symmetry

MDPI

Article

Optimizing Multi-Objective Federated Learning on Non-IID Data with Improved NSGA-III and Hierarchical Clustering

Jialin Zhong, Yahui Wu, Wubin Ma *, Su Deng and Haohao Zhou

Science and Technology on Information System Engineering Laboratory, National University of Defense Technology, Changsha 410073, China; zhongjialin20@nudt.edu.cn (J.Z.); yahui_wu@nudt.edu.cn (Y.W.); sudeng@nudt.edu.cn (S.D.); haohaozhou@nudt.edu.cn (H.Z.)

* Correspondence: wb_ma@nudt.edu.cn

Abstract: Federated learning (FL) can tackle the problem of data silos of asymmetric information and privacy leakage; however, it still has shortcomings, such as data heterogeneity, high communication cost and uneven distribution of performance. To overcome these issues and achieve parameter optimization of FL on non-Independent Identically Distributed (non-IID) data, a multi-objective FL parameter optimization method based on hierarchical clustering and the third-generation non-dominated sorted genetic algorithm III (NSGA-III) algorithm is proposed, which aims to simultaneously minimize the global model error rate, global model accuracy distribution variance and communication cost. The introduction of a hierarchical clustering algorithm on non-IID data can accelerate convergence so that FL can employ an evolutionary algorithm with a low FL client participation ratio, reducing the overall communication cost of the NSGA-III algorithm. Meanwhile, the NSGA-III algorithm, with fast greedy initialization and a strategy of discarding low-quality individuals (named NSGA-III-FD), is proposed to improve the convergence efficiency and the quality of Pareto-optimal solutions. Under two non-IID data settings, the CNN experiments on both MNIST and CIFAR-10 datasets show that our approach can obtain better Pareto-optimal solutions than classical evolutionary algorithms, and the selected solutions with an optimized model can achieve better multi-objective equilibrium than the standard federated averaging (FedAvg) algorithm and the Clustering-based FedAvg algorithm.

Keywords: federated learning; multi-objective optimization; NSGA-III; parameter optimization

Citation: Zhong, J.; Wu, Y.; Ma, W.; Deng, S.; Zhou, H. Optimizing Multi-Objective Federated Learning on Non-IID Data with Improved NSGA-III and Hierarchical Clustering. *Symmetry* **2022**, *14*, 1070. https://doi.org/10.3390/sym14051070

Academic Editors: Aviv Gibali and Mihai Postolache

Received: 11 April 2022
Accepted: 19 May 2022
Published: 23 May 2022

1. Introduction

While the rapid development of artificial intelligence has brought great convenience to society [1], such as smart healthcare [2] and intelligent transportation [3], its development is facing new difficulties and challenges [4], such as data silos [5] and privacy leaks. Traditional, centralized machine learning needs to gather scattered data before training, but in fact, it is difficult to aggregate data in many fields. For example, due to privacy considerations, hospitals cannot share data for machine learning, resulting in asymmetric information and data silos that put a limit on sharing data between different organizations. In addition, as people's awareness of privacy protection has gradually increased, countries around the world have also issued privacy protection laws and regulations, such as the General Data Protection Regulation (GDPR) [6] in the European Union and the Personal Data Protection Act (PDPA) [7] in Singapore, which place severe restrictions on the sharing of sensitive data.

Therefore, federated learning (FL) [8] has emerged as a viable solution to the problems of data silos of asymmetric information and privacy leaks. FL can train a global model without extracting data from a client's local dataset. After downloading the current global model from the server, each client trains the global model on the local data, and then uploads the trained local model to the server to aggregate as the latest global model. After

iterations, FL can finally obtain a global model while effectively protecting the user's privacy by avoiding sharing locally private data. In addition, the federated averaging algorithm (FedAvg) is a classical algorithm of federated learning, which is used to update the global model on the server by obtaining the average values of the parameters collected from clients.

The research on federated learning continues to deepen, but it still faces the challenges of data heterogeneity, high communication cost and structural heterogeneity [9]. Each client's data is usually non-independent and identically distributed (non-IID), that is, the number of data labels and data size between clients are asymmetric, which may damage the accuracy of FL. The parameter transmission between the server and clients consumes massive communication resources. Meanwhile, due to different computing, storage capabilities and network environments between clients, some clients may be offline, causing a loss of model parameters, which will affect the efficiency, accuracy and fairness of FL.

In order to solve the problem of non-IID data in federated learning, Yue Zhao et al. [10] created global shared data in the central server to improve the training accuracy of non-IID data, but this sharing of data essentially violates the principle of data privacy protection in FL. Jiang [11] first built a global model in a collaborative way, and then used the private data to personalize the global model for each client. Muhammad et al. [12] combined federated learning with a recommendation system and proposed an algorithm that uses a k-means method to cluster the similarity of different nodes, then randomly selects a certain number of nodes in different clusters to participate in training.

It is necessary to consider reducing the communication overhead of FL. In this regard, Chen [13] proposed a layered asynchronous update algorithm. The author layered the parameters into shallow parameters and deep parameters according to the structural characteristics of the deep neural network model. In the early global communication iteration process, only the shallow parameters are transmitted between the server and local clients, and the deep parameters of the global model are transmitted and aggregated in the last few rounds of communication. This algorithm reduces the communication overhead by reducing the size of the transmission model parameters and the update frequency of the deep parameters in the neural network. The disadvantage of this algorithm is that the accuracy of the model would be affected. Zhu [14] introduced the sparse evolution algorithm (SET) [15] into federated learning. By controlling the sparsity parameter between the fully connected layers of the neural network, the SET algorithm is able to control the connection sparsity between fully connected networks. In this way, the parameter size of the transmission model is reduced, and the communication cost is effectively reduced, but it may affect the global model accuracy.

Another challenge of federated learning is structural heterogeneity. Due to different computing, storage capabilities and asymmetric network environments between clients, some clients will possibly be offline and lose model parameters during transmission. In order to enhance the robustness of federated learning, scholars have conducted various studies on structural heterogeneity. Hao et al. [16] designed a secure aggregation protocol that allows clients to withdraw at any time, as long as the number of remaining clients can meet the FL update, which improves the fault tolerance and robustness of the system. Other scholars have studied how to rationally allocate heterogeneous equipment resources. Kang et al. [17] considered the differences in the costs of clients to encourage more high-quality clients to carry out FL training. Li et al. [18] used the variance of the global model accuracy distribution as a fairness measure and designed a q-FFL optimization algorithm, which increased the model aggregation weight of high-loss clients. Experiments show that the algorithm can improve the accuracy of low-accuracy participation, and the global model accuracy distribution between clients is more balanced, promoting fair resource allocation of federated learning.

The above-mentioned studies have been carried out on a certain aspect of communication cost or structural heterogeneity, but few studies address comprehensive considerations

on these issues. However, the application of FL often requires model accuracy, fairness and communication cost at the same time. To achieve the balance of multiple objectives and parameter optimization under the FL framework, some scholars have tried to combine intelligent optimization algorithms with federated learning. Zhu et al. [14] defined FL as a bi-objective optimization model with the goal of minimizing model error rate and communication cost, and used the NSGA-II algorithm to optimize the neural network structure parameters of FL. The Pareto-optimal solution evolved by the algorithm improves model performance and communication efficiency to a certain extent compared with the standard FedAvg algorithm. However, the algorithm does not consider the instability of the communication environment caused by the structural heterogeneity of federated learning or the imbalance of accuracy distribution among clients. Basheer et al. [19] used the particle swarm algorithm to update the number of hidden layers, the number of neurons and the global communication rounds of the neural network, but its optimization goal is a single goal without comprehensive consideration of the other goals of federated learning.

In response to the above problem, aiming at realizing multi-objective equilibrium and hyperparameter optimization of FL on non-IID data, this paper proposes a framework for optimizing the structure of neural network models in FL. Therefore, we first define FL as a three-objective optimization model, which aims to simultaneously minimize the global model error rate, the global model accuracy distribution variance and communication cost, and takes the learning rate, batch size and neural network structure parameters as decision variables. Before using an evolutionary algorithm to optimize FL, a hierarchical clustering algorithm is introduced to divide the clients into different clusters, and the clusters are proportionally sampled for the FL evaluation process of the evolutionary algorithm. Then, based on the characteristics of FL, this paper proposes an improved NSGA-III algorithm, namely NSGA-III-FD, with improves NSGA-III with Fast greedy initialization and Discarding strategy of abandoning low-quality individuals of the population in the late iterations. The experimental results show that the proposed NSGA-III-FD algorithm can achieve the balance of three objectives, improve the training efficiency and obtain an appropriate parameter for FL training, as it can effectively reduce the communication cost and the variance of the accuracy distribution while maintaining the overall performance of the FL model without serious loss. The contributions of this paper are as follows:

1. In the case of non-IID data, which can be regarded as an asymmetric data distribution, we construct a multi-objective FL optimization model and comprehensively consider the three minimization objectives of global model error rate, global model accuracy distribution variance and communication cost, which aims to achieve FL parameter optimization and realize the balance of model accuracy, fairness and communication cost.
2. Using a hierarchical clustering algorithm on non-IID data can accelerate convergence and improve the accuracy of the global model so that FL can carry out NSGA-III with a low FL client participation ratio without a serious loss of accuracy. It can reduce communication cost and improve the efficiency of the evolutionary algorithm, which would be more feasible in a practical application.
3. We propose the NSGA-III-FD algorithm. In order to quickly converge and obtain high-quality Pareto solutions, a fast greedy initialization for multi-objective FL and the strategy of discarding low-quality individuals in the late iterations are proposed to speed up NSGA-III evolutionary efficiency and to improve the applicability of individual solutions to the population.
4. Through CNN experiments on MNIST and CIFAR-10 datasets, it is verified that using a hierarchical clustering algorithm can accelerate convergence and improve FL accuracy on non-IID data. The Pareto solutions obtained by the proposed NSGA-III-FD algorithm are better than that of the NSGA-III algorithm and other classical evolutionary algorithms, such as MOEAD, NSGA-II and SPEA2. The results show that the Pareto solutions obtained by the NSGA-III-FD algorithm are of higher quality. Moreover, some Pareto solutions are selected for FL experiments. The optimized

neural network model can effectively reduce the communication cost and the variance of the global model accuracy while maintaining the accuracy of the federated learning global model.

The remainder of this paper is organized as follows: The Section 2 proposes a multi-objective optimization model of federated learning and discusses the proposed NSGA-III-FD algorithm in detail, the Section 3 presents a comparative analysis and summary of the experiments, finally, the Section 4 makes a conclusion.

2. Proposed Algorithm

2.1. Preliminaries

2.1.1. Federated Learning

Federated learning is a distributed privacy protection machine learning technology that allows clients to jointly train a global model without uploading local private data to the server. Suppose there are K clients with local datasets $\{D_1, D_2, \cdots, D_K\}$; the traditional centralized learning puts all the data together as $D = D_1 \cup D_2 \cdots D_K$ and uses D to train the model. In the training process of federated learning, federated learning aims to minimize the global loss function in a distributed scheme, that is, to minimize the weighted average of the local client loss function. The loss function of the k-th client with the local dataset D_k is shown below:

$$L_k(w) = \frac{1}{n_k} \sum_{i \in D_k} l_i(w) \tag{1}$$

where k is the serial number of the clients, w is the global model parameters, $L_k(w)$ is the loss function of the k-th client, $l_i(w)$ is the loss function of data sample i in local dataset D_k, and n_k is equal to the local dataset's size as $n_k = |D_k|$.

In addition, the global goal of federated learning is to minimize the global loss function $L(w)$:

$$\min_w L(w) = \min_w \frac{\sum_{i \in \cup_k D_k} l_i(w)}{|\cup_k D_k|} = \min_w \sum_{k=1}^{K} \frac{n_k}{n} L_k(w) \tag{2}$$

where n is the total data sample size of K clients. The goal of federated learning is to optimize the global loss function $L(w)$ by minimizing the weighted average of the client loss function $L_k(w)$. Federated learning is a collaborative process, as shown in Figure 1.

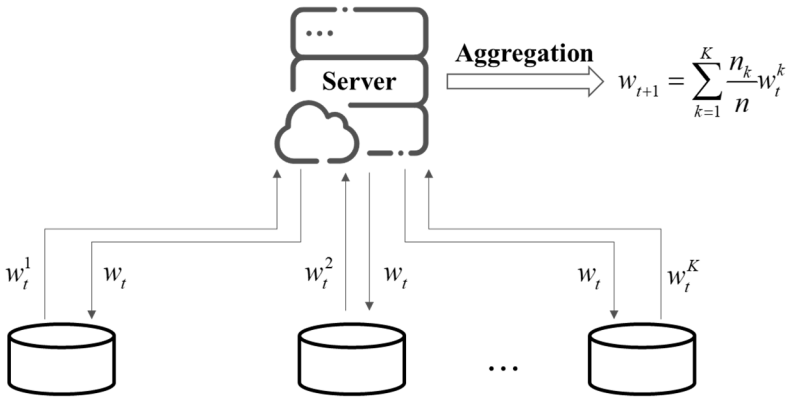

Figure 1. The training process of federated learning.

At t round of training, each client receives a global model w_t from the server and trains model w_t with local data. After training, the k-th client can obtain an updated local model w_t^k and then upload it to the server. The server aggregates the models with certain rules to obtain a new global model w_{t+1} for the next round of iterative training.

2.1.2. NSGA-III Algorithm

There are many studies on multi-objective optimization algorithms based on genetic algorithms and Pareto-optimal solutions, such as the second-generation non-dominated ranking genetic algorithm [20] (NSGA-II), the decomposition-based multi-objective evolutionary algorithm [21] (MOEA based on decomposition, MOEA/D), SPEA2 [22] and PAES [23] (Pareto archived evolution strategy). NSGA-II is a powerful and robust multi-objective evolutionary algorithm for problems with two or three objectives. If the number of objectives is greater than three, newer evolutionary algorithms can be used, such as the third generation of the reference-point-based non-dominated ranking genetic algorithm [24] (NSGA-III); NSGA-III outperforms NSGA-II for optimization problems with four or more objectives. In this paper, FL is defined as a three-objective optimization problem model. To ensure scalability, when the objectives of FL are extended to four and more, the NSGA-III algorithm is adopted in this paper.

2.2. Multi-Objective Federated Learning Optimization Model

We construct a three-objective optimization model of FL, and explain its objectives, decision variables and coding of decision variables. The three-objective optimization model of FL can be summarized as follows:

$$\begin{cases} \min F(v) = \min(f_1(v), f_2(v), f_3(v))^T \\ \quad s.t. \\ \quad \eta \in [1, Max_\eta], B \in [1, Max_B] \\ Conv \in [1, Max_{conv}], kc \in [1, Max_{kc}], ks \in (3 or 5) \\ L \in [1, Max_L], Fc \in [1, Max_{Fc}] \end{cases} \tag{3}$$

where $F(v)$ is the objective function of the model, v represents decision variables and $v = \{\eta, B, Conv, kc, ks, L, Fc\}$; constraints are the value range of each variable. The details of the objective function and decision variables are defined as follows:

1. Objective function

The model has three objective functions, including minimizing the global model test error f_1, the global model accuracy distribution variance f_2 and the communication cost f_3. The three minimization functions comprehensively consider the balance between the accuracy, fairness and communication cost of the FL model.

Accuracy is an important goal of the FL model. Since the objective function is a minimization function, the goal of maximizing accuracy is transformed into the goal of minimizing the global model test error rate. Traditional FL tends to lean toward some clients, resulting in a large accuracy gap between clients, especially on non-IID data. Therefore, the introduction of a fairness objective into the FL optimization model could lead to a more balanced distribution of accuracy among clients. In this paper, the fairness goal is represented by the global model accuracy distribution variance used in [18]. The third objective function is to minimize communication cost, as low cost means the feasibility and sustainability of FL. In FL, the communication cost is directly related to the model parameters transmitted by clients. Based on human experience, it is difficult to find a model with low model complexity and high accuracy. Therefore, the three objective functions form a large space, and there is no exact correlation among the three functions.

In this paper, the specific evaluation process of the three objectives of individuals in the evolutionary algorithm uses the clustering-based FedAvg algorithm. After FL training, we could test the trained global model w to obtain the accuracy of each client $\{a_1, a_2, \cdots, a_K\}$. The average test accuracy of the global model is calculated $A = \frac{\sum_{k=1}^{K} a_k}{K}$, from which the target global model test error rate can be calculated as $f_1 = 1 - A$.

The goal f_2 is the variance of the global model accuracy distribution $f_2 = \frac{\sum_{k=1}^{K} (a_k - A)^2}{K}$, which is also obtained based on the accuracy distribution of each client. Variance can be considered as one of the measures of fairness.

The goal f_3 is defined as the communication cost of the individual in the population. Generally, the communication cost of each client is only related to the size of the model parameters it transmits. A complex model means a higher communication cost. In this paper, the communication cost of the FL training with an evolutionary algorithm is related to the size of the model parameters, the proportion of clients participating in the training and the communication rounds. However, only the size of the model parameters is different among the individuals of the evolutionary algorithm, so the target communication cost f_3 can ignore the other two factors and only be expressed by the size of the model parameters σ, that is, $f_3 = \sigma$.

2. Model decision variables

Decision variables are represented by v; since FL is a process of collaborative training with a machine learning model, the optimized parameters in this paper include the learning rate η, batch size B, and neural network structure parameters. Among them, the neural network structure parameters directly determine the model complexity of FL and affect the three objective functions of communication cost, accuracy and variance of FL.

The neural network chosen for this article is the convolutional neural network (CNN). CNN parameters include the number of convolutional layers $Conv$, the number of kernel channels kc, kernel size ks, the number of fully connected layers L and the number of neurons in the fully connected layer Fc. That is, $\boldsymbol{v} = \{\eta, B, Conv, kc, ks, L, Fc\}$; the value range of each variable is set in the experimental part.

3. Decision variables coding

We use the NSGA-III-FD algorithm to optimize the learning rate, batch size and neural network structure parameters of FL. Chromosomes are the main body of algorithm operation. There are two types of decision variables: integer and real. All integers use binary encoding, and real numbers use real-value encoding. The number of convolution layers, number of convolution kernels, size of the convolution kernel, number of fully connected layers, number of neurons per layer and batch size of CNN are binary encoding, and the learning rate uses real-value encoding. An example of CNN encoding is shown in Figure 2.

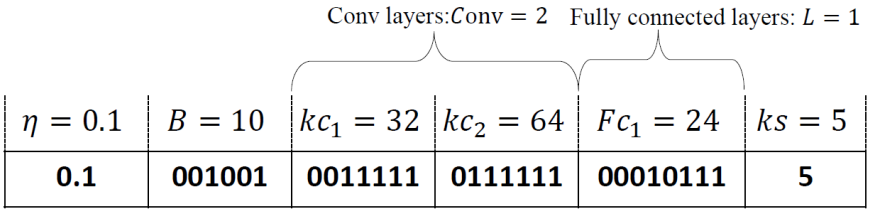

Figure 2. Coding example of CNN.

The decoding process is to automatically increase by 1 during binary decoding. For example, B = 001001 is decoded as 9, but the actual value is B = 10. For convenience, the size of the convolution kernel in CNN is only selected between 3 and 5, and the convolution output is always kept unchanged. In the neural network structure of CNN, only a pooling layer is added at the end of the convolution layer.

2.3. Modified NSGA-III-Based Multi-Objective FL Parameter Optimization Algorithm with Hierarchical Clustering

2.3.1. Federated Learning with Hierarchical Clustering

In this paper, FL is set with non-IID data. Before the evolutionary algorithm, the clients are clustered into clusters by hierarchical clustering. There are two parameters in a hierarchical clustering algorithm. The first is the distance measurement of cluster similarity, and the second is the link mechanism parameter. Since there may be outliers in the data setting of this paper, and both single linkage and complete linkage are easily affected by outliers, Euclidean distance and Ward's linkage are finally selected in this paper based

on experimental results in the literature [25]. Euclidean distance is a common measure to judge the similarity of vectors. Ward's linkage (which can only be combined with Euclidean distance measure) seeks to minimize the variance in the cluster when merging two clusters.

In this process, we refer to the literature [25] to conduct N rounds of FL global communication before clustering, and then conduct a round of global communication involving all clients to obtain the model parameters uploaded by all clients. The model parameters are transformed into vectors, and then the hierarchical clustering algorithm is used to iteratively merge the most similar clusters until the given distance threshold T to obtain the clustering results. The pseudo code is shown in Algorithm 1.

Algorithm 1: Hierarchical clustering-based FedAvg

Server:
Initialize w_0
For $t = 1, 2, \ldots, N+1$ do
  if $t == N+1$:
    for each $k \in K$ in parallel do
      $w_t^k \leftarrow$upload from $ClientUpdate(k, w_t)$
      $\vec{w^k} \leftarrow w^k$ to vector
    end for
    $C = \{c_1, c_2, \cdots c_M\} \leftarrow$HierarchicalClusteringAlgorithm $(\vec{w}, P_{HC})$
  else:
    Select $m = \max(\alpha \cdot K, 1)$ clients as S_t
    for each $k \in S_t$ in parallel do
      $w_t^k \leftarrow$upload from $ClientUpdate(k, w_t)$
    end for
    $w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_t^k$
End for
$ClientUpdate(k, w_t)$**:**
Download w_t as w^k
  For $e = 1, 2, \ldots, E_c$ do
for $b \in B$ do
    $w^k = w^k - \eta \nabla L_k\left(w^k, b\right)$
  end for
End for
Upload w^k to server

In Algorithm 1, N is pre-rounds before hierarchical clustering, K is the total number of clients, w_0 is the initial global model parameter, w_t^k is the parameter of the k-th transmission model at t round of training, $\vec{w^k}$ is the vector transformed from k-th model parameter, $\vec{w}$ is vectors of all clients, the client's fraction α is a random number between 0 and 1, and S_t is the selected m clients to participate at t round of FL training. P_{HC} is parameters of the hierarchical clustering algorithm, such as distance threshold T; $C = \{c_1, c_2, \cdots c_M\}$ is the clusters result, and M is the number of clusters. Within the $ClientUpdate(k, w_t)$ algorithm, E_c is the iteration round of the client's local training, B is the mini-batch size of the client's local training, and η is the learning rate of the mini-batch SGD.

2.3.2. Modified NSGA-III for Multi-Objective FL Parameter Optimization

In order to ensure the scalability of the algorithm objectives, this paper adopts the NSGA-III algorithm and names the improved NSGA-III as NSGA-III-FD.

The search space of decision variables in FL is large, and better initial solutions can accelerate the NSGA-III convergence speed and improve the quality of Pareto solutions. This paper improves the random initialization of NSGA-III to fast greedy initialization and appropriately discards the low-quality individuals in the population at the late iterations. For example, when the error rate is higher than a certain value, such as $f_1 > 85\%$, the

individual is deleted and the reserved solution of fast greedy initialization is used, so as to ensure the accuracy of individuals in final Pareto-optimal solutions.

The fast greedy initialization process is briefly described as follows: Firstly, randomly generate l-fold initial solutions, where l is the multiple of the initial solution. After randomly dividing all clients into groups of the same size based on clustering results, the FL evaluation process of the initial solutions is performed simultaneously within each group. In addition, each training parameter of FL is reduced for quick evaluation. Then, select the optimal population solution for each of the three objectives, respectively; After mixing and removing duplicate solutions, the solutions with the specified population number are randomly selected as P_0, and the remaining solutions are recorded as the initial reserved solutions RS_0.

The flow chart of the NSGA-III-FD algorithm based on hierarchical clustering results is shown in Figure 3.

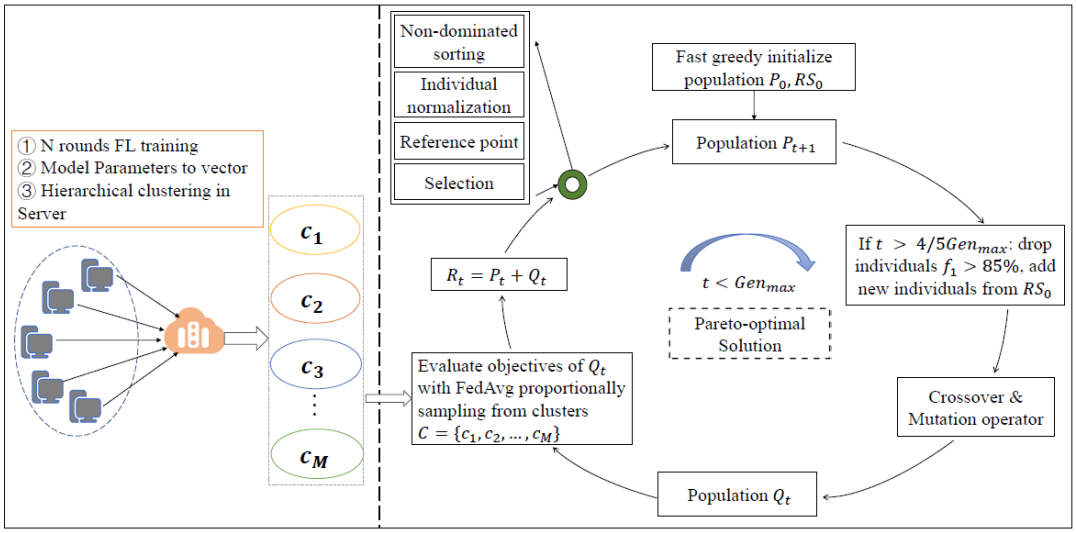

Figure 3. Flow chart of the NSGA-III-FD algorithm.

NSGA-III-FD first uses fast greedy initialization to generate the first-generation parent population and encodes the corresponding variables into binary and real-value chromosomes. The iterative process is mainly to select two parent individuals using a binary tournament to generate two offspring individuals. The crossover and mutation algorithm adopts single-point crossover and flip mutation on the binary chromosomes, simulated binary crossover (SBX) and polynomial mutation of the real-value chromosomes. This process is repeated until offspring population Q_t are produced.

Then, calculate the three objectives of the offspring population Q_t. Mix parent population and offspring population $R_t = P_t + Q_t$, and conduct non-dominated sorting of the mixed population R_t. Select the next generation P_{t+1}. Repeat these steps until the late iterations, for example, where $t \geq 4/5Gen_{\max}$ and $Gen_{\max}$ is the maximum number of generations, drop the low-quality individuals in the obtained population, such as the individuals with $f_1 > 85\%$, and randomly select new individuals to supplement the population from RS_0. Then, continue until meeting the maximum number Gen_{max}, and a set of Pareto-optimal solutions are obtained.

The specific evaluation process of the three objectives of an individual in the NSGA-III-FD algorithm is to conduct the FedAvg algorithm with clustering results $C = \{c_1, c_2, \cdots c_M\}$. The pseudo code is shown in Algorithm 2.

Algorithm 2: Clustering-based FedAvg evaluation algorithm of an individual

Input: Decision variable parameters corresponding to individual i, hierarchical clustering results $C = \{c_1, c_2, \cdots c_M\}$
Output: Three objectives value of individual i
Use parameters of individual i to initialize the neural network weight w_0^i
For $t = 1, 2, \ldots$ do
  for $c \in C$ do
    Select $m_c = \max(\alpha \cdot K_c, 1)$ clients
  end for
  $S_t = \cup m_c \leftarrow$ proportionally sampling from clusters C
  for each $k \in S_t$ in parallel do
    $w_t^k \leftarrow$ upload from $ClientUpdate(k, w_t)$
  end for
  $w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_t^k$
End for
Use the trained w^i to calculate the global model test accuracy A and model parameter size $\sigma^i = f\left(w^i\right)$
Calculate objective f_1^i global model test error $E = 1 - A$
Calculate objective f_2^i global model accuracy distribution variance V
Calculate objective f_3^i communication cost $Cost = \sigma^i$

In Algorithm 2, i is an individual of the population in the NSGA-III-FD algorithm, c is the index of clustering results C, and K_c is the set of clients in cluster c. After the i-th individual is decoded, the relevant parameter of FL is obtained. First, use the parameters to initialize the global model used in the clustering-based FedAvg algorithm. In each round of training, the model uses mini-batch SGD to train local data. After a certain number of rounds, three goals are calculated: the test error of the global model, the variance of the global model accuracy distribution and the communication cost.

3. Experiments

3.1. Experimental Setting

This section describes the experimental setting of this paper. It mainly includes the following parts:

1. Experimental environment and experimental dataset

The experimental environment was based on an Intel (R) core (TM) i9-9900KF CPU @ 3.60 ghz × 16 Ubuntu system. Training and testing on MNIST dataset [26] and CIFAR-10 dataset [27]: the MNIST dataset consisted of 28 × 28-pixel handwritten digital images, with 60,000 training images and 10,000 test images; the CIFAR-10 dataset consisted of 32 × 32-pixel images from 10 classes, and it had 50,000 training images and 10,000 test images.

2. Neural network model parameters

We chose the convolutional neural network (CNN) as the neural network model for FL training in this paper, and the standard CNN parameters were set empirically [8]. The CNN model was set with two 5 × 5 convolutional layers (the first with 32 channels and the second with 64 channels), followed by a 2 × 2 Max pool layer, a full connected layer of 128 neurons, and finally a 10 class SoftMax output layer (with a total of 1,659,146 parameters on the MNIST dataset, and 2,152,266 parameters on the CIFAR-10 dataset). The mini-batch SGD algorithm had a learning rate η of 0.05 and batch size B of 10.

3. Federated learning parameters setting

In FL, we set the total number of clients $K = 100$ and the client participation ratio $\alpha = 0.1$, that is, there were $100 \times 0.1 = 10$ clients in each round of communication. For client-local model training, the local iteration round was set to 3.

Since the data between clients are usually heterogeneous in size and distribution, we studied the following two non-IID data settings. The first was the extreme uneven data distribution (extreme non-IID). Each client had approximately only one label of data, but the data size of each node was the same, so the data skewness, which measures the asymmetry of data probability distribution, was high. The second was unbalanced non-IID. The number of labels and data size of each client were unbalanced, and only the upper and lower limits of client data size were set.

The above parameter settings were used as the standard FedAvg settings in the experiments of this paper.

4. NSGA-III-FD Parameters Setting

Next, we set the parameters for NSGA-III-FD with 20 iterations, where the population size was set to 20. The individual operators adopted the relevant setting in the literature [14]. The selection operator used a two-round tournament, and the crossover and mutation operators were empirically set to use a single-point crossover with probability 0.9 and a bit-flip mutation with probability 0.1 in the binary chromosome. In addition, a simulated binary crossover with probability 0.9 and $n_c = 2$ and a polynomial mutation with probability 0.1 and $n_m = 20$ were used in the real-value chromosomes.

3.2. Performance Analysis of Federated Learning with Hierarchical Clustering

The data distribution of FL was set as non-IID in this paper. In order to improve the accuracy on non-IID data and reduce communication cost, we carried out hierarchical clustering on FL to divide all clients into clusters before NSGA-III-FD, which made it possible to start the NSGA-III-FD algorithm with a low FL participation ratio without serious loss of accuracy.

We selected Euclidean distance and Ward's linkage parameters to use in the hierarchical clustering algorithm. Before clustering, we performed N global communication rounds of FL training, and after obtaining model parameter vectors of all clients, the hierarchical clustering algorithm iteratively merged the clients until the given distance threshold T to obtain the clustering results. The experiments were carried out on two non-IID data distributions on the MNIST and CIFAR-10 datasets. The hierarchical clustering parameters differed in different datasets and different non-IID settings. The selected parameters after a series of experiments are shown in Table 1.

Table 1. Related parameter settings of federated learning with hierarchical clustering.

	MNIST	**CIFAR-10**
Extreme non-IID	$N = 5, T = 6$	$N = 5, T = 5$
Unbalanced non-IID	$N = 1, T = 11$	$N = 1, T = 5$

Clients from the clusters were proportionally sampled to FL training, and the clustering-based FL (CFL) was compared with the traditional FL. In addition, the FL global communication round was set $E = 50$ and the client participation ratio was set $\alpha = 0.1$. In addition, it was reasonable to choose a low client participation ratio, such as $\alpha = 0.1$, to take FL training on non-IID data, especially on extreme non-IID data, as the literature [25] proves that varying the client fraction seems to have only a small effect on test set accuracy on non-IID data. The experimental results on the MNIST and CIFAR-10 datasets with the CNN model under two non-IID data settings are shown in Figure 4.

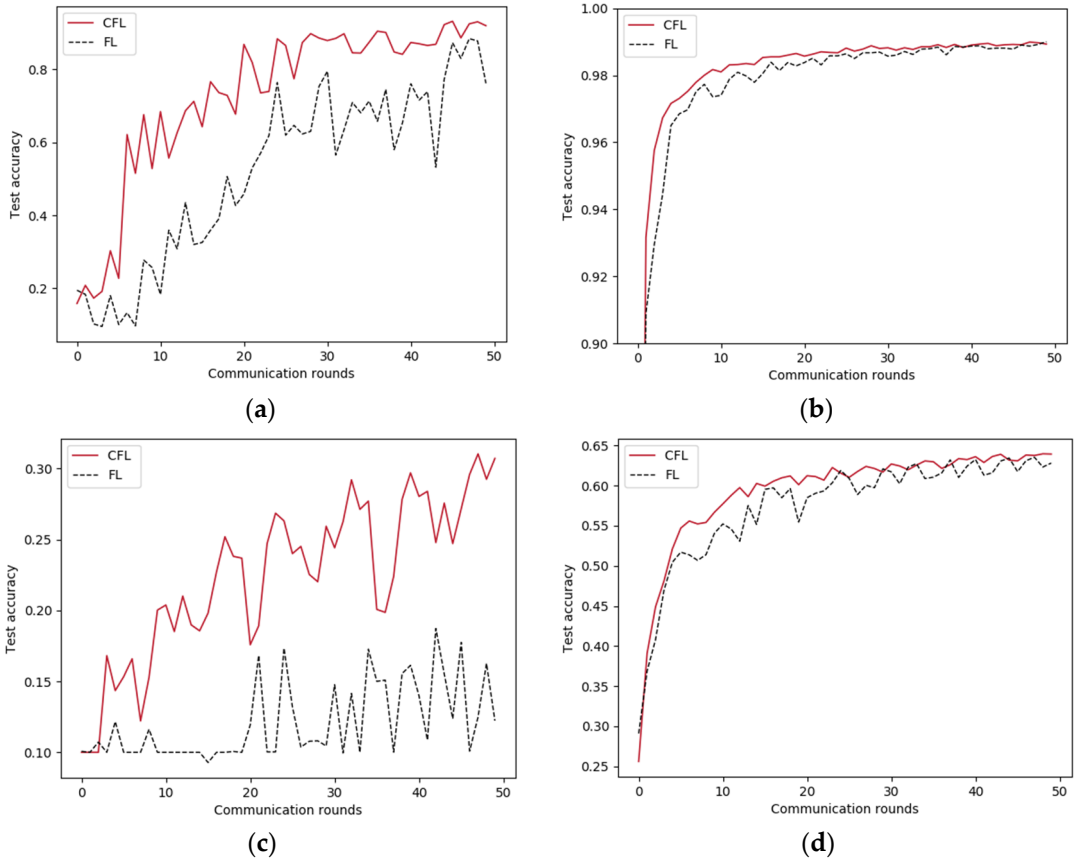

Figure 4. Iterative curve of federated learning with hierarchical clustering. (**a**) MNIST, extreme non-IID; (**b**) MNIST, unbalanced non-IID; (**c**) CIFAR-10, extreme non-IID; (**d**) CIFAR-10, unbalanced non-IID.

Overall, it is clearly seen that the global test accuracy of the CFL is higher than that of the FL in all situations, and CFL can speed up the convergence speed. Under extreme non-IID, CFL can significantly improve the accuracy. In the case of unbalanced non-IID, the accuracy of CFL is improved less, but it can at least accelerate the convergence speed under the condition of ensuring the accuracy.

Therefore, it is feasible to employ the multi-objective federated learning evolutionary algorithm with a low client participation ratio $\alpha = 0.1$ using clustering results, which can not only reduce the impact of non-IID data distribution, but also reduce the communication cost of the whole evolutionary algorithm.

3.3. Multi-Objective Federated Learning Evolutionary Algorithm Performance Analysis

3.3.1. Performance Analysis of NSGA-III-FD Algorithm

We employed NSGA-III-FD to optimize the parameters of FL, so as to achieve balance between the global model test error rate, global model accuracy distribution variance and communication cost. According to the above analysis, we conducted the FL evaluation process of NSGA-III-FD based on the hierarchical clustering results, and other compared evolutionary algorithms did the same, so as to verify the effectiveness of the proposed NSGA-III-FD.

Firstly, we compared the proposed NSGA-III-FD with NSGA-III to explore the effectiveness of the proposed fast greedy initialization method and the drop strategy in the NSGA-III-FD algorithm. A summary of the NSGA-III-FD algorithm experimental parameter settings is given in Table 2.

Table 2. Related parameter settings of the NSGA-III-FD algorithm.

Parameter	CNN
Population size	20
Generations	20
Learning rate	0.01–0.2
Batch size	1–64
Number of conv layers	1–3
Kernel channels	1–128
Kernel size	3 or 5
Number of fully connected layers	1–3
Number of Fc neurons	1–256

We set the population size to 20 and the generations to 20, and the number of communication rounds for each individual in the FL evaluation process was $E = 10$. We set the client participation ratio $\alpha = 0.1$ and the maximum batch size to 64. The range of the learning rate η was between 0.01 and 0.2, because a learning rate that is too large will harm the convergence of FL.

In the parameter settings of the CNN neural network, the maximum number of convolution layers was 3, the maximum number of kernel channels was 128, the maximum number of fully connected layers was 3, the maximum number of neurons in the fully connected layer was 256, and the size of the convolution kernel size was 3 or 5.

We used the NSGA-III-FD and NSGA-III algorithms to evolve the final Pareto solutions on the MNIST and CIFAR-10 datasets under extreme non-IID and unbalanced non-IID data. The result is presented in Figure 5, where each point represents a solution corresponding to a specific structural parameter in FL. The red points represent the Pareto-optimal solutions obtained by the proposed NSGA-III-FD algorithm, and the blue points represent the Pareto-optimal solutions obtained by the NSGA-III algorithm.

From Figure 5, we can see that the red solutions are basically better than the blue solutions, that is, the Pareto solutions obtained by the proposed NSGA-III-FD algorithm are better than the Pareto solutions of the NSGA-III algorithm. Moreover, it can be found that the red solutions of NSGA-III-FD converge more at the inflection point. The solution at the inflection point is characterized by small target values and higher solution quality. Moreover, due to the strategy of discarding low-quality individuals, the red solutions have fewer high error rate points. The blue solutions are more dispersed, and there are more and higher test error rate solutions.

In addition, Table 3 shows the relevant evaluation index results of the Pareto solutions finally obtained by NSGA-III-FD and NSGA-III, and we discuss the Pareto solutions of the two algorithms from multiple dimensions.

Hypervolume index [28] (HV), which calculates the sum of the hypervolume of the hypercube formed by all non-dominated solutions and reference points, is a comprehensive index for evaluating Pareto solutions. Generally speaking, the greater the HV value, the better the quality of the evaluated Pareto solution. It can be seen from Table 3 that the HV value of the NSGA-III-FD algorithm is always better than NSGA-III, showing better quality. In addition, by observing the hypervolume of generations, we can determine that the NSGA-III-FD algorithm can converge before the 20th generation.

The NSGA-III-FD algorithm obtains more Pareto non-dominated solutions and is more robust than the NSGA-III algorithm in terms of the number of solutions.

Coverage rate C (A, B) [29] calculates the proportion of solution set B that are dominated by at least one solution in A and measures the degree of overlap between the two solution sets. The larger the index C is, the better the quality of solution set A is versus that of solution set B. In Table 3, FD in C (FD, N) represents the solution set of NSGA-III-FD, and N represents the solution set of NSGA-III; C (FD, N) measures are basically greater than C (N, FD). In terms of coverage C, the solution of NSGA-III-FD is better than that of the NSGA-III algorithm.

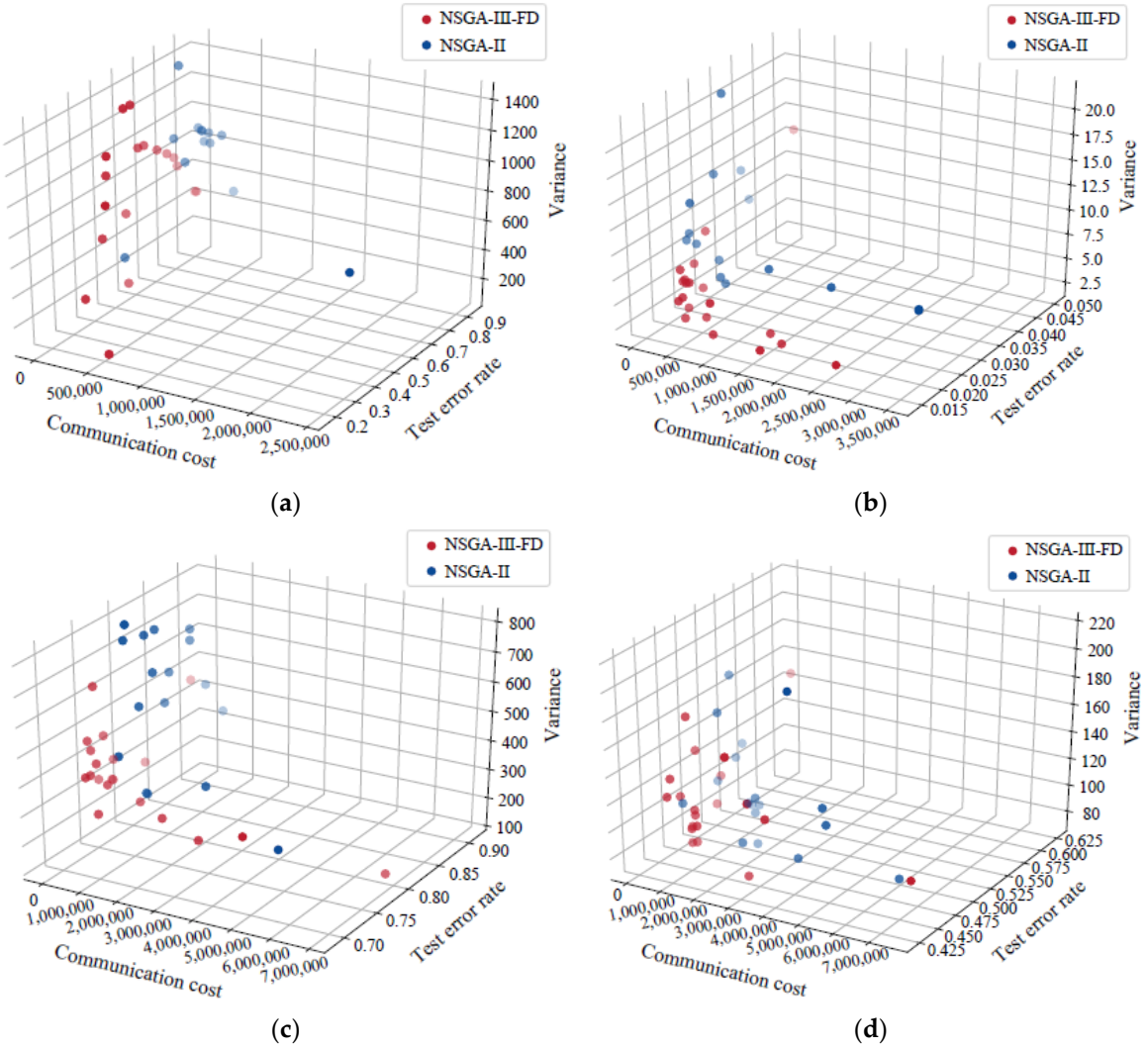

Figure 5. Pareto-optimal solutions distribution diagram of NSGA-III-FD and NSGA-III. (**a**) MNIST, extreme non-IID; (**b**) MNIST, unbalanced non-IID; (**c**) CIFAR-10, extreme non-IID; (**d**) CIFAR-10, unbalanced non-IID.

Table 3. Analysis of indicators of NSGA-III-FD algorithm and NSGA-III algorithm.

		MNIST		CIFAR-10	
		Extreme Non-IID	**Unbalanced Non-IID**	**Extreme Non-IID**	**Unbalanced Non-IID**
Hypervolume	NSGA-III-FD	**2,175,452,092**	**2,223,543**	**742,307,751**	**189,679,984**
	NSGA-III	1,042,304,455	1,496,982	573,310,061	168,263,685
Number of Pareto Solutions	NSGA-III-FD	**18**	**20**	**20**	**20**
	NSGA-III	13	15	17	18
Coverage, C/%	C (FD, N)	**100**	**80**	**76.47**	**66.67**
	C (N, FD)	0	5	0	0
Minimum Error Rate/%	NSGA-III-FD	**16.97**	**1.42**	**68.71**	**42.32**
	NSGA-III	35.67	1.86	74.08	43.3
Minimum Variance	NSGA-III-FD	**102.13**	**2.39**	**133.03**	74.91
	NSGA-III	498.51	6.47	243.63	**74.88**
Minimum Cost	NSGA-III-FD	**1971**	50,683	**3560**	88,633
	NSGA-III	49,977	**40,738**	22,259	**242,717**

The minimum value of a single objective reflects the extreme value of each objective function and reflects the optimization ability of the algorithm. It can be seen from Table that the minimum global model test error rate obtained by NSGA-III-FD is smaller than that of NSGA-III. In terms of minimum variance and minimum communication cost, there are a few cases where the minimum value of NSGA-III is better than that of NSGA-III-FD, but there is little difference between the two minimum values.

Based on the above analysis of the HV index, the number of non-dominated solutions, C index and the single objective minimum value, we can conclude that the Pareto-optimal solutions obtained by the proposed NSGA-III-FD algorithm are better-quality solutions than those obtained by the NSGA-III algorithm.

3.3.2. Comparison between NSGA-III-FD Algorithm and Classical Evolutionary Algorithms

In addition, we compared the proposed NSGA-III-FD with NSGA-II, MOEAD and SPEA2, from the index of HV, the number of Pareto solutions, coverage rate C, single-objective optimal value, etc. The experimental results are shown in Figure 6 and Table 4.

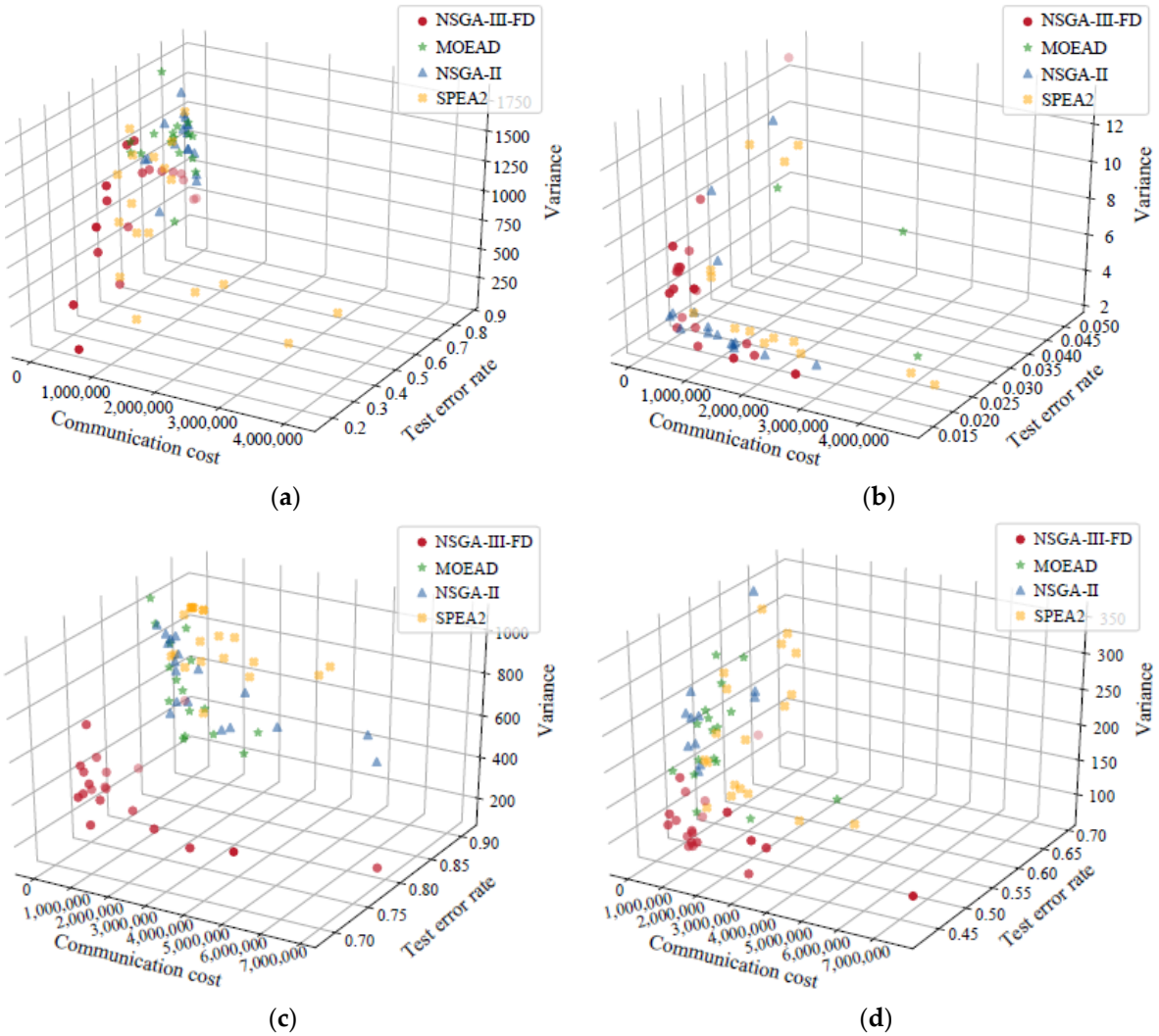

Figure 6. Pareto-optimal solutions of NSGA-III-FD, MOEAD, NSGA-II and SPEA2. (**a**) MNIST, extreme non-IID; (**b**) MNIST, unbalanced non-IID; (**c**) CIFAR-10, extreme non-IID; (**d**) CIFAR-10, unbalanced non-IID.

Table 4. Analysis of various indicators of the NSGA-III-FD algorithm and other evolutionary algorithms.

		HV	Number of Pareto Solutions	Coverage, C/%	Min Error Rate/%	Min Variance	Min Cost
MNIST, extreme non-IID	NSGA-III-FD	**4,548,688,737**	**18**		**16.97**	**102.13**	**1971**
	MOEAD	1,292,750,316	16	100/0	39.62	748.23	6806
	NSGA-II	1,085,050,947	17	82.35/0	60.49	601.4	4591
	SPEA2	3,103,427,927	18	100/0	26	320.81	68,688
MNIST, unbalanced non-IID	NSGA-III-FD	**1,789,828**	**20**		**1.42**	**2.38**	50,683
	MOEAD	357,216	3	100/0	1.89	4.41	995,112
	NSGA-II	1,705,197	16	62.5/45	1.61	2.62	**21,448**
	SPEA2	1,491,010	14	57.14/0	2	2.49	36,027
CIFAR-10, extreme non-IID	NSGA-III-FD	**1,100,617,191**	**20**		**68.71**	**133.03**	**3560**
	MOEAD	347,349,691	16	75/0	81.02	296.82	15,019
	NSGA-II	293,673,388	18	44.44/0	81.37	376.6	4451
	SPEA2	224,484,016	20	95/0	81.3	590.25	17,185
CIFAR-10, unbalanced non-IID	NSGA-III-FD	**520,407,254**	**20**		**42.32**	**74.91**	88,633
	MOEAD	417,464,425	17	41.18/0	44.19	100.69	43,715
	NSGA-II	293,334,515	11	72.73/0	48.21	153.95	**20,102**
	SPEA2	370,646,715	19	89.47/0	46.84	101.58	63,912

A simple analysis is shown in Table 4 and Figure 6. In Table 4, all indicators of the NSGA-III-FD algorithm are basically better than those of MOEAD, NSGA-II and SPEA2, except for the minimum communication cost. In Figure 6, the red solutions of the NSGA-III-FD algorithm are basically better than other algorithms. Except for MNIST, unbalanced non-IID, the results are slightly mixed with the solutions of other algorithms, but in terms of data indicators, the solution quality of the NSGA-III-FD algorithm is better.

In summary, the proposed NSGA-III-FD algorithm is better than the NSGA-III, MOEAD, NSGA-II and SPEA2 evolutionary algorithms, and the Pareto solution obtained by NSGA-III-FD has a higher quality.

3.3.3. NSGA-III-FD Pareto Solutions for FL Experiment

Since the communication rounds setting $E = 10$ in the FL evaluation process of NSGA-III-FD is very small, the FL accuracy performance has not been fully explored. Therefore, the solution obtained by the NSGA-III-FD algorithm was selected for an FL enhancement experiment to verify whether the algorithm realizes the multi-objective equilibrium and parameter optimization of FL.

For the Pareto-optimal solutions obtained by the NSGA-III-FD algorithm, two solutions were selected, one of which was the solution with the smallest global test error, and the other one was the inflection point solution. We performed CFL training on these two solutions and compared them with the Standard FL and CFL. The communication rounds were set to 100 rounds on MNIST, 500 rounds on CIFAR-10 (extreme non-IID) and 200 rounds on CIFAR-10 (unbalanced non-IID), and the client participation ratio was set to $\alpha = 0.1$. All results are listed in Table 5, and the global test accuracy is also listed in Figure 7.

From the results shown in Figure 7 and Table 5, we can observe the evolution of the selected solutions as follows. It can be seen from Figure 7 that, in general, CFL high of NSGA-III-FD works best, especially on CIFAR-10, extreme non-IID. We verified in the first experiment that CFL is superior to the standard federated learning algorithm.

Table 5. Experiment data of solutions obtained by NSGA-III-FD.

Parameter	Solution	η	B	Conv Layers	Kernel Size	Fc Layers	Test Accuracy/%	Variance	Cost
MNIST, extreme non-IID	CFL high	0.0167	60	[4, 90]	5	[82]	**97.04**	**10.25**	1,456,586
	CFL random	0.064	8	[9]	5	[85]	94.68	19.76	**151,119**
	Standard CFL	0.05	10	[32, 64]	5	[128]	95.64	15.52	1,659,146
	Standard FL	0.05	10	[32, 64]	5	[128]	95.64	20.52	1,659,146
MNIST, unbal-anced non-IID	CFL high	0.0163	8	[19, 40, 99]	5	[78, 221]	**99.33**	**0.93**	1,651,902
	CFL random	0.0158	13	[18, 40, 99]	5	[16, 214]	99.13	1.31	**433,875**
	Standard CFL	0.05	10	[32, 64]	5	[128]	98.97	1.14	1,659,146
	Standard FL	0.05	10	[32, 64]	5	[128]	98.82	1.71	1,659,146
CIFAR-10, extreme non-IID	CFL high	0.0472	40	[20]	5	[169]	**55.14**	224.02	**86,869**
	CFL random	0.0247	39	[49]	5	[165]	52.82	**143.39**	2,075,309
	Standard CFL	0.05	10	[32, 64]	5	[128]	51.68	262.72	2,152,266
	Standard FL	0.05	10	[32, 64]	5	[128]	34.04	981.65	2,152,266
CIFAR-10, unbal-anced non-IID	CFL high	0.011	10	[13, 125]	3	[55, 210]	**65.87**	**88.92**	1,789,039
	CFL random	0.013	20	[85, 107, 124]	3	[15, 241]	65.58	137.74	**662,371**
	Standard CFL	0.05	10	[32, 64]	5	[128]	64.66	100.79	2,152,266
	Standard FL	0.05	10	[32, 64]	5	[128]	64.31	106.62	2,152,266

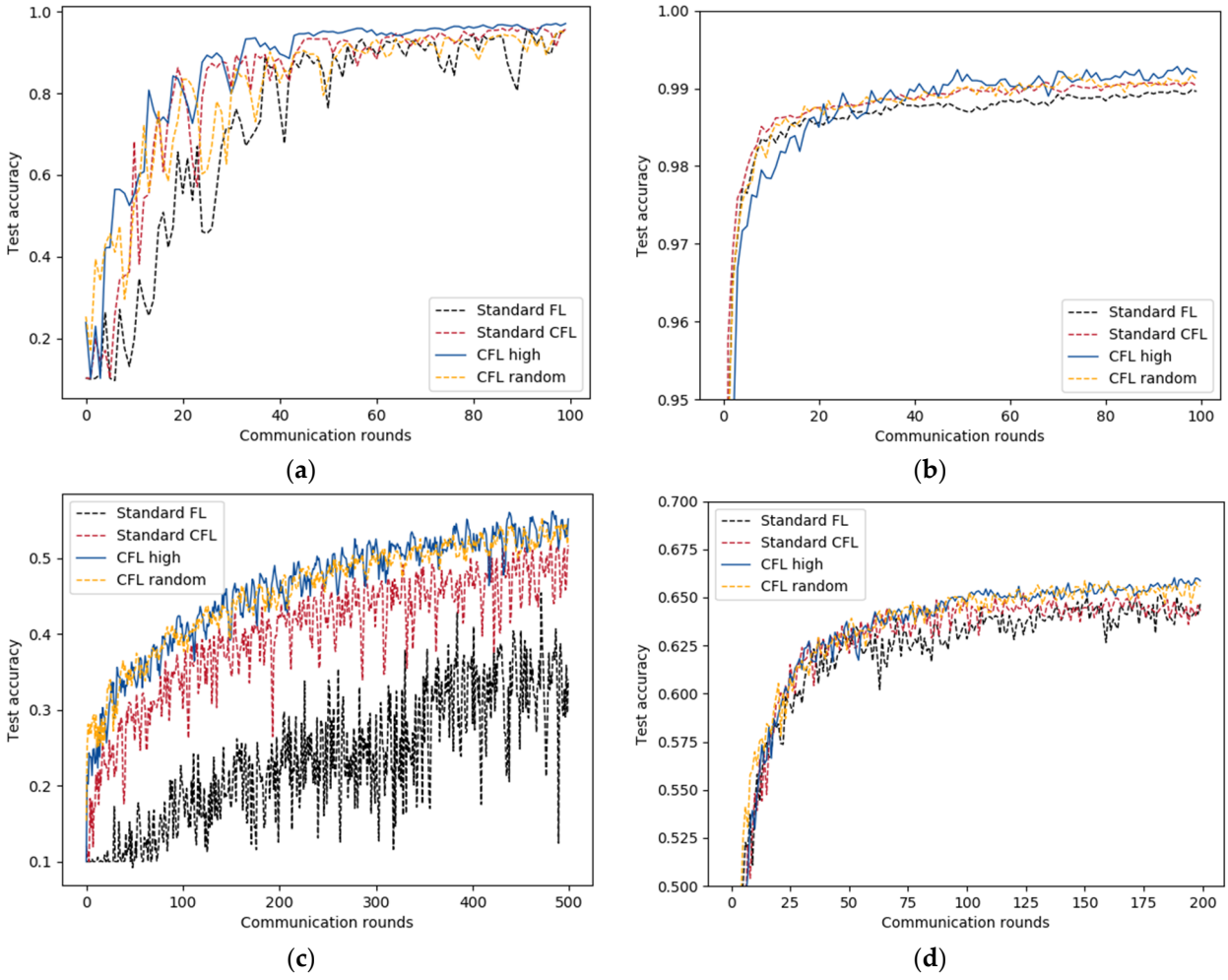

Figure 7. Iterative curve of solutions obtained by the NSGA-III-FD algorithm. (**a**) MNIST, extreme non-IID; (**b**) MNIST, unbalanced non-IID; (**c**) CIFAR-10, extreme non-IID; (**d**) CIFAR-10, unbalanced non-IID.

The solutions of the NSGA-III-FD algorithm in the figure may be superior to the standard FL and CFL for two reasons: one is the function of the federated learning algorithm based on clustering results, and the other is the function of the optimized model of the NSGA-III-FD algorithm. Through specific analysis, we can find that the lowest error rate solution from CFL high and the randomly selected inflection point solution from CFL random are basically above CFL, except for the CFL random on CIFAR-10, extreme non-IID, which is obviously inferior to CFL. It can be determined that the NSGA-III-FD evolutionary algorithm is effective for model optimization in terms of accuracy. The inflection point solution is inferior to CFL, perhaps because the positioning of the inflection point solution sacrifices accuracy in exchange for low communication cost.

From the data in Table 5, the communication cost of the CFL high solution of NSGA-III-FD is higher than the inflection solution (CFL random) and slightly lower than the standard FL, but the accuracy is higher. The inflection solution shows that the communication cost is significantly reduced, but the accuracy will be partially lower than the CFL. This reflects the multi-objective balance of federated learning. When the complexity of the model is reduced, it may be at the cost of accuracy or fairness. The target of the NSGA-III-FD evolutionary algorithm is to find the solution with a low cost and, as much as possible, to find the parameters solution with a low communication cost, high accuracy, high fairness and stable iteration. Through analysis, the NSGA-III-FD algorithm can basically realize multi-objective equalization of federated learning and parameter optimization.

4. Discussion

We have conducted hierarchical clustering experiments to verify the feasibility of federated learning with a low client participation ratio and compared the proposed NSGA-III-FD with NSGA-III and other classical evolutionary algorithms to verify the effectiveness of the proposed algorithm. The optimized model was tested to verify the effectiveness of the whole algorithm in multi-objective equilibrium and parameter optimization. Then, the practical implications and shortcomings of this work are discussed below.

4.1. Practical Implications

It can be considered from three aspects. The first is that the NSGA-III algorithm can ensure scalability when the objectives of FL are extended to four and more. Second, this paper carried out experiments on two non-IID data settings, as there are more non-IID data in real life. Thirdly, compared with the method referred to in the literature [14,19], which uses all clients to train, our method can reduce the cost of the whole evolutionary algorithm, as it can carry out the NSGA-III algorithm with a low client participation ratio while ensuring the accuracy of federated learning and training. Therefore, the algorithm in this paper has practical significance with the characteristics of low cost and high scalability.

4.2. Limitations

First, when the optimized network expands from a lower to a higher dimension, additional operations may be required, such as adding a PCA algorithm before clustering.

Second, there is room for further optimization of the communication costs. In general, the method proposed in this paper has certain advantages in communication cost. However, when data are expanded or a high-dimensional CNN neural network is used, the communication cost will still be very large. Therefore, it is necessary to consider how to reduce the communication cost in the process of federated learning training and to further enhance the scalability of the algorithm.

5. Conclusions and Future Work

This paper studies multi-objective equilibrium and parameter optimization of federated learning under a non-IID data setting. We first constructed a three-objective optimization model of federated learning, and the optimization objective was to minimize the distribution variance of the global model test error rate, communication cost and global

model test error rate. The decision variable was the parameter of the federated learning neural network (CNN in this paper). By introducing a hierarchical clustering algorithm to FL to solve the non-IID data skew problem, the FL accuracy was improved and made it feasible to carry out an FL evaluation process of an evolutionary algorithm with a low client participation ratio. Then, we improved NSGA-III with a fast greedy initialization and a strategy of discarding low-quality individuals to speed up the convergence and obtain high-quality solutions. The experimental results show that the improved NSGA-III-FD algorithm is better than the original NSGA-III algorithm and other classical evolutionary algorithms. Compared with the standard FL and clustering-based FL, the selected Pareto-optimal solutions optimized by the NSGA-III-FD algorithm can achieve balance between three objectives, which can effectively reduce the distribution variance of the accuracy of the global model and the communication cost without deteriorating the accuracy of the global model.

The multi-objective federated learning evolutionary algorithm implemented in this paper can achieve the balance between different objectives, reduce the overall communication cost of the evolutionary algorithm and improve operation efficiency. However, there is still a lot of work to be done in the future. It is necessary to further consider the problems of structural heterogeneity, such as intermittent availability of clients, communication loss, etc. Future work may focus on how to improve the fault tolerance and computational efficiency of multi-objective federated learning.

Author Contributions: Conceptualization, J.Z. and W.M.; methodology, Y.W. and W.M.; software, H.Z.; formal analysis, S.D.; data curation, H.Z. and S.D.; writing—original draft, J.Z. and Y.W.; visualization, J.Z.; funding acquisition, W.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Program of the National Natural Science Foundation of China, grant number 61871388.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MNIST and CIFAR-10 datasets used to support the findings of this study are available at https://www.tensorflow.org/ (accessed on 1 March 2020) and http://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz (accessed on 1 March 2020).

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

References

1. Ribeiro, M.; Grolinger, K.; Capretz, M. MLaaS: Machine Learning as a Service. In Proceedings of the IEEE International Conference on Machine Learning & Applications, Miami, FL, USA, 9–11 December 2015.
2. Waring, J.; Lindvall, C.; Umeton, R. Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [CrossRef] [PubMed]
3. Lopez, K.L.; Gagne, C.; Gardner, M.A. Demand-Side Management Using Deep Learning for Smart Charging of Electric Vehicles. *IEEE Trans. Smart Grid* **2018**, *10*, 2683–2691. [CrossRef]
4. Papernot, N.; Mcdaniel, P.; Sinha, A.; Wellman, M. Towards the Science of Security and Privacy in Machine Learning. *arXiv* **2016**, arXiv:1611.03814.
5. Mehmood, A.; Natgunanathan, I.; Xiang, Y.; Hua, G.; Guo, S. Protection of Big Data Privacy. *IEEE Access* **2016**, *4*, 1821–1834. [CrossRef]
6. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *Int. Data Priv. Law* **2017**, *7*, 76–99. [CrossRef]
7. Chik, W.B. The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform. *Comput. Law Secur. Rep.* **2013**, *29*, 554–575. [CrossRef]
8. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, PMLR, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
9. Li, L.; Fan, Y.; Tse, M.; Lin, K.-Y. A review of applications in federated learning. *Comput. Ind. Eng.* **2020**, *149*, 106854. [CrossRef]

10. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated Learning with Non-IID Data. *arXiv* **2018**, arXiv:1806.00582. [CrossRef]
11. Jiang, Y.; Konecný, J.; Rush, K.; Kannan, S. Improving Federated Learning Personalization via Model Agnostic Meta Learning. *arXiv* **2019**, arXiv:1909.12488.
12. Muhammad, K.; Wang, Q.; O'Reilly-Morgan, D.; Tragos, E.Z.; Smyth, B.; Hurley, N.J.; Geraci, J.; Lawlor, A. FedFast: Going beyond Average for Faster Training of Federated Recommender Systems. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 6–10 July 2020; Association for Computing Machinery: New York, NY, USA, 2020.
13. Chen, Y.; Sun, X.; Jin, Y. Communication-Efficient Federated Deep Learning With Layerwise Asynchronous Model Update and Temporally Weighted Aggregation. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 4229–4238. [CrossRef] [PubMed]
14. Zhu, H.; Jin, Y. Multi-objective evolutionary federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1310–1322. [CrossRef] [PubMed]
15. Constantin Mocanu, D.; Mocanu, E.; Stone, P.; Nguyen, P.H.; Gibescu, M.; Liotta, A. Scalable Training of Artificial Neural Networks with Adaptive Sparse Connectivity inspired by Network Science. *arXiv* **2017**, arXiv:1707.04780.
16. 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.* **2020**, *16*, 6532–6542. [CrossRef]
17. Kang, J.; Xiong, Z.; Niyato, D.T.; Yu, H.; Liang, Y.-C.; Kim, D.I. Incentive Design for Efficient Federated Learning in Mobile Networks: A Contract Theory Approach. In Proceedings of the 2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS), Singapore, 28–30 August 2019; pp. 1–5.
18. Li, T.; Sanjabi, M.; Smith, V. Fair Resource Allocation in Federated Learning. *arXiv* **2020**, arXiv:1905.10497.
19. Qolomany, B.; Ahmad, K.; Al-Fuqaha, A.; Qadir, J. Particle Swarm Optimized Federated Learning For Industrial IoT and Smart City Services. In Proceedings of the GLOBECOM 2020–2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; pp. 1–6.
20. Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
21. Zhang, Q.; Li, H. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. Evol. Comput.* **2007**, *11*, 712–731. [CrossRef]
22. Zitzler, E.; Laumanns, M.; Thiele, L. *SPEA2: Improving the Strength Pareto Evolutionary Algorithm*; ETH: Zurich, Switzerland, 2001.
23. Knowles, J.D.; Corne, D.W. The Pareto archived evolution strategy: A new baseline algorithm for Pareto multiobjective optimisation. In Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406), Washington, DC, USA, 6–9 July 1999; Volume 1, pp. 98–105.
24. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Trans. Evol. Comput.* **2014**, *18*, 577–601. [CrossRef]
25. Briggs, C.; Fan, Z.; András, P. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–9.
26. Seng, L.M.; Chiang BB, C.; Salam ZA, A.; Tan, G.Y.; Chai, H.T. MNIST Handwritten Digit Recognition with Different CNN Architectures. *J. Appl. Technol. Innov.* **2021**, *5*, 7.
27. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, University of Tront, Toronto, ON, Canada, 2009.
28. Zitzler, E.; Thiele, L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **1999**, *3*, 257–271. [CrossRef]
29. Zitzler, E.; Thiele, L. Multiobjective Optimization Using Evolutionary Algorithms–A Comparative Case Study. In Proceedings of the International Conference on Parallel Problem Solving from Nature, Amsterdam, The Netherlands, 27–30 September 1998; Springer: Berlin/Heidelberg, Germany, 1998.