited
1744-0084
DOI 10.1108/IJWIS-04-2022-0080

## 1. Introduction

Machine learning has been widely studied in various research fields for its powerful performance in data analysis. It was possible to derive better results through machine learning methods by learning the hidden multi-dimensional characteristics of given data that were difficult for humans to distinguish. This structure of machine learning in the medical imaging field, where it is crucial to capture fine features in images, has been very helpful in strengthening the existing diagnostic approaches. For example, support vector machines, deep neural networks, convolutions and clustering techniques have been applied in the medical field to effectively search those human-unidentifiable correlations from medical data.

Through the active use of machine learning approaches, the medical field was able to expand its scope to specific medical fields such as radiology, pathology, neuroscience, genetics and even mental disorders. However, the biggest issue in the field of medical artificial intelligence (AI) is not the accuracy of diagnosis, but the protection of patients' personal information.

Federated learning, a machine learning algorithm based on the distributed data environment, has emerged under stricter data regulations laws around the world. When the concept of federated learning was first introduced, data privacy regulations such as the EU's General Data Protection Regulation, California's (CA's) Privacy Rights Act and China's Personal Information Protection were representative rules, but now more countries around the world are implementing efficient regulations, such as Brazil's Lei Geral de Prote,cão de Dados, Canada's Digital Charter Implementation and Singapore's Personal Data Protection Act, to protect their citizens' personal information. Thus, centralized machine learning methods, that collect and learn based on the proper amount of data, are no longer applicable under the personal data protection regulations. In particular, for medical data, researchers and business providers should follow the Health Insurance Portability and Accountability Act (HIPAA), which comprehensively protects the medical records and independently identifiable health information of patients and medical information providers. With numerous increasing data regulations, researchers have applied various solutions to prevent invasion of privacy.

First, the most common solutions to adopt for data privacy issues are to process and import variables that can identify individual users when collecting their data. Primary information leakage can be prevented through the measures such as secure aggregation, pseudonymization, data reduction, data suppression and data masking in normal data environments. However, personal health information (PHI) is difficult to apply these security methods, as it contains any format of information that can identify the data owner. PHI is a wider concept of personal identifiable information (PII), which means sensitive information such as health insurance records, medical numbers, health status, medical images and mental health records are included on top of basic user variables. Therefore, data security methods for PHI are difficult to completely protect the information and to use medical data efficiently and safely, structural solutions are required to learn without violations.

Federated learning is a structural solution for the existing data privacy violation problems of machine learning methods. It has a unique structure and characteristics compared to centralized machine learning. Traditional machine learning approaches require a large volume of training data collected from local data owners to the server for model generation. Federated learning, a decentralized learning structure, generates and develops deep neural network models without local data collection to the server. The core concepts used in the neural network model learning process are as below.

Individual clients' data stored in the local environment does not move, and the server generates the initial training model and delivers it to each participating client. The transferred initial training model goes through a model update process through learning within each client's data environment, and the server collects all of the corresponding results to create a better performance learning model. The most commonly used federated learning framework is achieved with the iteration of the above step. The structure of the federated learning process described above has considerable advantages in the application of machine learning in the medical field. Even if researchers want to create a better performance prediction model for patients visiting their hospitals, the work can be done without collecting or observing patients' data, and there is no need to share or infringe on patient data when cooperating between hospitals. Detailed explanations about basic federated learning algorithms and medical federated learning structures will be described in detail in the next section of the paper.

The paper is organized as follows. In Section 1, we describe the concept of federated learning and specific structures in medical environments. In Section 3, how federated learning is applied to the medical fields is described. Section 4 contains open problems in medical federated learning and existing solutions in the following order: data/system heterogeneity, client management, traceability and responsibility and security issues. Finally, the federated learning framework that our research team is currently working on is introduced with the functional explanations in section five.

## 2. Federated learning

This section summarizes two main categorizations of federated learning: the type of data and features and the overall structure of federated learning. First, we divided federated learning into Horizontal Federated Learning (FL), Vertical FL and Transfer FL, depending on the characteristics of the data and features. Second, we also categorized the overall architecture of FL into hospital-patient, hospital-hospital and combined architecture.

### *2.1 Federated learning structures*

The federated learning can be divided into three structures, horizontal, vertical and transfer, by data and feature partitions. Horizontal FL uses the data set with a shared feature space but with different sample spaces across the participating clients. Figure 1 visualizes the structure of the horizontal FL. Assuming that the clients are multiple hospitals, the feature space of the dataset might be similar to each other in terms of medical data. In this case, the global model can be trained collaboratively taking advantage of the data sets sharing the same feature space. Vertical FL, on the other hand, uses the data set with a shared sample space but with different feature spaces. Figure 2 visualizes the structure of the vertical FL. Bank statement data set and health information data set of the same group of people could be an example. This case might be beneficial in collaboratively training a global model by referring to the different or diverse categories of data of the same sample space. Transfer FL uses the data set with neither unshared sample and feature spaces, having a little overlap with both spaces. Assume that different

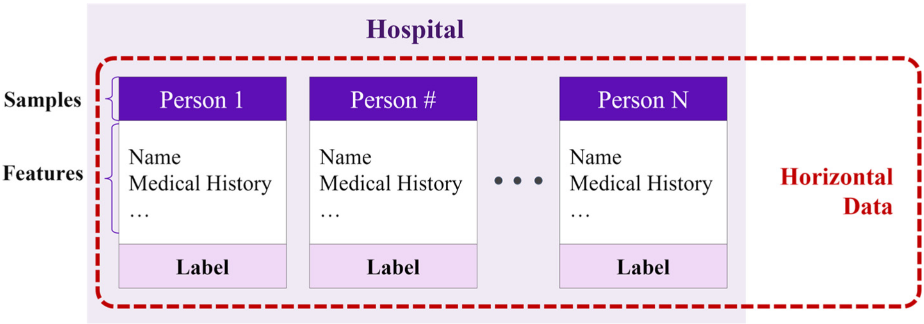


**Figure 1.** Horizontal federated learning

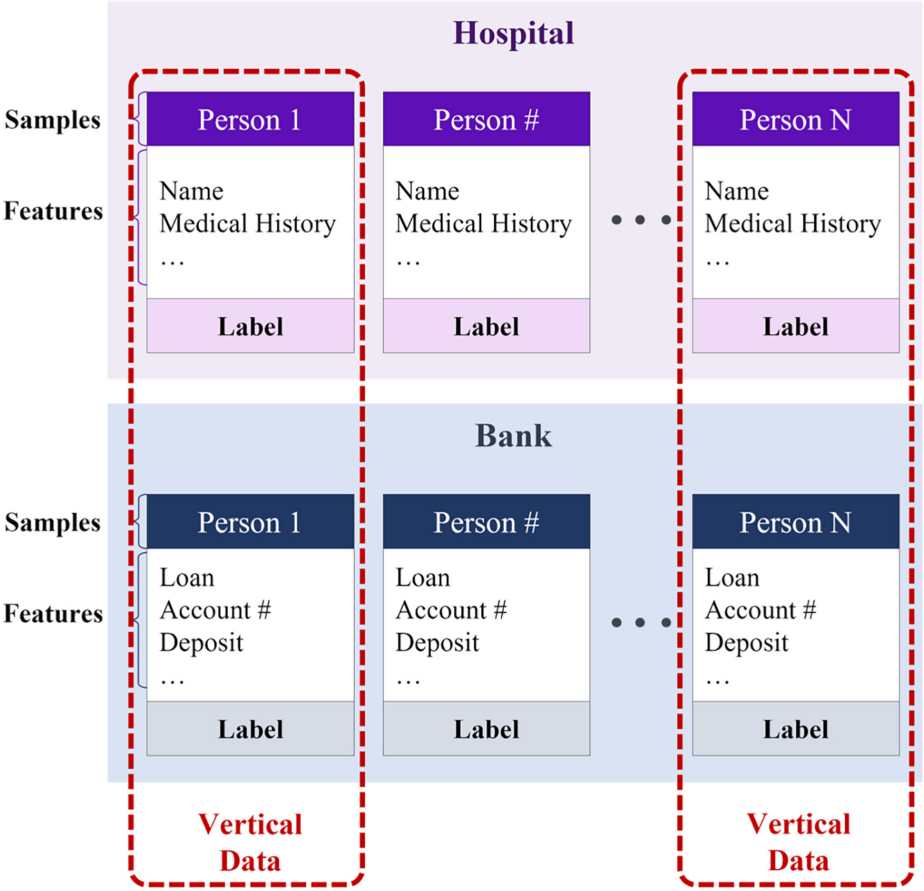


**Figure 2.**
Vertical federated learning

department of two physically distant hospitals participate in FL. In this case, the user spaces are different, as different patients from different hospitals, and the feature spaces are also minimally overlapped as distinct departments using different mechanism during the training phase.

### *2.2 Medical federated learning structures*

Federated learning is widely applied in health-care area, as it allows the application of various machine learning techniques without data collection from local agents. A server transmits its initial model to each client; hospital or patient. Each local client trains the received model, and then sends back the trained model parameters. Then, the server aggregates all the received parameters to update the global model; consequently, this collaborative and distributed learning can have the same effect as centralized learning. Most importantly, the global model is updated via aggregating multiple local models, so the local data privacy is isolated. Figure 3 describes three representative architectures of medical federated learning.

The figure on the upper left shows the most typical medical federated learning architecture, when patients are the local clients and the hospital works as a central server.

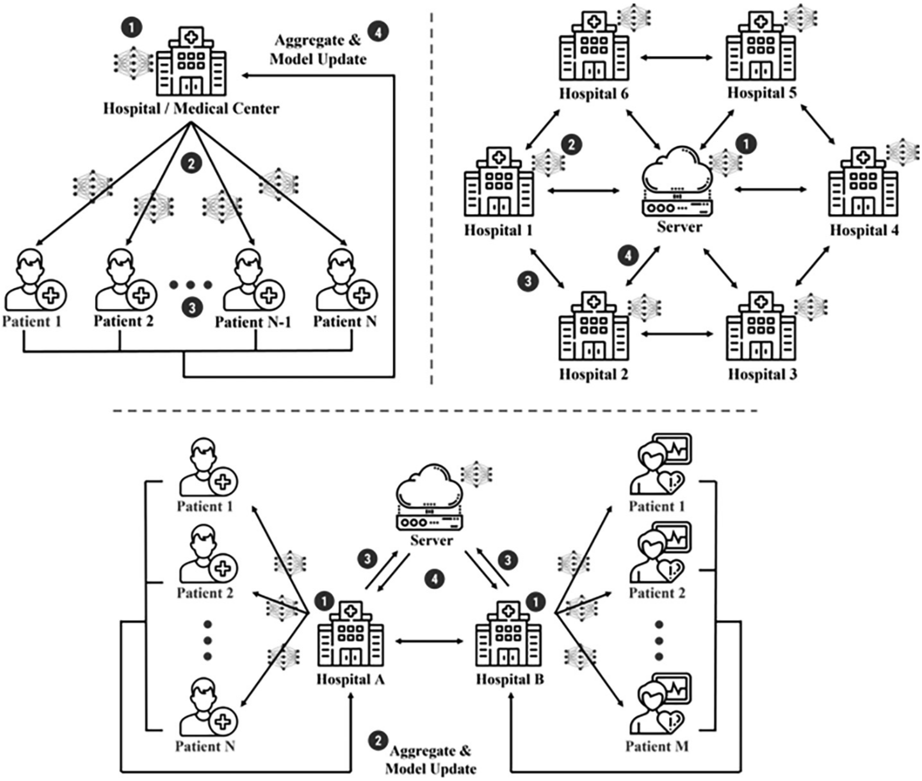


**Figure 3.**
Federated learning architecture in the field of medicine

The upper right figure is when individual hospitals or medical centers are participating as local clients, and a reliable server controls the whole architecture. Owkin connect, one of the most famous medical federated learning platforms also adopts this type of structure. Finally, the figure at the bottom shows the integrated design. While a reliable server controls the entire architecture, each hospital separately applies federated learning to its patient data. This structure has the advantage of training both global and local features to generate a higher performance model.

## 3. Federated learning in medical applications

When practitioners predict critical tasks using medical data, many data sets are required, and it is vital to use them to increase the accuracy of diagnosis or treatment without data privacy invasion. Using the structure of distributed machine learning (DML) allows the researchers to use data from various hospitals to build massive data sets. Then AI algorithm is applied to medical fields, resulting in performance improvement. However, transferring raw data caused sensitivity problems, and federated learning solved these challenges. Cases of improving performance have also increased by taking advantage of not providing all original data to the centralized server. This section describes papers using federated learning to utilize medical data. We summarize the articles using federated learning on the following categorizations; federated learning for Imaging, Signal and COVID-19.

### *3.1 Federated learning for imaging*

As the model's performance for image classification improves exponentially, machine learning plays a crucial role in predicting diseases or analyzing patients' conditions as a clinical decision support system. Using patients' computer tomography (CT) and magnetic resonance imaging images as data, machine learning models predict fatal diseases connected to the patient's life.

Linardos, for example, preprocessed M&M and ACDC Dataset as N4 via field correction showed higher accuracy than in the DML environment (Linardos *et al.*, 2022). They used federated learning with the ResNet model for hypertrophic cardiomyopathy diagnosis. The machine learning model became more robust when combined with federated learning. Furthermore, Kaissis used a pediatric pneumonia data set with ResNet to classify pediatric chest radiographs as viral pneumonia, bacterial pneumonia and normal (Kaissis *et al.*, 2021). They proved that federated learning performance was comparable to DML and even safer from inference attacks. In addition to simple applications, cases of proposing various methods for solving privacy problems are increasing as mentioned above. This will be further announced in Section 4.

As the model performances with unsupervised learning approaches are great, researchers tried the generative model for medical database generation. Li applied the generative adversarial network (GAN) algorithm on the Cardiac data set T. Li *et al.*, 2020b. When cardiac CT volumes imitating original real-world data were used to train the UNet-3D binary segmentation model, they found that the performance of federated learning was further improved. Also, in specific environments, the use of GAN has helped improve performance better for disease diagnosis.

In addition to improving the accuracy of federated learning, there are efforts to reduce communication costs and research when there is a distance between customers by using the network structure. Tedeschini focused on brain tumor segmentation using MQ telemetry transport for real-time networking in a federated learning environment (Tedeschini *et al.*, 2022). Experimental results showed that the federated learning model has a similar dice similarity coefficient value with distributed machine learning environment even in real-time conditions. The value of the dice similarity coefficient was calculated as the predicted and measured values.

### *3.2 Federated learning for signal*

Several hospitals diagnose patients' diseases using signal data and image data. Raza *et al.* (2022) uses the MIT beth israel hospital (MIT BIH) Arrhythmia database to learn a model by applying convolutional neural network architecture. They trained a model for classifying arrhythmias using electrocardiography data and showed that it has a higher performance than distributed machine learning settings. They also demonstrated remarkable results by modifying the model to apply explainable AI (XAI) to time-series data such as signal data.

Brophy used the generative adversarial network to explore the relationship between arterial blood pressure (ABP) and photoplethysmogram (PPG), then learned the time series to time series generative adversarial network model that generates ABP with PPG (Brophy *et al.*, 2021). As measuring ABP is expensive, they have produced significant results and seemingly similar results to DML.

Various researchers have used federated learning not just for medical applications but also for diagnosing diseases. Nandi and Xhafa (2022) proposed federated learning for real-time emotion state classification from multimodal data streaming, a real-time emotion classifier, using the dataset for emotion analysis using physiological and audiovisual recordings data set containing the physiological signals data. They conducted feature extraction and fusion with wavelet decomposition and trained the classifier using a three-layer feed-forward natural network. In addition, they compared the classification accuracy by increasing the number of clients used in federated learning and stated that FedREMCS could be adopted in various health-care fields based on these results. Moreover, Yoo generated a major depressive disorder severity classifier with heart rate variability data collected at Seoul Samsung Medical Center (Yoo *et al.*, 2021). They used the clustering-based federated learning method, Personalized Federated Cluster Model, to mitigate the nonidentically distributed (IID) problem and demonstrated higher accuracy compared to Federated Averaging.

### *3.3 Federated learning for COVID-19*

Since the outbreak of COVID-19, a worldwide pandemic disease in 2020, it has attracted attention as a new research topic in the medical AI field. Common symptoms in COVID-19 patients are lung tissue damage, which leads to cell destruction and pulmonary fibrosis. Chest X-rays or CT scan results are commonly examined to identify the phase of illness, and various research trials are being made to classify the collected medical images through machine learning algorithms. Through chest image analysis, COVID-19 symptoms are efficiently distinguished from pneumonia that are difficult to classify, and at the same time, personal information protection with federated learning techniques is also satisfied.

Zhang introduced a dynamic fusion-based federated learning algorithm to diagnose COVID-19 infections, using a set of medical image data sets collected from Kaggle and GitHub (W. Zhang *et al.*, 2021). An experiment was conducted with a structure in which three clients participated in learning, and test accuracy and convergence time were evaluated using three different training methods: GhostNet, ResNet50 and RestNet101. Liu *et al.* (2020b) newly applied a federated learning-based Covidnet algorithm to distinguish chest X-ray images of pneumonia patients and COVID-19 patients. Although it is lower than the ResNet models, it shows similar classification performance to MobileNet, a lightweight model and shows the possibility of medical imaging analysis through federated learning.

Unlike the previous two studies, Dayan's research team demonstrated the advantages of federated learning through the participation of multiple medical institutions for COVID-19 patient data analysis (Dayan *et al.*, 2021). A total of 20 institutions are participating in the creation of a federated learning medical data classification model, which is the same architecture as the concept shown on the right side of Figure 1. Compared to the case when generating a classifier with one local institution, using 20 institutions showed an average performance improvement of 13.9%. Kumar *et al.* (2021) also developed a framework that fused capsule networks and blockchain-based federated learning for diagnosis through lung CT imaging of patients collected in various hospitals. The datasets were collected from three hospitals, and the sensitivity and specificity performances improved by more than 7% than the existing benchmark machine learning models.

From the above experimental results, the performance degradation is not significant compared to centralized machine learning when federated learning algorithms consider PHI protection for COVID-19 diagnosis. Since centralized machine learning performed better when using the same amount of data for training, federated learning can employ more data to achieve higher performance with the advantage of training through distributed client data. Various mutant types of COVID viruses are emerging in countries around the world, and it will be important to form a federated learning-based diagnostic structure on data collected by each institution. While many studies in the medical field emphasize the benefits of applying federated learning, there are still research issues to be solved for better use.

## 4. Research issues

### *4.1 Heterogeneity issues*

An independent and independent and identically distributed (IID) data environment is the most commonly adopted assumption in machine learning. IID refers to an environment in which the data points for the system are independent and collected in an identical distribution rather than a skewed form. However, data distribution does not follow IID assumptions in most real data analysis environments, including the medical field. We classified these nonIID situations in medical federated learning as data heterogeneity and system heterogeneity.

*4.1.1 Data heterogeneity.* Data heterogeneity refers to the environment when data held by participating clients in federated learning has a heterogeneous data distribution or characteristics, which is also called nonIID data distribution. We can classify nonIID ness of data distribution into nonidentical and not independent distributions. Nonidentical data distribution is generally separated into five specific cases. Feature distribution skew, label distribution skew, same label but different features, same features but different labels and quantity skew are the five possible nonIID cases (Kairouz *et al.*, 2019).

Researchers should identify where their data environment belongs among the five cases and apply the heterogeneity issue solutions based on this. We provide specific examples and descriptions of the above five nonIID cases in a medical federated learning architecture in Table 1.

Besides nonidentical data distribution cases, not-independent distribution is a violation of the consistency of data depending on the other factors. Such violations are introduced when the data changes over time or geolocation ibid.

*4.1.2 System heterogeneity.* The heterogeneity of the data environment of federated learning participants should be considered first, and problems that may arise from their participating equipment should also be identified. Devices may cause system heterogeneity issues depending on their hardware setting, computer power, communication cost and network connectivity.

The federated learning with multiple medical centers participating in the training can cause differences in database or infrastructure between each hospital. A few hospitals, for example, are trying to generate a global machine learning prediction model for COVID-19 lung lesions. Samsung Medical Center has established a system infrastructure that can efficiently manage patient data by operating its database and digital therapeutics laboratory, and based on this, they intend to participate in the federated learning system.

| Non-IID case | Description and examples |
|---|---|
| Feature distribution skew | Marginal distributions of data features differ<br>ex) Even if two individuals wear the same smartwatch model and exercise for the same time duration, the features of measured values are unique due to the personal characteristics difference, such as their gait |
| Label distribution skew | Marginal distributions of data labels differ<br>ex) Frostbite is a disease that frequently occurs in cold areas because it is caused by exposure to severe cold resulting in tissue damage to body parts. Therefore, it is rare in places with relatively warm temperatures |
| Same label but different features | Conditional distributions of data features differ<br>ex) Medical devices are used to measure healthcare data such as neuroimages and biomarkers of patients. However, hospitals do not use the identical medical device brands |
| Same feature but different labels | Conditional distributions of data labels differ<br>ex) Lung imaged by the recent pandemic COVID-19 virus are difficult to distinguish from the pneumonia because they have similar features in many lesions |
| Quantity skew | Amount of each patients/hospital data differs<br>ex) Suppose five times more patients have visited hospital A than hospital B. The quantity of data each hospital has will also significantly differ |

**Table 1.**
Explanations and examples for five non-IID data distribution

However, it is unlikely that all other hospitals participating in the same system have established such infrastructure and equipment. These differences raise system heterogeneity issues in federated learning processes.

*4.1.3 Approaches for heterogeneity issues.* Data and system heterogeneity issues incur misleading results in a federated learning environment. When the local data distribution varies, the global model weights hardly converge to an optimal point; as clients have different capabilities, the global model may be biased into specific dimensions. In a federated learning environment that requires convergence to the optimal gradient, the unstable convergence issue due to nonIID data is defined as client drift (Karimireddy *et al.*, 2020). Studies have been conducted to solve the problem, and our team classifies the works into three main categories: clustering, optimization and model fusion.

4.1.3.1 Clustering methods. Clustering-based machine learning algorithms are the representative unsupervised learning methods for finding similarities with peripheral data for unlabeled data sets. It is used as a form of grouping unlabeled data and grasping its core characteristics and is also adopted as a method to pre-process and use the remaining data when the number of data with labels is relatively small. Many researchers leveraged clustering methods to solve data heterogeneity issues because gathering data points with similar patterns from unlabeled data is analogous to grouping clients with similar weight distribution to which the server is inaccessible in federated learning.

The main concept of the clustering-based approach is to identify a group of clients with similar distributions and to compromise the heterogeneity of datasets that each client has. Sattler proposed Clustered Federated Learning that adopts a clustering method by measuring the cosine similarity of each local model (Sattler *et al.*, 2021). They swapped the labels to fit the data set for the nonIID environment. Experimental results demonstrated that the proposed work achieved a reasonable performance even in extreme nonIID situations.

Briggs applied a hierarchical clustering-based method to improve the performance when the clients have nonIID data (Briggs *et al.*, 2020). They introduced a method to generate optimized clusters by comparing L1, L2 and cosine similarity distance metrics between clusters, demonstrating that the proposed method can reach the desired performance faster and more accurately than traditional federated learning methods. Based on (ibid.), Yoo *et al.* (2021) used heart rate variability data of patients to diagnose depression severity. They applied a clustering-based federated learning algorithm called personalized federated learning with clustering for new incoming participants to improve prediction accuracy and solve nonIID issues.

Various clustering-based techniques were tried to deal with the performance degradation caused by data heterogeneity, and there are also research studies to solve the issues of model aggregation time delay caused by heterogeneous data.

Chen *et al.* (2020a) introduced the FedCluster to address the problem of slow convergence that occurs when the federated averaging on heterogeneous local data. Each participating local client is not included in the model update at a time but clustered according to certain criteria, in which the cluster participates in the federated learning process for each round. Depending on the target federated learning environment, it applies the best clustering scenario of random uniform clustering, timezone-based clustering and availability-based clustering. Experiments with MNIST and CIFAR-10 benchmark data sets demonstrated that the cyclic federated learning structure through FedCluster showed a faster convergence time than the conventional FedAvg algorithm.

As the server has no information about the participants' data distribution, researchers tried to handle the issue with unsupervised learning algorithms. Clustering-based method, one of the most widely spread unsupervised learning, is applied to federated learning to solve the heterogeneity issues by gathering clients. However, as there is concern that this approach rather induces biased results for certain clusters, researchers should be careful not to bias the distribution in model generation.

4.1.3.2 Optimization methods. A different approach for solving the heterogeneity issue is using optimization algorithms. The only use of FedAvg limits the models from converging as a single global model cannot properly represent nonIID data. This is because the data derived from different distributions diverge in various directions to represent the features of the distributions to which they belong.

Xie *et al.* (2020) proposed federated Stochastic Expectation-Maximization (SEM), a multicenter federated learning framework, which allows optimization function to find multiple local optima points. The SEM makes it possible to find local optima points in a variety of distributions of clients in each clustered multi-center, which solves the problems that do not converge to a single global model. Reddi *et al.* (2020) applied various adaptive optimizers called FEDADAGRAD, FEDYOGI and FEDADAM, which are advantageous in hyperparameter coordination and improving the convergence rate of the federated learning model over the vanilla FedAvg method.

SCAFFOLD also addresses client-drifting issues in federated learning. By introducing a client control variable, Karimireddy *et al.* (2020) adjusts each local update in the direction of the optimal global model. FedProx from Li *et al.* (2018) modifies the conventional FedAvg method in two directions; tolerating partial work and adding proximal terms. Each participating device will have a systematic difference, such as computing power, and accordingly, the research solves the issue by giving each device an iteration that can be performed on its device. Li added a proximal term to the algorithm, to prevent the heterogeneity problems arising from excessive local update iterations.

4.1.3.3 Mixture model of global and local. In addition to the optimization method and the clustering method on clients with similar parameters, techniques have also emerged to create an optimized model form by engaging in the structure of the training model layer. Hanzely and Richt¨arik (2020) and Arivazhagan *et al.* (2019) solved heterogeneity issues by mixing global and local models generated by federated learning. The global model is too general for all clients' data and the local model is too specific to generalize, they combine representative learning layers from both global and local to generate a fusion model. In this way, it is possible to adopt general features from other participating clients, while adding their personalized features from their local data.

### *4.2 Client management issues*

Unlike the centralized machine learning model that collects and analyzes user data in one server, federated learning requires each client to join a system using their local data. Therefore, the system to manage clients must be adopted for efficient training. The server has to determine the list of participating clients based on limited information and confirm their exact contribution, to avoid any free-riding clients who are looking for the benefits without any contribution to the model training.

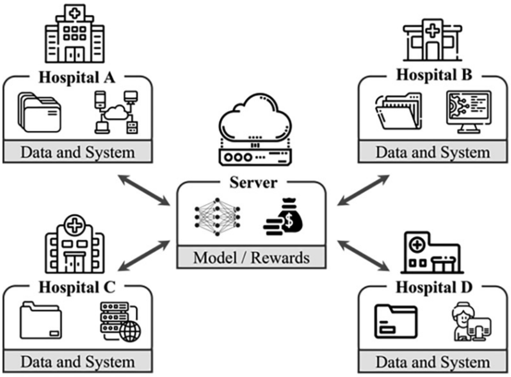
In line with this, the concept of an incentive mechanism that encourages more active participation based on user contribution has recently been introduced to solve client management aspects of federated learning. The incentive mechanism was first introduced by system architects aiming for improved performance of repetitive loops of manufacturers by rewarding the participants. Although many studies have been conducted so far, they have not used a proper incentive mechanism in the federated learning system. Standards must be set to evaluate the user contribution. In the following section, we describe the evaluation criteria from two different perspectives: data quality and resource usage (Figure 4).

*4.2.1 Client management based on data quality.* With the advent of the big data era, the value of client data is increasing regardless of the research field. Unlike when awareness of privacy was weak, high-quality data became an incomparable asset under strengthened data protection regulations. Therefore, to generate a better performance model in a federated learning environment, clients who participate in the system hope to receive appropriate rewards for their data quality. Depending on the data quantity and quality of the local clients, the server should accurately calculate their contribution to keep them participating in the training system. However, in federated learning, the application of incentive mechanism-based methods has been mentioned as an alternative because the server does not have the authority to verify user data.

It is important to evaluate the quality and quantity of data held by companies or institutions as well as data held by individual users. Unlike when training with patient data in one hospital as shown on the left side of Figure 3, the server should evaluate the data of each institution when several hospitals aim to jointly create a better performance model as shown in other structures in Figure 3. Chen *et al.* (2020b) deals with the client management issue that occurs in multi-parties federated learning. From a business perspective, the research team pointed out that if other companies can grow further through the high-quality data of one good institution, they will no longer participate in federated learning because it will threaten their profits in the future.

In addition to the necessity of considering the incentive mechanism approach for calculating contributions, the number of participating clients and the amount of data held by them affect the model performances. Zhao *et al.* (2021) presents a new stochastic gradient method called FedPAGE, comparing the number of clients and the amount of datasets each client has with the results how the difference occurs in reaching the optimal performance. It was found that when the number of data held by one client was large, the convergence time is relatively faster, but when the total number of clients participating in the learning process is large, it was not easily converge.

Generally in federated learning, participants are randomly selected and participate in the training round. In the FedCS paper, a temporal threshold was set in each training round of federated learning, and an experiment was conducted in the form of removing all clients that were not reflected in learning due to reasons such as computational limitation or data limitation (Nishio and Yonetani, 2019). As a result, in both nonIID and IID data



**[Environments]**

- A server only possesses model and limited rewards.
- Hospitals only possesses heterogeneous data and system.

**[Process]**

1. Different hospitals participate in Federated Learning with their own data and structured system.
2. A server uses client evaluation methods to determine whether hospitals are allowed to continue to participate.
3. As the process repeats, the server utilizes incentive mechanism to distribute rewards to hospitals according to their contributions.

**Figure 4.**
Client management in the field of medicine

environments, it efficiently selects clients and converges faster to the desired model performance.

*4.2.2 Client management based on resource usage.* In addition to the importance of client data, it is critical to identify the computational capabilities or systems they use to participate in federated learning. As mentioned in the system heterogeneity section, the computational environment of each user may be different in federated learning, resulting in differences in model training outcomes. Zeng described possible multi-dimensional resource differences in Mobile Edge Computing (MEC) (Zeng *et al.*, 2020). Central servers have to recognize dedicated resources and provide suited incentives to encourage good clients. In addition, the aforementioned FedCS research also introduces the process of determining the federated client to participate in learning based on the resource ability held by each client. Most federated learning systems, on the other hand, assumed that each client participating in model training has the same dedication.

*4.2.3 Approaches for client management issues.* Generating a better performance federated learning model requires contributions from multiple data providers with proper quality of data and resources, though not all clients equally contribute to federated learning. Therefore, an algorithm that can identify each contribution needs to be applied. The two most widely studied approaches leveraged Shapley value and Stackelberg's game theory which we will discuss in the following.

4.2.3.1 Shapley value. IA suggested how to perform data valuations through Shapley Value (Jia *et al.*, 2019), which has been widely used in game theory. They listed how Shapley Value enables data evaluation in multiple machine learning analytics environments and demonstrates their approach's scalability. Similar to the approaches of game theory with Shapley Value, Lim used contract theory to identify the data quality and quantity of each data owner and applied a hierarchical incentive mechanism in the federated crowd-sourcing network (Lim *et al.*, 2020).

4.2.3.2 Stackelberg game theory. The Stackelberg game theory is widely used to assess each participant's contribution and construct a reward system. Sarikaya and Ercetin (2019) solved the problem caused by heterogeneous worker performance through the Stackelberg game-based method. This measures the time it takes each participant to complete a given task for an updated gradient transfer and assigns a proper reward for each computing power based on the Stackelberg game theory. Khan also adopted a Stackelberg game-based approach in L. U. Khan *et al.* (2020). Each edge node delivers its own computation energy and latency to the model aggregator, which is in charge of incentives. The goal of the model aggregator is to minimize learning time while maximizing model performance, so it adjusts client learning level based on the clients' Stackelberg results.

Pandey *et al.* (2020) proposed a two-stage of Stackelberg game by developing an optimal learning model through maximizing the utility of participating devices and MEC servers. When the MEC server announces the objectives of the optimized global model it wants to create and rewards accordingly, each device participates in the federated learning by optimizing the global learning model and maximizing the yield through the local data it possesses.

Similarly to studies that deal with data values by Stackelberg game theory, auction systems were also applied to solve clients' heterogeneous resource issues. To address the problem of client management, Le *et al.* (2021) used a primary-dual greedy auction mechanism. When the server is assigned the task of federated learning training, each client submits a bid based on their own computation resource and transmission power. Subsequently, the server selects clients who can develop optimal models based on the bid list and provides customized rewards after completing the learning task. Zeng *et al.* (2020)

also applied auction-based techniques on various scoring functions to allow devices with high-quality data to participate at a relatively low cost.

Giving no or the same level of incentive to all local clients will result in some participants earning rewards for providing their low-quality data and resource. Others, on the other hand, will suffer from losses while contributing high-quality data. Hence, designing federated learning without explicit incentive mechanisms may violate the purpose of federated learning, which is to collaboratively develop a high-performance learning model. Researchers must develop a more sophisticated incentive mechanism to manage local clients in a real-world federated learning environment systematically.

### *4.3 Traceability and responsibility issues*

In federated learning, the server cannot directly investigate the local clients' data. However, these structural advantages cause considerable difficulties in that the server cannot track the results or hold them accountable for learning outcomes. The inability to check the learning process of machine learning is a problem that has arisen as the application of deep neural network techniques has expanded. This problem is due to the black-box nature of the neural network, and federated learning should consider taking another step such as in Figure 3. Machine learning has a black box issue that makes it difficult to accurately determine the process of the algorithm that produced the result for provided data. However, in addition to this traceability issue, there is a double black box problem that federated learning cannot even investigate the data each client has, as shown on the right side of Figure 5.

In addition to protecting patient data through federated learning applications, another crucial factor to be considered in the medical machine learning field is explainability. That is

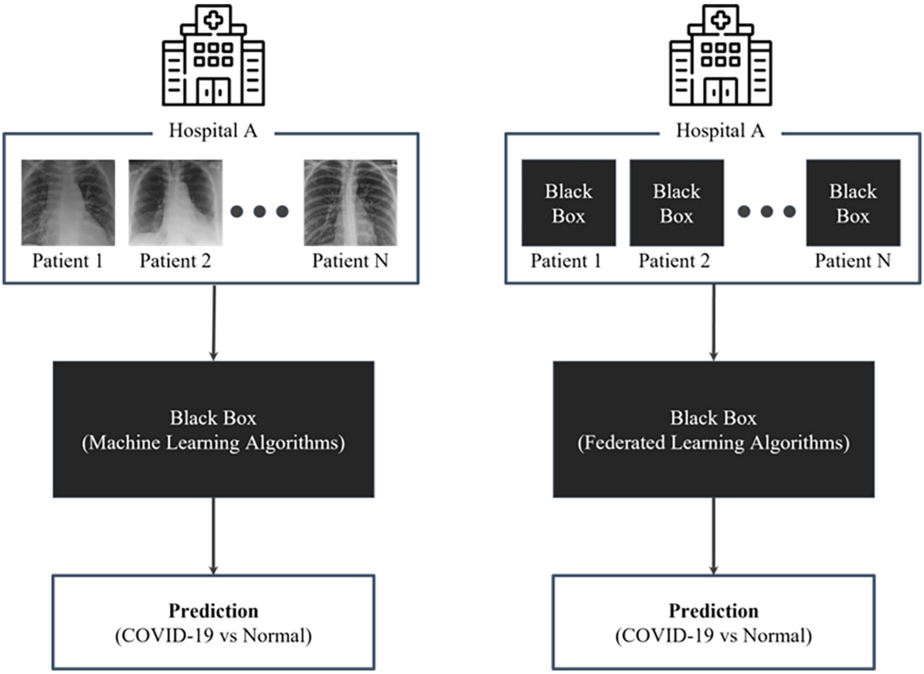


**Figure 5.**
Black box nature of machine learning and federated learning

why precision and recall are adopted more for performance evaluation than accuracy in the medical field. It is crucial to understand the exact results of each prediction class, not just how accurate the model is. As the medical field is directly related to human life, it should be possible to explain on what basis they make decisions. However, neither the existing machine learning methods nor the federated learning methods show flawless predictive performance, and explanation power for predictive values is insufficient for actual usage.

It would be ideal if the predictive or diagnostic model obtained by the federated learning consistently demonstrates professional-level performance with proper descriptions, but this is not the case. Researchers should consider the process when it produces false-positive results. If the false-positive rate is high in the federated learning task, the server must determine which participant or training round is responsible for the problem. It may be necessary to redesign the entire architecture of the medical federated learning model when the specific client who caused the error or the training process is not defined. Rieke also described issues of determining the subject of responsibility for the unexpectedly faulty results of the medical analysis, caused by the federated learning in health care (Rieke *et al.*, 2020).

Medical experts can provide sufficient information and advice during centralized machine learning data preprocessing, such as noise filtering, segmentation and even data labeling (J. Xu *et al.*, 2021). Federated learning cannot perform the mentioned preprocessing step, so it can be a trade-off structure in the field of medicine. Protecting PHI from arbitrary intrusion is significant, but the advantages of data protection are fatal disadvantages for medical applications. In line with this, XAI was applied to solve the issue (Lundberg and Lee, 2017), allowing researchers to figure out which parts of the neural networks are responsible for the performance degradation. Characteristics of XAI must also be applied when federated learning is adopted in the field of medicine to prevent medical errors caused by false-positive rates.

Few studies applied XAI to medical federated learning, but they have attempted to increase accountability while maintaining the advantages of data privacy by combining XAI and federated learning. Raza *et al.* (2022) combined XAI with Federated Transfer learning to design an electrocardiography monitoring healthcare system, by adding Gradient-weighted Class Activation Mapping (Grad-CAM) module on federated learning architecture to provide signal classification. However, more extensive researches on XAI and federated learning remain an open problem (Selvaraju *et al.*, 2017).

### *4.4 Privacy and security issues*

Although deep neural network models brought huge advancement in the medical field, and federated learning prevents a model from private information leakage, various privacy and security attacks remain unsolved problems. For instance, a medical image deep neural networks are especially susceptible to adversarial attacks due to ambiguous ground truth, highly standardized format and many other reasons (Finlayson *et al.*, 2018). At the same time, however, the attacks can be easily detected because of the biological characteristic of the images (i.e. manipulation occurring outside the pathological region) (Ma *et al.*, 2021). This section will introduce various attack and defense approaches, especially those studied in federated learning environments (Table 2).

*4.4.1 Attacks.* Federated learning is especially vulnerable to adversarial attacks due to the absence of raw data inspection and collaborative training using private local data. As generally known, machine learning can be divided into two phases: the training phase and the inference phase. Nevertheless, due to the zero knowledge distributed nature of federated learning, the training phase attacks are more severe than those of the inference phase; as

**Table 2.** A summary of privacy and security attacks

| Attack category | Attack types | Attack target | Attack methods | Attacker role |
|---|---|---|---|---|
| Poisoning attacks | Data poisoning | Security (data integrity) | Label flipping | Client |
| | | | Backdoor | Client |
| | Model poisoning | Security (model integrity) | Gradient manipulation | Client |
| | | | Training rule Manipulation | Client |
| Inference attacks | Membership inference | Privacy (information leak) | Membership inference | Client and server |
| | | | Properties inference | Client and server |
| | GAN reconstruction | Privacy (information leak) | Class representative inference | Client and server |
| | | | Inputs and labels Inference | Client and server |

neither centralized property (i.e. server) nor the other participating clients are allowed to investigate each other's private data.

4.4.1.1 Poisoning attacks. As the central aggregator is inaccessible to each distributed training data, an attacker on the client-side can intentionally send the poisonous model update. For example, an attacker may poison the data or model updates to simply degrade the model performance or to bias, the model against certain demographics to cause a scandal or toward a direction to over-recommend certain treatments.

Poisoning attacks can be categorized into data poisoning and model poisoning attacks. The two types of poisoning attacks are different in that the former aim to compromise the integrity of the training data, while the latter aim to compromise the integrity of the model.

Data poisoning attacks include label flipping or data backdoor attacks. Label flipping attacks are one of the client-side data poisoning attacks that flip the labels of the attacker-chosen data classes to attacker-chosen labels to misclassify the specific data classes. Tolpegin *et al.* (2020) simulate and analyzes label flipping attacks. In their experiment, the class label of airplane images is flipped to bird, so the global model misclassifies airplane images to bird at inference time. Hayes introduces a contamination attack that is essentially manipulating a small set of training data (Hayes and Ohrimenko, 2019), compromising the integrity of the data. The author suggested adversarial training as a defense, which will be discussed in Section 4.4.2 in detail.

The model poisoning attacks involve model backdoors and gradient and/or training rules manipulation. Although poisoning attacks can be differently categorized into two, model poisoning attacks generally include data poisoning attacks as the poisoned data ultimately leads the model to be poisoned. Therefore, we here introduce numerous previous works that are not limited to data poisoning but the hybrid approach of data and model poisoning attacks as well.

Bagdasaryan *et al.* (2020) proposed model replacement to introduce a backdoor into the global model. Their proposed attack kept high accuracy for both main and backdoor tasks to improve its persistence by evading anomaly detection. Fang *et al.* (2020) manipulated the local model parameters before sending them to the global server. As a result of the manipulation, the local models deviate toward the inverse direction of the global model before the attack. Bhagoji *et al.* (2019) introduced a targeted model poisoning attack that poisons the model updates by explicit boosting and remains stealthy by alternating minimization. Xie *et al.* (2019) proposed a distributed backdoor attack that breaks down a global trigger pattern into distinct local patterns and embeds them in the training sets of several adversarial parties. Their work showed that the distributed attack is more effective than the centralized backdoor attacks. Fung pointed out the vulnerabilities of federated learning, especially against the Sybil-based poisoning attacks (Fung *et al.*, 2018). The authors mentioned that the distributed nature increases attack effectiveness, especially when multiple malicious parties participate.

4.4.1.2 Inference attacks. Unlike poisoning attacks, inference attacks typically hamper the privacy of private information. Even though federated learning alleviates the privacy leakage issues, there still exist some privacy threats. For example, exploiting the fact that the communicating model parameters necessarily include the encrypted information about the private training data, an attacker may approximate or even reconstruct the data samples by extracting and decoding the model parameters. Inference attacks include membership inference and GAN-based reconstruction attacks that lead the system to leak information about the training data unintentionally. The recent trend of inference attacks is moving toward the GAN-based method due to its stealth and detection evasion ability.

Wang *et al.* (2019) achieves user-level privacy leakage by incorporating GAN with a multitask discriminator. Their proposed method discriminates category, reality and client identity of input data samples and recovers the user-specific private data. In GAN poisoning attack research (J. Zhang *et al.*, 2019), an attacker first acts as a benign participant and stealthy trains a GAN to mimic the other participants' training samples. With the generated samples, the attacker manipulates the model update with a scaled poisoning model update to compromise the global model ultimately.

*4.4.2 Defense methods.* The defense mechanism against the attacks in federated learning includes minimizing the influence of malicious clients and preventing the malicious clients' model parameters from incorporating into the global model. Also, for privacy, the defense includes preventing private information from being leaked.

4.4.2.1 Information leak prevention. Several techniques, such as Multi-Party Computation (MPC), Homomorphic Encryption and Differential Privacy (DP), are adopted to the FL to protect the models and data. MPC is a type of cryptographic method that which the final output is computed across multiple participants by exchanging secret shares. Any specific local model is then impossible to be reconstructed, therefore, the local models are kept private. HE permits arithmetic operations on encrypted parameters without decryption, to protect the exchanging model parameters. Using this technique, the global server can aggregate the encrypted model parameters sent from the local clients without decoding. This method may prevent various attacks from both the honest-but curious server and malicious clients. DP is adopted in federated learning not only for the local data protection but the local or global model protection. This technique adds perturbation or random noise (e.g. Gaussian noise) to the training data and the model parameters to prevent the adversaries from inferring the data or models.

Privacy-enhanced Federated Learning(PEFL) research solved the vulnerability by leveraging MPC(Hao *et al.*, 2019a). Their proposed method has strength in the situation that multiple clients collude to prevent private data from being leaked.

The following previous works leverage a combination of those mentioned above three popular techniques. Augenstein demonstrated generative models that are trained using federated methods with DP. They applied their model to both text and image, using differentially private federated GANs (Augenstein *et al.*, 2019). Ghazi *et al.* (2019) also exploits DP for secure aggregation via shuffled model. Their proposed methods preserve privacy and relax computational complexity, which is one of the necessities for better scalability. (Hao *et al.* (2019b) combined DP and additive homomorphic encryption to obtain both performance and security. Their proposed method is especially robust when not only the clients but the server is honest but curious. Truex *et al.* (2019) approached similarly, combined DP with multiparty computation balances the trade-off between the performance and availability.

4.4.2.2 Model protection. In the perspective of security, various works have been done to prevent model from corruption. There are broadly two approaches: robust aggregation and anomaly detection. Typically, robust aggregation aims to train the global model robustly, even if malicious clients participate in the federated learning process. Anomaly detection aims to detect malicious or anomalous clients so that their model updates are not aggregated in the server and/or block them from further participation.

Fu *et al.* (2019) offered a robust aggregation technique with residual-based reweighting. Their reweighting strategy used iteratively reweighted least squares to integrate repeated median regression. On the other hand, there are various defense attempts to prepare for possible attacks in the form of classifying anomalies (Li *et al.*, 2020a; Shen *et al.*, 2016; Fang *et al.*, 2020; Tolpegin *et al.*, 2020; Jeong *et al.*, 2021). Li proposed spectral anomaly detection mechanism based on models' low-dimensional embeddings (Li *et al.*, 2020a). The central server learns to detect and remove the malicious model updates by removing noisy features and retaining essential features, leading to targeted defense. A notable point of their proposed approach is that it worked in both semi-supervised and unsupervised learning and designed the protocol for encryption. FoolsGold mitigated poisoning attacks (Wu *et al.*, 2020). Their approach comes from the idea that the malicious clients' updates are different from those of the benign ones; thereby, the Sybils are distinguishable by measuring the contribution similarity.

The aforementioned work's objective, anomaly detection, remains the same, but several approaches leveraged clustering-based and thresholding-based approaches (Shen *et al.*, 2016; Fang *et al.*, 2020; Tolpegin *et al.*, 2020; Jeong *et al.*, 2021; Cao *et al.*, 2020; Sun *et al.*, 2019). Auror dealt with targeted poisoning attacks leveraging clustering and thresholding techniques. Shen created clusters of clients and measured the pairwise distance between them, which will be used to distinguish two distinctive clusters. Within each cluster, if more than half of the clients are determined to be malicious by the predefined threshold, all the clients belonging to that cluster are then classified as malicious (Shen *et al.*, 2016). Fang *et al.* (2020) proposed a defense mechanism, a combination of error rate-based rejection and loss function-based rejection. The idea comes from that as the malicious clients tamper with the models' performance, the error rate and the loss impact are greater than those of benign clients. Therefore, if a client greatly impacts the higher error and loss rate, the clients are identified as malicious so as to aggregate only benign clients' weights. Along with the data poisoning attack, Tolpegin *et al.* (2020) used principal component analysis to visualize the spatially separable malicious clients' model updates and those of benign ones.

Sun defended against backdoor attacks by leveraging norm bounding and weak DP (Sun *et al.*, 2019). They noted that the norm values of malicious clients are relatively greater than those of benign ones, so they detected malicious clients by thresholding based on the calculated norm value of each clients' weight updates. FLTrust protected the global model against byzantine attacks by making use of a ReLU-clipped cosine similarity-based trust score(Cao *et al.*, 2020). In their works, however, the global server had been trained on an innocent dataset called root data set; in other words, the server did have the knowledge of benignity and malice of the client data set, which is a breach of the no-raw-data-sharing assumption. Based on Cao *et al.* (2020) and Sun *et al.* (2019), Jeong *et al.* (2021) proposed anomalous and benign client classification in federated learning that leverages feature dimension reduction, dynamic clustering and cosine-similarity-based clipping to detect and classify anomalous and benign clients where the benign ones have nonIID data and IID data. Similar to the aforementioned approaches, only the benign clients' model weights are aggregated in the global server.

## 5. Framework under development

Like many different cases introduced in the study, federated learning is in the stage of building and applying an actual data environment beyond theoretical research not only in the medical field but also in various other fields. As a specific task-optimized federated learning system design is not simple, several open-source frameworks are available for researchers. Depending on the data environment and analysis objectives, frameworks such as federated learning AI technology enabler (FATE) (Q. Yang *et al.*, 2019), TensorFlow Federated, NVIDIA Clara and IBM Federated Learning can help to expand the research scope.

Our medical federated learning research team is also developing a customized lightweight federated learning framework. The core goal of the framework under development is to allow users to participate in federated learning systems and receive predicted results based on their data set in any environment, including mobile devices and computers. There have been quite a few studies that apply federated learning based on data measured through smartphones or wearable devices and conduct learning under data protection, but not many open-source frameworks support the mobile training environment. FATE focuses on industrial usage, NVIDIA Clara focuses on health-care institutes usage, and even IBM Federated Learning also focuses on enterprise usage. Therefore, our framework focuses primarily on the customized use of modular structures and provides practical implementation and support for lightweight devices.

The core component of a federated learning system mainly controls the process between a cloud and participating clients. It consists of seven detailed management modules about the database, model, weight, participant, cluster, state and aggregate as shown in Figure 6. Model and weight management modules are in charge of initializing a global model and distributing it to each client based on the participant management module. When participating devices update received model gradients with their local data, the aggregate management module collects and updates the previous model gradients for better performance. The entire process and functionality of the configured module follow the basic federated learning algorithm explained in section two. However, the main distinction of our framework is that it adopts efficient APIs designed for training mobile devices. The framework contains six different management APIs; log and configure, HTTP server, database, model, weight and cluster, which helps system clouds effectively communicate with participating devices. With designed APIs, the cloud controls communication protocols, balances shared resources, authorizes the clients and authenticates the clients.

In line with the growing wearable device market, our research team looks forward to the development of the medical field that analyzes user data widely and conducts daily diagnosis and evaluation in the medical field through a modular framework that can also be applied to mobile devices. When the development of the framework of our research team is completed, a comparison of model performance with other federated learning frameworks such as FATE and TensorFlow Federated will be conducted based on benchmark data sets. MedMNIST data (J. Yang *et al.*, 2021), which are medical benchmark data, will be mainly used to test the issue environment such as data heterogeneity, system heterogeneity and client management described in this study. We will also present how much the proposed framework can solve each federated learning issue through the system of modularized architecture. The explained framework will be open to the public after the development completes.

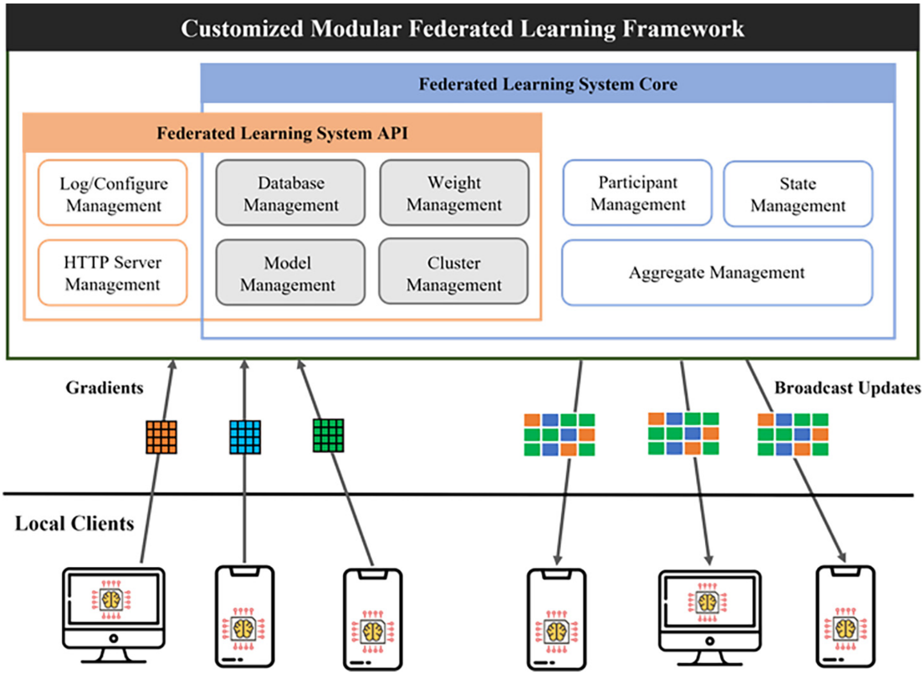


**Figure 6.**
The architecture of the modular FL currently in development

## 6. Conclusion and future works

Google introduced federated learning that enables AI learning without collecting local data in 2017. It has been actively studied, especially in the medical field. Training methods without client data collection is an attractive advantage under data privacy perspectives. Federated learning techniques, however, still have a variety of unresolved open problems due to their characteristics, such as different data distributions, client participating structures and even vulnerable training environments.

In this research, current unsolved issues of federated learning and emerging solutions are discussed with specific cases in medical machine learning approaches. Open problems of medical federated learning are related to data/system heterogeneity, client management, traceability, accountability and security perspectives. Existing federated learning studies have dealt with issues of the federated learning environment itself, but there was a lack of content on how these points were applied in the real-world working environment to cause problems. Therefore, we categorize the critical issues of the field and help researchers easily understand through examples of actual medical machine learning environments. Various attempts have been applied to address different issues that occur in the medical field, but there remain some other open questions besides the issues mentioned above.

### *6.1 Explainable artificial intelligence in federated learning*

The explainability of the deep neural network models has been an important issue in machine learning due to its black-box nature. After the input data is entered into the model, the result is output through a complex neural network structure, but it is not possible to accurately grasp the decision that occurs in the neural network process. Medical federated learning, where the patients' health information data cannot be investigated or collected, proving the explainability of the output result is even more difficult as mentioned in the accountability and traceability section.

The XAI field is peeling off the existing black-box characteristics by identifying which part of the input data affects the deep neural network output values. Similar to the Grad-CAM-based XAI study conducted by Raza in ECG data classification, there are studies using XAI, but only a few exist. XAI techniques such as Grad-CAM (Selvaraju *et al.*, 2017), DeepLIFT (J. Li *et al.*, 2021) and SmoothGrad (Smilkov *et al.*, 2017) use a salience mapping-based method that finds and displays the parts that have affected network output. However, this approach also requires data access for validation, which does not match the basic concept of federated learning. Therefore, to strengthen the explainability of diagnostic results, XAI research designed for the federated learning environment will be needed.

### *6.2 Data heterogeneity, attack and federated learning*

Studies on attack and defense mechanisms have been published from data security perspectives, and attacks using heterogeneous data distribution have not yet been fully studied by researchers. However, as explained in the heterogeneity issues section, the data environment faced by researchers is not the identical and independently distributed environment. To deal with the actual implementation of medical federated learning, careful defense methods against heterogeneous data environment attacks must be further considered.

## References

Arivazhagan, M.G., *et al.* (2019), "Federated learning with personalization layers", arXiv preprint arXiv:1912.00818.

Augenstein, S., *et al.* (2019), "Generative models for effective ML on private, decentralized datasets", arXiv preprint arXiv:1911.06679.

Bagdasaryan, E., *et al.* (2020), "How to backdoor federated learning", *International Conference on Artificial Intelligence and Statistics. PMLR*, pp. 2938-2948.

Bhagoji, A.N., *et al.* (2019), "Analyzing federated learning through an adversarial lens", *International Conference on Machine Learning. PMLR*, pp. 634-643.

Briggs, C., Fan, Z. and Andras, P. (2020), "Federated learning with hierarchical clustering of local updates to improve training on nonIID data", *2020 International Joint Conference on Neural Networks (IJCNN). IEEE*, pp. 1-9.

Brophy, E., *et al.* (2021), "Estimation of continuous blood pressure from PPG via a federated learning approach", *Sensors 21*, Vol. 18, p. 6311.

Cao, X., *et al.* (2020), "FLTrust: byzantine-robust federated learning via trust bootstrapping", arXiv preprint arXiv:2012.13995.

Chen, C., *et al.* (2020a), "Fedcluster: boosting the convergence of federated learning via cluster-cycling", *2020 IEEE International Conference on Big Data (Big Data). IEEE*, pp. 5017-5026.

Chen, M., *et al.* (2020b), "Mechanism design for multi-party machine learning", arXiv preprint arXiv:2001.08996.

Dayan, I., *et al.* (2021), "Federated learning for predicting clinical outcomes in patients with COVID-19", *Nature Medicine*, Vol. 27 No. 10, pp. 1735-1743.

Fang, M., *et al.* (2020), "Local model poisoning attacks to byzantinerobust federated learning", *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pp. 1605-1622.

Finlayson, S.G., *et al.* (2018), “Adversarial attacks against medical deep learning systems”, arXiv preprint arXiv:1804.05296.

Fu, S., *et al.* (2019), “Attack-resistant federated learning with residual-based reweighting”, arXiv preprint arXiv:1912.11464.

Fung, C., Yoon, C.J. and Beschastnikh, I. (2018), “Mitigating sybils in federated learning poisoning”, arXiv preprint arXiv:1808.04866.

Ghazi, B., Pagh, R. and Velingker, A. (2019), “Scalable and differentially private distributed aggregation in the shuffled model”, arXiv preprint arXiv:1906.08320.

Hanzely, F. and Richt¨arik, P. (2020), “Federated learning of a mixture of global and local models”, arXiv preprint arXiv:2002.05516.

Hao, M., Li, H., Luo, X., *et al.* (2019a), “Efficient and privacyenhanced federated learning for industrial artificial intelligence”, *IEEE Transactions on Industrial Informatics*, Vol. 16 No. 10, pp. 6532-6542.

Hao, M., Li, H., Xu, G., *et al.* (2019b), “Towards efficient and privacy-preserving federated deep learning”, *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. *IEEE*, pp. 1-6.

Hayes, J. and Ohrimenko, O. (2019), “Contamination attacks and mitigation in multi-party machine learning”, arXiv preprint arXiv:1901.02402.

Jeong, H., Hwang, J. and Chung, T.M. (2021), “ABC-FL: anomalous and benign client classification in federated learning”, arXiv: 2108. 04551[cs.LG].

Jia, R., *et al.* (2019), “Towards efficient data valuation based on the shapley value”, *The 22nd International Conference on Artificial Intelligence and Statistics*. *PMLR*, pp. 1167-1176.

Kairouz, P., *et al.* (2019), “Advances and open problems in federated learning”, arXiv preprint arXiv:1912.04977.

Kaissis, G., *et al.* (2021), “End-to-end privacy preserving deep learning on multi-institutional medical imaging”, *Nature Machine Intelligence*, Vol. 3 No. 6, pp. 473-484.

Karimireddy, S., Praneeth, *et al.* (2020), “Scaffold: stochastic controlled averaging for federated learning”, *International Conference on Machine Learning*. *PMLR*, pp. 5132-5143.

Khan, L.U., *et al.* (2020), “Federated learning for edge networks: resource optimization and incentive mechanism”, *IEEE Communications Magazine*, Vol. 58 No. 10, pp. 88-93.

Kumar, R., *et al.* (2021), “Blockchain-federated-learning and deep learning models for covid-19 detection using CT imaging”, *IEEE Sensors Journal*, Vol. 21 No. 14, pp. 16301-16314.

Le, T.H.T., *et al.* (2021), “An incentive mechanism for federated learning in wireless cellular network: an auction approach”, *IEEE Transactions on Wireless Communications*, Vol. 20, pp. 4874-4887.

Li, J., *et al.* (2021), “Deep-lift: deep label-specific feature learning for image annotation”, *IEEE Transactions on Cybernetics*, Vol. 52 No. 8.

Li, S., *et al.* (2020a), “Learning to detect malicious clients for robust federated learning”, arXiv preprint arXiv:2002.00211.

Li, T., Kumar Sahu, A., Talwalkar, A., *et al.* (2020b), “Federated learning: challenges, methods, and future directions”, *IEEE Signal Processing Magazine*, Vol. 37 No. 3, pp. 50-60.

Li, T., Sahu, A.K., Zaheer, M., *et al.* (2018), “Federated optimization in heterogeneous networks”, arXiv preprint arXiv:1812.06127.

Lim, W.Y.B., *et al.* (2020), “Hierarchical incentive mechanism design for federated machine learning in mobile networks”, In *IEEE Internet of Things Journal*, Vol. 7 No. 10, pp. 9575-9588.

Linardos, A., *et al.* (2022), “Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease”, *Scientific Reports*, Vol. 12 No. 1, pp. 1-12.

Liu, B., *et al.* (2020), “Experiments of federated learning for covid-19 chest x-ray images”, arXiv preprint arXiv:2007.05592.

Lundberg, S.M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768-4777.

Ma, X., *et al.* (2021), "Understanding adversarial attacks on deep learning based medical image analysis systems", *Pattern Recognition*, Vol. 110, p. 107332.

Nandi, A. and Xhafa, F. (2022), "A federated learning method for realtime emotion state classification from multi-modal streaming", *Methods*, Vol. 204.

Nishio, T. and Yonetani, R. (2019), "Client selection for federated learning with heterogeneous resources in mobile edge", *ICC 2019-2019 IEEE international conference on communications (ICC)*, *IEEE*, pp. 1-7.

Pandey, S.R., *et al.* (2020), "A crowdsourcing framework for on-device federated learning", *IEEE Transactions on Wireless Communications*, Vol. 19 No. 5, pp. 3241-3256.

Raza, A., *et al.* (2022), "Designing ECG monitoring healthcare system with federated transfer learning and explainable AI", *Knowledge-Based Systems 236*, Vol. 236, p. 107763.

Reddi, S., *et al.* (2020), "Adaptive federated optimization", arXiv preprint arXiv:2003.00295.

Rieke, N., *et al.* (2020), "The future of digital health with federated learning", In *NPJ Digital Medicine*, Vol. 3 No. 1, pp. 1-7.

Sarikaya, Y. and Ercetin, O. (2019), "Motivating workers in federated learning: a Stackelberg game perspective", *IEEE Networking Letters*, Vol. 2 No. 1, pp. 23-27.

Sattler, F., Müller, K.-R. and Samek, W. (2021), "Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32 No. 8, pp. 3710-3722.

Selvaraju, R.R., *et al.* (2017), "Grad-cam: visual explanations from deep networks via gradient-based localization", *Proceedings of the IEEE international conference on computer vision*, pp. 618-626.

Shen, S., Tople, S. and Saxena, P. (2016), "Auror: defending against poisoning attacks in collaborative deep learning systems", *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pp. 508- 519.

Smilkov, D., *et al.* (2017), "Smoothgrad: removing noise by adding noise", arXiv preprint arXiv:1706.03825.

Sun, Z., *et al.* (2019), "Can you really backdoor federated learning?", arXiv preprint arXiv:1911.07963.

Tedeschini, B.C., *et al.* (2022), "Decentralized federated learning for healthcare networks: a case study on tumor segmentation", IEEE Access.

Tolpegin, V., *et al.* (2020), "Data poisoning attacks against federated learning systems", *European Symposium on Research in Computer Security*. *Springer*, pp. 480-501.

Truex, S., *et al.* (2019), "A hybrid approach to privacy-preserving federated learning", *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 1-11.

Wang, Z., *et al.* (2019), "Beyond inferring class representatives: user-level privacy leakage from federated learning", *IEEE INFOCOM 2019 – IEEE Conference on Computer Communications*. *IEEE*, pp. 2512-2520.

Wu, C., *et al.* (2020), "Mitigating backdoor attacks in federated learning", arXiv preprint arXiv:2011.01767.

Xie, C., *et al.* (2019), "Dba: distributed backdoor attacks against federated learning", *International Conference on Learning Representations*.

Xie, M., *et al.* (2020), "Multi-center federated learning", arXiv preprint arXiv:2005.01026.

Xu, J., *et al.* (2021), "Federated learning for healthcare informatics", *Journal of Healthcare Informatics Research*, Vol. 5 No. 1, pp. 1-19.

Yang, J., *et al.* (2021), "Medmnist v2: a large-scale lightweight benchmark for 2d and 3d biomedical image classification", arXiv preprint arXiv:2110.14795.

Yang, Q., *et al.* (2019), “Federated machine learning: concept and applications”, *ACM Transactions on Intelligent Systems and Technology*, Vol. 10 No. 2, pp. 1-19.

Yoo, J.H., *et al.* (2021), “Personalized federated learning with clustering: Non-IID heart rate variability data application”, arXiv preprint arXiv:2108.01903.

Zeng, R., *et al.* (2020), “Fmore: an incentive scheme of multi-dimensional auction for federated learning in MEC”, *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). IEEE*, pp. 278-288.

Zhang, J., *et al.* (2019), “Poisoning attack in federated learning using generative adversarial nets”, *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE*, pp. 374-380.

Zhang, W., *et al.* (2021), “Dynamic-fusion-based federated learning for COVID-19 detection”, In *IEEE Internet of Things Journal*, Vol. 8 No. 21, pp. 15884-15891.

Zhao, H., Zhize, L. and Richt¨arik, P. (2021), “FedPAGE: a fast local stochastic gradient method for communication-efficient federated learning”, arXiv preprint arXiv:2108.04755.

**Corresponding author**
Tai-Myoung Chung can be contacted at: tmchung@skku.edu