SHFL

Asad, Muhammad; Aslam, Muhammad; Jilani, Syeda Fizzah; Shaukat, Saima; Tsukada, Manabu

Published in:
Future Internet

DOI:
10.3390/fi14110338

Published: 18/11/2022

Document Version
Publisher's PDF, also known as Version of record

future internet

MDPI

Article

SHFL: K-Anonymity-Based Secure Hierarchical Federated Learning Framework for Smart Healthcare Systems

Muhammad Asad [1,*,†], Muhammad Aslam [2,†], Syeda Fizzah Jilani [3,†], Saima Shaukat [1,†] and Manabu Tsukada [1,†]

[1] Graduate School of Information Science and Technology, Department of Creative Informatics, The University of Tokyo, Tokyo 113-8654, Japan
[2] School of Computing, Engineering, and Physical Sciences, University of the West of Scotland, Glasgow G72 0LH, UK
[3] Department of Physics, Aberystwyth University, Aberystwyth SY23 3FL, UK
* Correspondence: asad@g.ecc.u-tokyo.ac.jp
† These authors contributed equally to this work.

Abstract: Dynamic and smart Internet of Things (IoT) infrastructures allow the development of smart healthcare systems, which are equipped with mobile health and embedded healthcare sensors to enable a broad range of healthcare applications. These IoT applications provide access to the clients' health information. However, the rapid increase in the number of mobile devices and social networks has generated concerns regarding the secure sharing of a client's location. In this regard, federated learning (FL) is an emerging paradigm of decentralized machine learning that guarantees the training of a shared global model without compromising the data privacy of the client. To this end, we propose a K-anonymity-based secure hierarchical federated learning (SHFL) framework for smart healthcare systems. In the proposed hierarchical FL approach, a centralized server communicates hierarchically with multiple directly and indirectly connected devices. In particular, the proposed SHFL formulates the hierarchical clusters of location-based services to achieve distributed FL. In addition, the proposed SHFL utilizes the K-anonymity method to hide the location of the cluster devices. Finally, we evaluated the performance of the proposed SHFL by configuring different hierarchical networks with multiple model architectures and datasets. The experiments validated that the proposed SHFL provides adequate generalization to enable network scalability of accurate healthcare systems without compromising the data and location privacy.

Keywords: federated learning; K-Anonymity; privacy-preserving; hierarchical clustering

Citation: Asad, M.; Aslam, M.; Jilani, S.F.; Shaukat, S.; Tsukada, M. SHFL: K-Anonymity-Based Secure Hierarchical Federated Learning Framework for Smart Healthcare Systems. *Future Internet* **2022**, *14*, 338. https://doi.org/10.3390/fi14110338

Academic Editor: Hamid Mcheick

Received: 25 October 2022
Accepted: 18 November 2022
Published: 18 November 2022

1. Introduction

Modern communication infrastructure has recently undergone a paradigm shift owing to the development of the Internet of Things (IoT) [1]. Facilitating the interoperability of the digital nervous system has revolutionized several elements of urban living in smart cities [2]. IoT applications have considerable computational power and data processing capabilities, and are integrated with sophisticated computing devices such as sensors, wristbands, smartphones, and actuators. [3]. The service-driven computational abilities of intelligent devices can be effectively harnessed to design future smart healthcare systems [4,5]. Smartphones and pervasive technology-based healthcare systems provide a huge range of fitness applications and have the scalability to introduce clinical and medical systems [6]. According to the patient's health status, these clinical and medical healthcare systems offer remote healthcare services such as telemedicine and telecare. Integrating IoT-enabled healthcare systems enables the realization of many robust telehealth patient care applications and raises concerns over the secure and efficient communication of critical personal data [7–9]. The large scale of highly valued personal data increases privacy and security risks, and requires innovative networking techniques to maintain industrial-level communication efficiency [10,11].

The effective utilization of personal healthcare data are a critical requirement. Conventional machine-learning distributed algorithms are trained on existing datasets to track health information [12–14]. Three significant research challenges in modern healthcare systems need to be considered. First, real-time data are more critical and difficult to access because of the different privacy policies of isolated organizations. Secure and open access to these isolated data hubs is challenging when managing a central global learning model [15]. A distributed federated learning (FL) framework was designed to optimize the learning model, and has shown promising results [16]. Second, personalization is the most crucial issue. Different clients of smart healthcare systems use multiple devices with varying processing capacities [17]. Therefore, standard server-based learning models are unsuitable for maintaining personalization, as FL distributes the learning model among all the connected clients and provides flexibility to adjust the learning model according to the client's resources. However, personalization requires further investigation to enhance the performance of the distributed learning techniques. Third, all network nodes can never directly connect to the global learning model because of the network range and other communication constraints, which is not the case in FL because the global model is transmitted only to the clients participating in the network [18].

However, it is practically difficult for a global model to maintain direct connectivity with the client models [19]. A major problem in FL is scalability, as it distributes the learning model over all the devices and aggregates all the updates from the client models [20]. Currently, the advancement in positioning technology has led to the introduction of location-based service (LBS) technologies such as GPS, Baidu Map, and other wireless positioning technologies in intelligent devices, which facilitate the network administrator to create distance-based clusters of the participating devices [21]. These LBS technologies also compromise the location and pose a severe risk to clients with healthcare ratings. Location privacy, along with data privacy, is a critical challenge, particularly in the case of distributed FL [22]. Therefore, it is essential to design a solution that can hide the location to secure the personal information of clients.

To this end, we designed and investigated K-anonymity-based secure hierarchical federated learning (SHFL). Unlike existing solutions, we consider a realistic deployment of various healthcare application-aware IoT devices in which the local client (LC) devices directly communicate with the healthcare cloud/server. Every LC is an intermediate bridge for a few end devices called intra-local clients (ILCs). This unique learning technique provides network scalability by sharing the learning models of directly and indirectly connected devices with a global model trained from rich data. SHFL provides a loose federation of directly connected clients and allows the client devices to communicate and accommodate local models of the connected sub-clients. Each local model of the client and subclient devices has a local dataset to train the learning models on the local devices without interacting with the cloud server. The significant contributions of this study are summarized as follows.

1. We propose the SHFL framework for smart healthcare systems. In particular, we added K-anonymity-based location privacy along with data privacy of FL to anonymize the identity of the participating clients.
2. The proposed SHFL framework leverages the centralized server that communicates with multiple directly and indirectly connected devices hierarchically. Moreover, the proposed SHFL formulates the hierarchical clusters of LBSs to execute hierarchical FL.
3. The performance of the proposed SHFL was evaluated through extensive simulation experiments conducted with multiple model architectures and datasets.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss the studies that motivated us to conduct this research. In Section 3, we explain the system models of the proposed SHFL framework. In Section 4, we propose a K-anonymity-based secure hierarchical framework and demonstrate the application of the proposed SHFL. In Section 5, we conduct simulation experiments to demonstrate the performance of the

proposed SHFL in comparison with state-of-the-art approaches. Finally, we present the conclusions of this study in Section 6.

2. Related Work

FL is widely preferred over conventional centralized machine learning schemes, as it can guarantee privacy [16]. In FL, training is performed on local devices using their local datasets and aggregating them to the cloud server to produce a new global model. The method is iteratively repeated until the accuracy reaches the desired level [23]. Despite this safe training, FL suffers from considerable propagation delays, particularly in large-scale networks. The devices are distributed over large distances, resulting in communication bottlenecks [24]. A plethora of studies has been conducted to reduce the communication costs. A lossy compression technique to minimize the cloud-to-device communication overhead has been proposed [25]. The authors proposed communication-mitigated FL, where they selected only relevant updates to be forwarded for aggregation and achieved high performance [26]. Extensive communication often results in higher computational costs. To mitigate this challenge, the authors proposed a hybrid approach that simultaneously manages the communication cost through compression and reduces the computation cost through differential privacy [27]. Several studies have been conducted to enable distributed FL frameworks, which focus on the fixed architecture of the distributed FL [28,29]. In large-scale networks, many mobile devices participate in the training; these devices differ from each other in terms of datasets, computation capacity, and battery level, and can easily be managed through proper hierarchical task division [30]. In Table 1, we summarize the existing studies and their limitations.

Table 1. Summary of existing schemes and their limitations.

Reference	Limitations	Summary
[5]	Limited discussion on healthcare	A study on security vulnerabilities in the field of healthcare.
[7]	Single case study	A case study on patient health monitoring.
[9]	Non-personalized FL model	A study on client-edge-based FL framework for in-home health monitoring.
[20]	Limited to general architecture	A study on incentive mechanism for interaction between the crowdsourcing platform and the model training approach adopted by the client to optimize the communication efficiency.
[22]	Limited accuracy due to PSI protocol	A study to control COVID using cryptographic protocols to reduce the information revealed regarding the traces of positive users.
[23]	Non-personalized FL model	A study on FedAvg algorithm for non-identical and non-independent data.
[27]	Non-personalized FL model	A study on communication efficiency using sparse compression and privacy preservation using differential privacy in FL.
[28]	No-application specific	The study uses cosine similarity between the gradient updates for clustering in FL.
[30]	General hierarchical architecture	A study on hierarchical FL framework that uses edge nodes for additional model aggregation for faster communication.
[31]	Non-personalized FL model	A study on privacy preservation of FL using Pallier homomorphic cryptosystem.

Considering these limitations, we propose the SHFL framework for smart healthcare systems. The proposed SHFL framework is leveraged with a hierarchical architecture that distributes the network load evenly among all connected devices, which results in the efficient management of computational resources. To secure sensitive client information, we introduce K-anonymity in FL, which helps minimize security threats from adversaries in the network. Finally, the hierarchical architecture facilitates efficient communication between the clients and the cloud server. The experimental results prove that the proposed SHFL achieves higher accuracy than the existing solutions.

3. System Models

In this section, we define the essential system models for the proposed SHFL framework. We first describe the FL model that rigorously trains the local models on local devices

and obtains a global model. Second, we define the threat model, which is the primary motivation for this study. Finally, we describe the network model of the proposed SHFL.

3.1. FL Model

FL is a novel paradigm for building collaborative machine learning models through on-device distributed training on local datasets. This distributed training ensures the security and privacy of client data [32]. The proposed SHFL leverages the synchronous model update in each round of communication between the cloud server and clients, which minimizes the overall loss function by executing it in a distributed manner [33]. The synchronous model update involves the following three steps.

1. The LCs in the network independently train their models using their data.
2. After training the local models, the trained models are uploaded to the cloud server, which aggregates them to obtain a new global model.
3. The newly obtained global model is sent to the clients, who train their local models independently using new global parameters.

This process is repeated until the convergence condition is met.

3.2. Threat Model

In the proposed SHFL architecture, we assume that every client in the network is an honest but inquisitive adversary [31,34]. In other words, while attempting to access client-specific information in the training data during local model updates, all clients abide by the legal directives issued by the network administrator via the FL task. A malicious cloud server may deduce the client's sensitive data, or the infected client may leak the auxiliary information of other clients. Based on this assumption, the main goal of the proposed framework is to secure the client's private information, including location and personal information, while ensuring convergence accuracy.

3.3. Network Model

In the network model, we assume the realistic deployment of different healthcare application-aware IoT devices, in which LC devices directly communicate with the healthcare cloud/server. Every LC acts as an intermediate bridge for a few end-devices called ILCs. Figure 1 shows the general network settings, in which the network configurations are flexible according to the network requirements. The deployed devices are based on modern IoT infrastructure that can run local learning models over the rich data generated by the healthcare applications of individual devices.

We assume the deployment of $L = i : i = 1, 2, ...L$ IoT smart healthcare devices of intra-local clients (ILCs), which communicate with the global healthcare server with the help of $M = j : j = 1, 2, ...M$ intermediate local clients (LCs). M_j aggregates the updates from a particular associated set of L_i devices and uploads them to the cloud server, denoted as G. In a particular network communication round R, only $L_i \subseteq L$ online ILC devices actively participate in the hierarchical FL process. Each ILC and LC device contains its own dataset to train the local data models, represented as $S_i = (x_k, y_k)_{k=1}^{|S_i|}$ and $S_j = (x_k, y_k)_{k=1}^{|S_j|}$, respectively. Here, x_k represents the $k - th$ input sample, and y_k is the corresponding labeled output of x_k of the hierarchical FL. In the proposed hierarchical FL (HeirFL) model, all clients are static and remain active throughout the learning and aggregation processes. The network nodes share the learning parameters to support two-level aggregation, first at the intermediate CL devices and then at the cloud server. To measure the computational cost of the proposed model, the formulation of the energy and delay overhead is vital for cloud server aggregation.

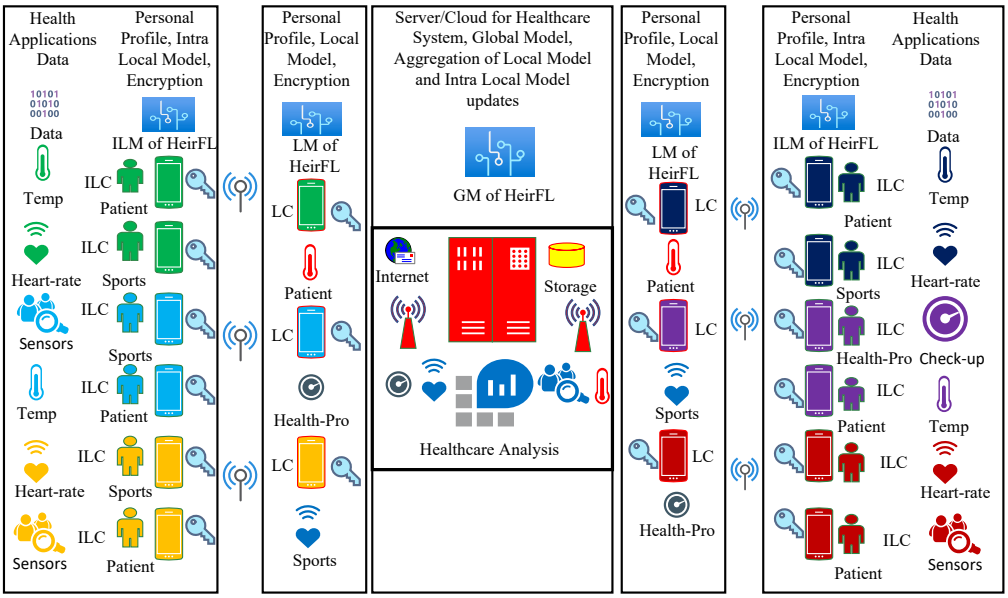


Figure 1. The proposed secure hierarchical federated learning (SHFL) architecture. A realistic deployment of different healthcare application-aware IoT devices is assumed, in which the LC devices directly communicate with the healthcare cloud/server. All the assumed devices are based on modern IoT infrastructure that can run local learning models over the rich data generated by the healthcare applications.

Based on the aforementioned FL model for the stepwise learning process, the model computation of the ILC devices is the first step to be considered. An active L_i ILC device computes the machine-learning model parameters ω, which classify each output value y_k for the input x_k of the dataset $S_i = (x_k, y_k)_{k=1}^{|S_i|}$. The loss function on $S_i = (x_k, y_k)_{k=1}^{|S_i|}$ of the L_i device is computed as

$$F_i(\omega)S_i = \frac{1}{|S_i|}\sum_{K=1}^{S_i} f_i(x_k, y_k, \omega). \tag{1}$$

To compute the ILC model accuracy $\Theta\epsilon(0,1)$, the ILC device L_i runs for a wide range of iterative algorithms according to the following local iterations:

$$Q(\Theta) = \mu \log(\frac{1}{\Theta}). \tag{2}$$

Furthermore, the data size S_i and machine-learning tasks affect the value of μ. L_i computes the local update at the $c-th$ iteration as

$$\omega_i^c = \omega_i^c - \eta \bigtriangledown F_i(\omega_i^{c-1}). \tag{3}$$

This learning process continues until the following computational results are obtained:

$$\|\bigtriangledown F_i(\omega_i^c)\| \leq \Theta\left\|\bigtriangledown F_i(\omega_i^{c-1})\right\|, \tag{4}$$

where η denotes a predefined learning rate. The number of required CPU cycles to process a single data sample of the device L_i is defined as D_i. Similarly, the total number of CPU cycles is calculated as $D_i|S_i|$. Thus, the total delay can be computed as

$$t_i^{cmpt-time} = Q(\Theta)\frac{D_i|S_i|}{f_n}. \tag{5}$$

Therefore, the computation can be implemented as

$$t_i^{cmpt-time} = \mu \log(\frac{1}{\Theta}) \frac{D_i \Big| S_i = (x_k, y_k)_{k=1}^{|S_i|} \Big|}{f_n}. \tag{6}$$

Similarly, the computational energy cost is calculated as follows:

$$e_i^{cmpt-cost} = Q(\Theta) \frac{\alpha_i}{2} f_n 2 D_i |S_i|, \tag{7}$$

$$e_i^{cmpt-cost} = \mu \log(\frac{1}{\Theta}) \frac{\alpha_i}{2} f_n 2 D_i \Big| |S_i = (x_k, y_k)_{k=1}^{|S_i|} \Big|. \tag{8}$$

When all the L ILC devices complete their local iterations, each device L_i uploads the model parameters ω of the ILC to the associated intermediate local client M_j. The intermediate M_j LC receives the updates from a connected set of active A_i devices and performs the aggregation process as follows:

$$\overline{\omega_i} = \frac{\sum i \epsilon A_{Ai} |S_i| \omega_i^c}{\left|S_{A_i}\right|}, \tag{9}$$

where the aggregated dataset of A_i connected to the M_j intermediate LC is

$$S_{A_i} = \cup_{i \epsilon A_i} S_i. \tag{10}$$

The intermediate LC M_j multicasts $\overline{\omega_i}$ to A_i for the computation process in the next communication round R. In particular, M_j repeats the process and guides the ILC devices iteratively to compute the model for better accuracy. The aggregation process of M_j continues until it reaches the accuracy level of all the other LC devices connected to their M. The required model accuracy of the general convex machine learning tasks and the overall number of iterations on M_j are

$$I(\varepsilon, \Theta) = \delta \frac{(\log(\frac{1}{\varepsilon}))}{1 - \Theta}, \tag{11}$$

where δ is a predefined constant of the learning tasks. HeirFL shows that the intermediate LCs access only the parameters instead of the actual data of the end devices, which significantly improves the security and privacy of personal data. After $I(\varepsilon, \Theta)$ iterations over M_j devices, the total energy cost of dealing with A_i ILC can be computed as

$$E_{j:i}^{cmpt-cost} = \sum_{j:i} \epsilon A_i I(\varepsilon, \Theta)(e_{j:i}^{cmpt-cost} + e_i^{cmpt-cost}). \tag{12}$$

Likewise, the computation and communication delay for M_j can be calculated as

$$T_{j:i}^{cmpt-cost} = I(\varepsilon, \Theta) max_{i \epsilon A_i} t_{j:i}^{cmpt-cost} + t_i^{cmpt-cost}. \tag{13}$$

In the last step, the cloud server gathers all the updated models received from the edge servers as follows:

$$\overline{\omega} = \frac{\sum j \epsilon M \left|S_j\right| \omega_j}{\left|S_j\right|}, \tag{14}$$

where the aggregated dataset of M_j connected to the global cloud server G is

$$S = \cup_{j \epsilon M_j} S_j. \tag{15}$$

Therefore, we may derive the system-wide energy and latency under a single global iteration, disregarding the cloud's substantially shorter aggregation time than that of mobile devices, as shown below.

$$E = \sum_{j} \epsilon M_j I(\varepsilon, \Theta)(E^{server} + E_j^{cmpt-cost}), \tag{16}$$

$$T = MAX\left\{T^{server} + T_j^{cmpt-cost}\right\}. \tag{17}$$

To meet the convergence conditions, the local and global-level aggregation can be repeated consistently to maximize the accuracy and reduce the loss.

4. K-Anonymity-Based SHFL

Different LBSs are integrated with modern IoT-enabled mobile devices; hence, the natural hierarchical distribution of learning tasks is based on LBSs. However, healthcare data demand a higher level of privacy protection. Hence, in this section, we present the SHFL framework for innovative healthcare systems. The proposed SHFL framework develops a secure hierarchical connection between the server model G and the local models of the directly connected devices M_j. In addition, the framework allows these devices to extend the network connectivity by configuring the clusters of intra-local models of the nearby devices L_i. The proposed SHFL framework has two technical tasks: first, develop LBS-based suitable hierarchical clusters of the participating M_j and L_i devices; second, maintain the location and data privacy of all active devices.

4.1. Secure Hierarchical Distributed Architecture

M_j devices act as the primary entities in the distributed architecture, and the global model is directly connected to a set $M_j = M_jID, LOC, APP, PPV, CS$, where M_jID represents the personal identification number, LOC is the location of M_j, APP is the application of the client being trained, PPV indicates the privacy-preserving value, where the nodes can define the degree of anonymity in a distributed manner. In contrast, CS represents the cluster size managed by a specific M_j. Similarly, $L_i = L_iID, LOC, APP, PPV$ represents the set of intra-local model clients. The magnitude of distributed clustering is based on LBSs, and it is dynamic in a realistic network environment. Furthermore, the global model fixes the cluster size to optimize the efficiency of the hierarchical architecture. The other two major factors of cluster formulation are similarity index and privacy index of the intra-local model devices. The similarity index of the nodes can be computed using the factors of communication cost, energy cost, application similarity, and previous participation score in FL. The security index can be calculated using the degree of K-anonymity as PPV.

4.1.1. Clustering Index of SHFL

The proposed SHFL framework formulates centralized clustering using a global model trained on the server. GPS-equipped smart devices voluntarily participate in hierarchical FL and exchange key parameters to initialize the clustering process. The primary task of the global model is to select the anchor device that can directly transmit the information of the nodes to the cluster members for training after cluster aggregation. As mentioned in the network model of SHFL, the global model receives the computation and energy calculations from the network nodes. The global model computes the similarity index by using the following equation:

$$Similarity - Index = \frac{R_e \times A_{Pc}}{d}, \tag{18}$$

where R_e is the residual energy resource, which indicates the primary availability index of the device, A_{Pc} represents the available processing capacity, and d is the distance of the node from the global network model. The global model utilizes the threshold of the similarity index to determine the rule of the network device as an anchor node or member

node. APP and PPV also play critical roles in selecting the anchor nodes. All anchor nodes act as local model nodes M_j, and member nodes act as intra-local model nodes L_i. This initial cluster formulation provides the basic setting for K-anonymity-based regrouping to establish anonymous clusters.

4.1.2. K-Anonymity-Based Regrouping of Anonymous Clusters of SHFL

In the SHFL architecture, we integrate the K-anonymous central server, which ensures the anonymity of LBS-enabled smart devices. This K-anonymous architecture also contains a GPS server and database server, where the global network model operates throughout the network operation. Figure 2 shows the detailed process of the integrated K-anonymous central server. The communication between the LCs and K-anonymous central server is protected through encryption technology. The following five-step process performs this anonymity query, which is formalized in Algorithm 1:

1. The directly connected LCs M_j transmit the initial dialogue with a trusted anonymous server and deliver their demand of anonymity with the degree of PPV and its corresponding CS cluster size.
2. For anonymous re-clustering of the devices, the anonymous server computes the anonymous results for the set M_j using the K-Anonymity algorithm and sends the results to the LBS server.
3. According to the position information provided by M_j and the recommendation of the K-anonymous algorithm, the LBS and database server continue the anonymous query procedure. The LBS server regroups the nodes and formulates the set U_j. Similarly, the LBS processes the re-clustering demand of the intra-local devices connected to M_j in the form of V_i.
4. The outcome of the LBS server and database server is delivered to the K-anonymous server.
5. Finally, the K-anonymous server re-checks the sets U_j and V_i according to the actual locations. It transmits the K-anonymous replies to the network devices.

Algorithm 1: Generation of anonymous result sets.

Input : Adapted sets of results, $V^{'} = \{V_1^{'}, V_2^{'}, ..., V_m^{'}\}$, $|V_i^{'}| \geq k_{min}$, i = 1,..., m

Output: Anonymous result sets, $V^{''} = \{V_1^{''}, V_2^{''}, ..., V_m^{''}\}$, $|V_i^{''}| \geq k_{min}$, i = 1,..., m

Initialization

Steps

for $V_i^{''}$ **do**

- $V_i^{''}$ obtains a set of client identifications m_i and a set of queries in the anonymous group $C_i^{'}$ by determining the highest value A_{min} to compare and obtain r as $A_{min} = \pi \times r2$;
- $e = Get\_Head\_Item\ (V_i^{''})$;
- if($r > e.radius$) e.radius = r; // Adjust the radius;
- The object e identifier is included in m_i; the query information is included in S_i, and S_i is fixed to the area of a circle having a radius equivalent to e.radius. Now, e can be anchored to the construct $V^{''}$;

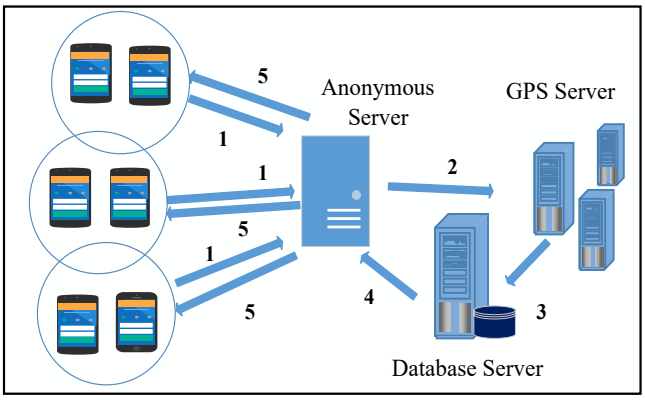


Figure 2. Privacy preserving K-Anonymity architecture integrated with SHFL.

After ensuring privacy preservation using the K-anonymity method, the following anonymous set $U_j = u_jID, LOC, AS$ of the local models appears as the available network statistics. u_jID is the anonymous ID, LOC is the updated location, and AS is the area of K-anonymity. Similarly, $V_i = u_iID, LOC, AS$ is a set of intra-local models that updates the anonymous profile to save the privacy of the client device. The K-anonymity algorithm performs significantly better once the clustered and corresponding anonymous groups contain minimum outliers. The outliers are devices with different query statistics that make the anchor device selection process of the cluster complex. The elimination of outliers simplifies the anonymous grouping process. To detect outliers, we represent the client space as ZT. d is the radial distance of x from the corresponding y, which can be denoted as $K - dist\ (d)$. Similarly, $K - dist\ (r)$ neighborhood of the object d is denoted by $V_iK(r)$. Any other device within the radius d of k is denoted as $dist\ (x, y)$.

1. **Reachability distance** (ReachDist) is calculated by the following equation:

$$ReachDist_Kd(x, y) = Max(K - dist(x), d(x, y). \tag{19}$$

2. **K-nearest neighbor (K-NN) distance** in the space ZT can be defined as follows:

$$DistKN(x, ZT) = \frac{\sum_{I=K}^{K} DistK(i)}{k}. \tag{20}$$

3. **The density of object** x in the client space ZT is defined as follows, where k denotes the density and x denotes the object:

$$DenK(x, ZT) = \frac{1}{DistKN(x, ZT)}. \tag{21}$$

4. **The Local Reachability Density** (LRD) of the device is the opposite of the average RD of K-NN based on the device x. The following equation calculates the LRD:

$$LRD_k(x, ZT) = \frac{|N_k(x)|}{\sum_{y \epsilon N_k(x) Reachdist_k(x,y)}}. \tag{22}$$

5. **Local Outlier Factor** (LOF) characterizes x as an outlier, which is calculated by the following equation:

$$LOF_k(x) = \frac{\sum_{y \epsilon N_k(x)} \frac{LRD_k(y,ZT)}{LRD_k(x,ZT)}}{|N_k(x)|}. \tag{23}$$

6. **Centrifugal degree of the anonymous group** is the average distance between the anchor and other points in the anonymous group. If the anonymous group C uses m as its anchor, the centrifugal degree (Cd) of m can be calculated as follows:

$$Cd(c) = \frac{\sum_{x \epsilon C} distance(x, z)}{|N_k(x)|}. \quad (24)$$

In the proposed SHFL, a cluster loop is constructed, and the dataset is divided equally among all U_j nodes and V_i member nodes of the anchors. The anchor nodes represent the LCs with local models, and the member nodes are V_i with intra-local models. The datasets were initialized over all nodes for training and testing. The anchor nodes react according to the queries of the member nodes, aggregate the updates, and forward them to the cloud server after achieving the required degree of anonymity. The clustering process continues for newly arrived nodes over the network, along with their anonymity handling. In addition, the anchor nodes utilize the above-mentioned parameters to check the node's location, neighbor distance, and k-neighbor density to determine the LOF and centrifugal degree. The initial anonymous group dataset V_i and completion of the initial partition of the anonymous group follow the traversal of all these parameters. Anonymous groups can be modified at runtime to remove the outliers to enhance the query service quality. Additionally, the anonymous groups beneath the U_j anchors are ranked in the reverse order according to the centrifugal degree. The cluster radius and dynamic locations of the nodes also play crucial roles in adjusting the group anonymity. This anonymous group adjustment is executed periodically after the anonymity process of the anonymous server and LBS servers.

5. Experiments

Model Definition: In the experiment, we performed the image classification tasks over distributed federated settings. We selected the full-size MNIST [35] and CIFAR-10 [36] benchmark datasets to investigate the proposed SHFL and validate the usability of the smart devices in the IoT. In addition, we utilized the COVID-19 [37] dataset to develop a healthcare application of SHFL. We performed extensive simulation experiments by varying the computation rate of LCs and ILCs. To evaluate the performance of SHFL in real-time healthcare applications, we altered different parameters to analyze the performance comprehensively. We constructed a CNN client model for all three datasets by creating 5×5 convolutional layers. The model divides the dataset into equal-sized shards according to the number of LCs and ILCs. Table 2 lists the benchmark hyperparameters.

System Configuration: Our simulation experiments were conducted on a CPU i9-9980HK @ 2.40 GHz with 32 GB RAM. The SHFL framework was designed using Python in TensorFlow.

Data Distribution: In all our experiments, we used the pathological $non - iid$ distribution of data, where each LC can only receive images corresponding to eight labels; thus, each LC receives minimum 300 samples. The data were divided into distinct clusters using the $non - iid$ distribution, and each cluster was evenly split across numerous LCs and ILCs.

Table 2. Hyperparameters used in the proposed experiments.

	MNIST	**CIFAR-10**	**COVID-19**
Parameter		Values	
Model	CNN	CNN	AlexNet
Momentum	0.5	0.3	0.1
Optimizer	SGD	SGD	SGD
Batch size	10	20	30
Learning rate	0.25	0.5	0.5
Clients	500	200	200
Client transmission power	200 mW	200 mW	200 mW
Communication rounds	300	500	200
Local epochs	200	200	100
Local update size	20,000 nats	20,000 nats	20,000 nats

5.1. Convergence

To explore the convergence behaviors of the proposed SHFL framework, we considered two scenarios of the number of LCs as $\{10, 15\}$ and four scenarios of the number of ILCs as $\{2, 5, 8, 10\}$, and tested the accuracy and loss on the MNIST and CIFAR-10 datasets. Each LC and ILC were assigned an equal-sized random subset of the training data. We continuously re-divided the dataset into equal parts to train the local model of all devices. We simulated these experiments for 300 and 500 communication rounds on MNIST and CIFAR-10, respectively, to investigate the trends of accuracy and loss of the distributed FL. Each client's LBS was altered by randomly swapping out eight labels depending on the cluster to which the client belonged, to maintain the anonymity of the LC and ILC structures. If all clients are part of the first cluster, the data points labeled "1" and "7" might be switched. Similarly, for the clients in the second cluster, the data points labeled "3" and "5" will be switched, and so on. In Figure 3, we present a realistic picture of the proposed SHFL, which shows the hierarchical architecture for multiple levels of LCs and ILCs.

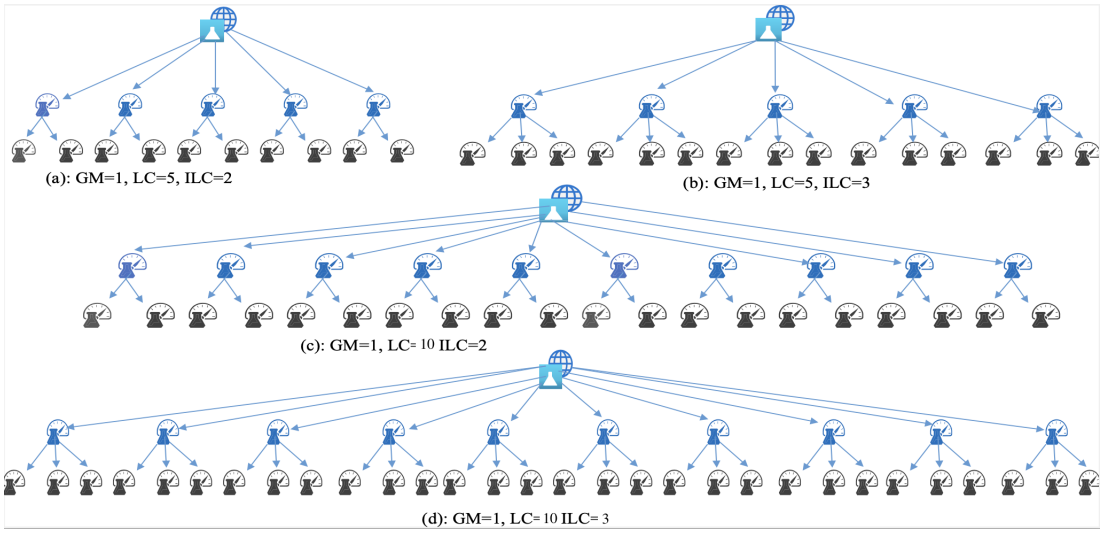
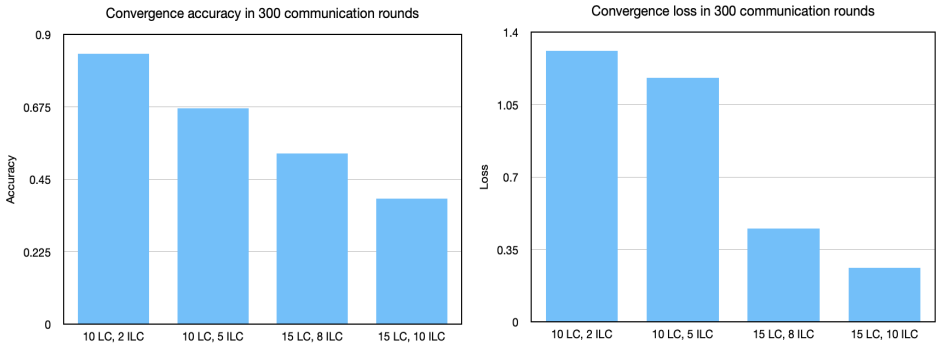


Figure 3. Network topologies for distributed federated learning of SHFL; the hierarchical architecture for multiple levels of LC and ILC is shown.

Figure 4 shows the simulation results of the four configuration settings of the LC and ILC in terms of accuracy and loss for the communication rounds on the MNIST dataset.

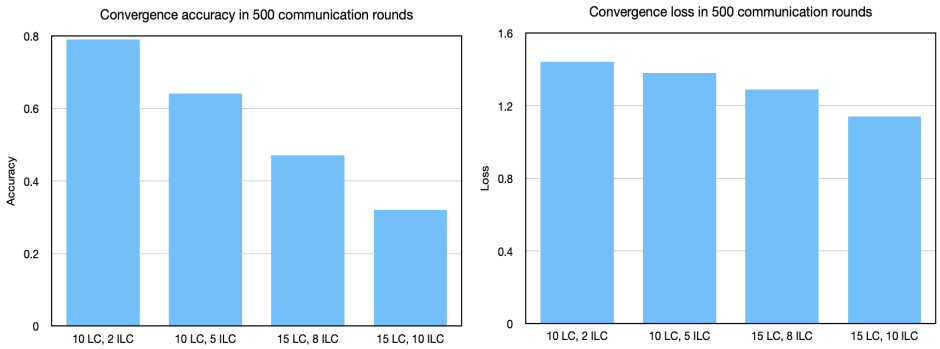
The proposed model performs well in terms of high accuracy rate and low loss in the case of a less complicated network architecture with a limited clustering formulation size. This is because the network nodes grow with the number of clusters and the cluster size increases proportionally. In this regard, the proposed SHFL provides higher network scalability and higher accuracy rate with increasing network communication rounds. In Figure 4a, the network topology of 10 LCs and 2 ILCs results in 55% accuracy in only 300 communication rounds; furthermore, 15 LCs and 10 ILCs produce 85% accuracy. Similarly, in Figure 4b, the network topology of 10 LC, 2 ILC shows a higher loss, as a greater number of ILCs requires more computational resources. In the case of 10 LC, 2 ILC, the graph shows the minimum loss, which increases with increasing numbers of LCs and ILCs.



(**a**) (**b**)

Figure 4. Comparison of SHFL training accuracy and training loss for different numbers of LCs and ILCs in the MNIST dataset, (**a**) training accuracy over MNIST; (**b**) training loss over MNIST.

The experimental results for the CIFAR-10 dataset are shown in Figure 5. In Figure 5a, we present the training accuracy, and in Figure 5b, we present the training loss of SHFL in 500 communication rounds. To provide a better comparison, we considered the same scenarios as for the MNIST dataset. The remaining parameters were the same as those described in Table 2. Figure 5a shows higher accuracy for higher numbers of LCs and ILCs; the accuracy decreases with decreasing numbers of LCs and ILCs. We attribute this performance gain to the hierarchical architecture of SHFL, which divides the computational resources equally among the clients. Furthermore, the degree of K-anonymity helps secure the maximum accuracy. Similarly, Figure 5b shows the minimum expected loss for the scenario of 10 LCs and 2 ILCs. As described above, a higher number of LCs and ILCs requires greater computational resources. Therefore, we can say that the optimal number of LCs and ILCs is between ≈ 12 and ≈ 7.



(**a**) (**b**)

Figure 5. Comparison of SHFL training accuracy and training loss for different numbers of LCs and ILCs on the CIFAR-10 dataset, (**a**) training accuracy over CIFAR-10; (**b**) training loss over CIFAR-10.

To further prove the accuracy of the proposed SHFL, we used the healthcare dataset related to X-rays for COVID-19. The dataset was divided into three classes: COVID-19, Pneumonia, and No findings. We utilized the 5×5 convolution layer cross-validation procedure for both binary and triple classification problems. Twenty percent of the X-ray images were used for testing, and the remaining was used for training the model. As shown in Figure 6, we trained the LCs and ILCs for 100 local epochs and tested the accuracy of 200 communication rounds for three network topologies: {5 LCs and 5 ILCs}, {15 LCs and 5 ILCs}, and {15 LCs and 15 ILCs}. Different topologies provide different levels of fluctuations in terms of accuracy. The end clients of distributed learning can provide their X-ray data in real time and obtain an initial probabilistic diagnosis. These results are for initial examinations and are mainly applicable to medical professionals. In Figure 6, the overlapped confusion matrix (CM) is shown for the following network topologies: (a) {5 LCs and 5 ILCs}, (b) {15 LCs and 5 ILCs}, and (c) {15 LCs and 15 ILCs}. The overlapped confusion matrix was created using the sum of the client models of all the folds. The SHFL model utilized the AlexNet client model to classify the training and testing of the COVID-19 dataset. In the case of the network topology of {5 LCs and 5 ILCs}, the proposed SHFL achieved 93% accuracy for COVID-19 detection and 88% accuracy for normal or no findings. Similarly, we achieved 81% accuracy in pneumonia detection. In the case of the network topology of {15 LCs and 5 ILCs}, we achieved 96% accuracy for COVID-19 detection and recognition. The network topology of {15 LCs and 15 ILCs} is very dense but still maintains an accuracy of 93% for COVID-19 detection. We are currently training and testing our proposed SHFL framework for a healthcare application for COVID-19 detection and have achieved high accuracy. However, we note that our research is currently independent. COVID-19 detection is a sensitive research field, and thus far, we have utilized only simulation-based experiments on publicly available COVID-19 datasets.

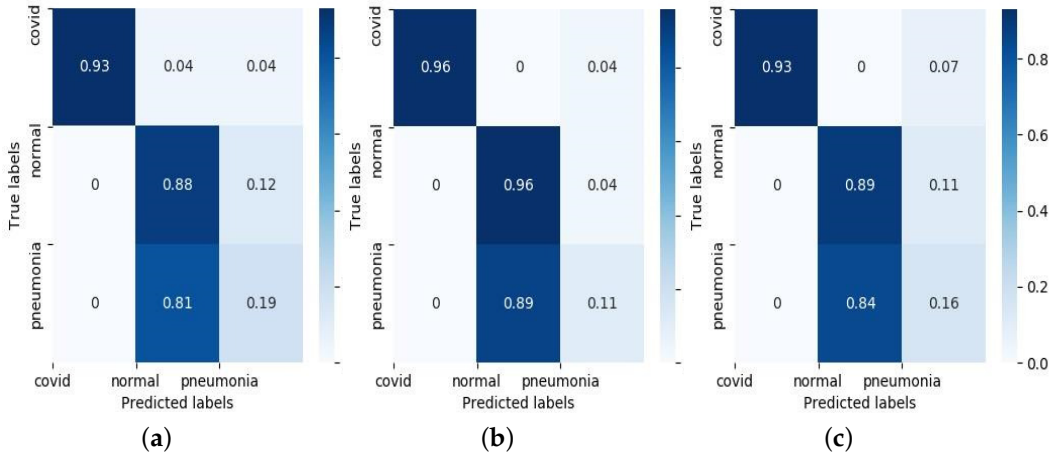
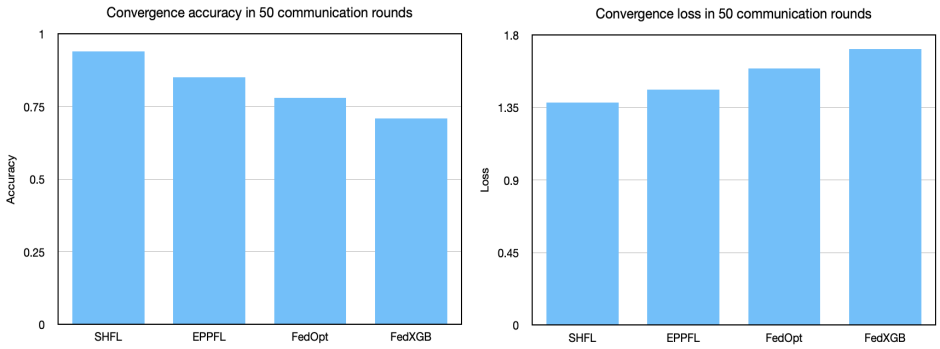


Figure 6. Overlapped and 5-fold confusion matrix results for (**a**) 5 LCs, 5 ILCs, (**b**) 15 LCs, 2 ILCs, and (**c**) 15 LCs, 5 ILCs.

5.2. Comparison

We selected three schemes related to privacy preservation in FL for comparison with the proposed SHFL. The schemes selected for comparison are Efficient and Privacy-Preserving Federated Learning (EPPFL) [38], Federated Optimization (FedOpt) [27], and Extreme Boost Federated Learning (FedXGB) [39]. We compared the existing schemes with the proposed SHFL in terms of the training accuracy and expected loss. For comparison, we selected the MNIST dataset, as it is utilized in existing schemes. In addition, we used the same CNN architecture for all compared schemes. A random distribution divided the MNIST dataset into 25% for testing and 75% for training. The remaining parameters were set to the values given in Table 2. Figure 7 shows the training accuracy and expected loss with an increasing number of communication rounds. Figure 7a shows that the proposed SHFL

achieves 3%, 7%, and 9% greater accuracy than EPPFL, FedOpt, and FedXGB, respectively. Similarly, in Figure 7b, SHFL has less expected loss than EPPFL, FedOpt, and FedXGB by approximately 6%, 9%, and 12%, respectively. We attributed this performance to the hierarchical architecture of the proposed SHFL. In addition, the anonymity of LCs and ILCs enables the SHFL in securing higher performance.



(**a**) (**b**)

Figure 7. Convergence comparison of the proposed SHFL with state-of-the-art techniques, (**a**) comparison accuracy over MNIST; (**b**) comparison loss over MNIST.

6. Conclusions

This paper presented an SHFL framework for smart healthcare systems. The proposed SHFL adopts a hierarchical FL approach, in which a centralized server communicates with participating clients who are further connected with the sub-clients. All clients are distributed in clusters using a similarity-based index and privacy-based index to execute distributed FL. In addition, the proposed SHFL introduces a K-anonymity method that hides the location and identity of the clients in the cluster. The communication between the K-anonymous central server and clients is protected using encryption technology. We conducted detailed experiments on the commonly used FL datasets to demonstrate the performance of the proposed SHFL in terms of convergence accuracy. The results prove that the proposed SHFL performs significantly better than the state-of-the-art FL approaches. Communication efficiency is a significant concern in FL, where SHFL needs to minimize the communication cost by using either compressed updates or sparse data, which could be the focus of our future work. Moreover, privacy attacks will be considered in our future work.

Author Contributions: Conceptualization, S.S.; Methodology, M.A. (Muhammad Asad) and S.S.; Formal analysis, M.A. (Muhammad Aslam); Writing—original draft, M.A. (Muhammad Asad); Writing—review & editing, M.A. (Muhammad Asad), M.A. (Muhammad Aslam), S.F.J., S.S. and M.T.; Visualization, M.A. (Muhammad Aslam), S.F.J. and S.S.; Supervision, M.T.; Project administration, S.F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was commissioned by the National Institute of Information and Communications Technology (NICT), JAPAN.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest regarding the publication of this article.

References

1. Bi, Z.; Da Xu, L.; Wang, C. Internet of things for enterprise systems of modern manufacturing. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1537–1546.
2. Chiuchisan, I.; Chiuchisan, I.; Dimian, M. Internet of Things for e-Health: An approach to medical applications. In Proceedings of the 2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), Prague, Czech Republic, 29–30 October 2015; pp. 1–5.

3. Khatua, P.K.; Ramachandaramurthy, V.K.; Kasinathan, P.; Yong, J.Y.; Pasupuleti, J.; Rajagopalan, A. Application and assessment of internet of things toward the sustainability of energy systems: Challenges and issues. *Sustain. Cities Soc.* **2020**, *53*, 101957. [CrossRef]
4. Jeong, J.S.; Han, O.; You, Y.Y. A design characteristics of smart healthcare system as the IoT application. *Indian J. Sci. Technol.* **2016**, *9*, 52. [CrossRef]
5. Chen, C.M.; Liu, S.; Chaudhry, S.A.; Chen, Y.C.; Khan, M.A. p A Lightweight and Robust User Authentication Protocol with User Anonymity for IoT-Based Healthcare. *CMES-Comput. Model. Eng. Sci.* **2022**, *131*, 307–329. [CrossRef]
6. Manogaran, G.; Varatharajan, R.; Lopez, D.; Kumar, P.M.; Sundarasekar, R.; Thota, C. A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. *Future Gener. Comput. Syst.* **2018**, *82*, 375–387. [CrossRef]
7. Abawajy, J.H.; Hassan, M.M. Federated internet of things and cloud computing pervasive patient health monitoring system. *IEEE Commun. Mag.* **2017**, *55*, 48–53. [CrossRef]
8. Patel, W.; Pandya, S.; Mistry, V. i-MsRTRM: Developing an IoT based Intelligent Medicare system for Real-Time Remote Health monitoring. In Proceedings of the 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), Tehri, India, 23–25 December 2016; pp. 641–645.
9. Wu, Q.; Chen, X.; Zhou, Z.; Zhang, J. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans. Mob. Comput.* **2020**, *21*, 2818–2832. [CrossRef]
10. Ali, W.; Din, I.U.; Almogren, A.; Guizani, M.; Zuair, M. A lightweight privacy-aware iot-based metering scheme for smart industrial ecosystems. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6134–6143. [CrossRef]
11. Islam, A.; Al Amin, A.; Shin, S.Y. FBI: A federated learning-based blockchain-embedded data accumulation scheme using drones for Internet of Things. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 972–976. [CrossRef]
12. Gribbestad, M.; Hassan, M.U.; A Hameed, I.; Sundli, K. Health Monitoring of Air Compressors Using Reconstruction-Based Deep Learning for Anomaly Detection with Increased Transparency. *Entropy* **2021**, *23*, 83. [CrossRef]
13. Malik, S.; Rouf, R.; Mazur, K.; Kontsos, A. The Industry Internet of Things (IIoT) as a Methodology for Autonomous Diagnostics in Aerospace Structural Health Monitoring. *Aerospace* **2020**, *7*, 64. [CrossRef]
14. Harweg, T.; Peters, A.; Bachmann, D.; Weichert, F. CNN-Based Deep Architecture for Health Monitoring of Civil and Industrial Structures Using UAVs. *Multidiscip. Digit. Publ. Inst. Proc.* **2019**, *42*, 69.
15. Zhu, X.; Li, H.; Yu, Y. Blockchain-based privacy preserving deep learning. In *Proceedings of the International Conference on Information Security and Cryptology*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 370–383.
16. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. Federated learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2019**, *13*, 1–207.
17. Rauch, G.; Röhmel, J.; Gerß, J.; Scherag, A.; Hofner, B. Current challenges in the assessment of ethical proposals-aspects of digitalization and personalization in the healthcare system. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **2019**, *62*, 758–764. [CrossRef] [PubMed]
18. Duan, M.; Liu, D.; Chen, X.; Liu, R.; Tan, Y.; Liang, L. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Trans. Parallel Distrib. Syst.* **2020**, *32*, 59–71. [CrossRef]
19. Sun, L.; Qian, J.; Chen, X.; Yu, P.S. Ldp-fl: Practical private aggregation in federated learning with local differential privacy. *arXiv* **2020**, arXiv:2007.15789.
20. Pandey, S.R.; Tran, N.H.; Bennis, M.; Tun, Y.K.; Manzoor, A.; Hong, C.S. A crowdsourcing framework for on-device federated learning. *IEEE Trans. Wirel. Commun.* **2021**, *19*, 3241–3256. [CrossRef]
21. Thrun, M.C.; Ultsch, A. Using Projection-Based Clustering to Find Distance-and Density-Based Clusters in High-Dimensional Data. *J. Classif.* **2020**, 1–33 . [CrossRef]
22. Berke, A.; Bakker, M.; Vepakomma, P.; Raskar, R.; Larson, K.; Pentland, A. Assessing disease exposure risk with location histories and protecting privacy: A cryptographic approach in response to a global pandemic. *arXiv* **2020**, arXiv:2003.14412.
23. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv* **2019**, arXiv:1907.02189.
24. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]
25. Caldas, S.; Konečny, J.; McMahan, H.B.; Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv* **2018**, arXiv:1812.07210.
26. Luping, W.; Wei, W.; Bo, L. CMFL: Mitigating communication overhead for federated learning. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 954–964.
27. Asad, M.; Moustafa, A.; Ito, T. FedOpt: Towards communication efficiency and privacy preservation in federated learning. *Appl. Sci.* **2020**, *10*, 2864. [CrossRef]
28. Sattler, F.; Müller, K.R.; Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3710–3722. [CrossRef]
29. Chen, M.; Poor, H.V.; Saad, W.; Cui, S. Wireless communications for collaborative federated learning. *IEEE Commun. Mag.* **2020**, *58*, 48–54. [CrossRef]
30. Liu, L.; Zhang, J.; Song, S.; Letaief, K.B. Client-edge-cloud hierarchical federated learning. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–6.

31. Zhang, J.; Chen, B.; Yu, S.; Deng, H. PEFL: A privacy-enhanced federated learning scheme for big data analytics. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
32. Chen, Y.; Luo, F.; Li, T.; Xiang, T.; Liu, Z.; Li, J. A training-integrity privacy-preserving federated learning scheme with trusted execution environment. *Inf. Sci.* **2020**, *522*, 69–79. [CrossRef]
33. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, H.B.; et al. Towards federated learning at scale: System design. *arXiv* **2019**, arXiv:1902.01046.
34. Hao, M.; Li, H.; Luo, X.; Xu, G.; Yang, H.; Liu, S. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Trans. Ind. Inform.* **2019**, *16*, 6532–6542. [CrossRef]
35. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision Proceedings of the Identity Mappings in Deep Residual Networks*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
37. Zhao, J.; Zhang, Y.; He, X.; Xie, P. Covid-ct-dataset: A ct scan dataset about covid-19. *arXiv* **2020**, ArXiv:2003.13865.
38. Hao, M.; Li, H.; Xu, G.; Liu, S.; Yang, H. Towards efficient and privacy-preserving federated deep learning. In Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–6.
39. Liu, Y.; Ma, Z.; Liu, X.; Ma, S.; Nepal, S.; Deng, R. Boosting privately: Privacy-preserving federated extreme boosting for mobile crowdsensing. *arXiv* **2019**, arXiv:1907.10218.