AgriEngineering

MDPI

Article

PCE-FL: A Personalized, Clustered, and Communication-Efficient Federated Learning Framework for Robust Tomato Leaf Disease Detection

Pradeep Gupta [1], Sonam Gupta [1], Lipika Goel [2], Abhay Kumar Agarwal [3], Arjun Singh [4], Vijay Shankar Sharma [4,*], Chiranji Lal Chowdhary [5,*] and Ruchita Chowdhary [6]

1 Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad 201015, India; gupta.pradeep85@gmail.com (P.G.); guptasonam6@gmail.com (S.G.)
2 Department of AIML, Sridevi Women's Engineering College, Hyderabad 500019, India; lipika.bose@gmail.com
3 Department of Computer Science and Engineering, Kamla Nehru Institute of Technology, Sultanpur 228118, India; abhaykumaragarwal@knit.ac.in
4 Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur 303007, India; arjun.singh@jaipur.manipal.edu
5 School of Computer Science Engineering & Information Systems, Vellore Institute of Technology, Vellore 632014, India
6 School of Computer Science & Engineering, Vellore Institute of Technology, Vellore 632014, India; ruchita.chowdhary2022@vitstudent.ac.in
* Correspondence: vijayshankar.sharma@jaipur.manipal.edu (V.S.S.); chiranji.lal@vit.ac.in (C.L.C.)

Abstract

Tomato leaf diseases represent a persistent threat to global food security, causing annual crop losses of 20% to 40%. Although deep learning models achieve accuracies exceeding 95% in centralized settings, their deployment across distributed farms is constrained by data privacy concerns, communication bottlenecks, and heterogeneous data quality. This paper proposes Personalized, Clustered, and Communication-Efficient Federated Learning (PCE-FL), a framework that integrates three synergistic components: (1) server-side client clustering to group farms with similar data distributions for personalized model training; (2) federated knowledge distillation to reduce communication overhead by over 91%; and (3) reputation-based aggregation to ensure robustness against unreliable contributions. Extensive experiments on realistic non-IID simulations of the PlantVillage tomato dataset Dirichlet($\alpha \in \{1.0, 0.5, 0.1\}$) demonstrate that PCE-FL achieves 89.1% accuracy under extreme heterogeneity ($\alpha = 0.1$), surpassing FedAvg by 10.9 and IFCA by 4.8 percentage points, while maintaining a 91% reduction in communication cost. All improvements are statistically significant ($p < 0.001$). These results advance the practical deployment of privacy-preserving collaborative AI in resource-constrained agricultural environments.

Keywords: federated learning; plant disease detection; tomato leaf classification; knowledge distillation; clustered aggregation; precision agriculture

Academic Editor: Chrysanthos Maraveas

Received: 8 February 2026
Revised: 19 March 2026
Accepted: 24 March 2026
Published: 6 May 2026

1. Introduction

Tomato (*Solanum lycopersicum*) is a vegetable crop of significant economic importance, cultivated globally across diverse climatic conditions and geographical regions. Nevertheless, there are constant threats to tomato production due to various fungal, bacterial, and viral pathogens. The Food and Agriculture Organization (FAO) reports that plant diseases result in annual losses of up to 20% to 40% in this crop worldwide [1], translating to

billions of dollars in economic losses and significant threats to food security in developing regions. Foliar symptoms should be identified at the earliest stage so that they can be subjected to timely management measures, which include fungicide treatments, removal of affected plants and environmental manipulation, in order to mitigate the spread of the disease [2].

Historically, the state of agriculture, especially the detection of diseases, has been examined through manual inspection by trained agricultural experts—a process that is labor-intensive, lengthy, expensive, and subject to errors [3]. The combination of computer vision and deep learning (DL) provides a transformative option towards automating this task. The efficacy of deep learning models, especially convolutional neural networks (CNNs) and sophisticated architectures like ResNet, Vision Transformers, and EfficientNet, has proven to be impressive in detecting plant diseases based on images of leaves [4]. These models have often been used to solve centralized training problems, where large aggregated datasets are used, and the classification accuracy is 95–99% on benchmark datasets, which is significantly higher than that of traditional machine learning algorithms using hand-crafted features [5].

Although the efficacy of centralized DL models is proven, data privacy and ownership can be seen as a fundamental limitation to the implementation of these models in a multi-farm environment. Farming businesses and individual farms are starting to consider their working data, such as crop health records, yield statistics and farm management, as sensitive competitive assets [6]. The necessity to communicate raw data to the central server is incompatible with the interests of the stakeholders and even results in the continually existing data silos that impede the creation of generalized models.

One of the most important paradigms to overcome this challenge has been federated learning (FL). FL allows several clients (farms) to jointly train one model without sharing confidential information. The algorithm has a cyclical nature: (1) a central server sends the existing global model to selected clients; (2) each client trains the model on its own data locally; and (3) clients only send new model parameters to the server to be aggregated [7]. This mechanism maintains data locality and privacy and utilizes a variety of distributed data resources [8].

Nonetheless, the simple extension of conventional FL algorithms like FedAvg to agriculture indicates that there are a variety of intertwined issues that can cause a major drop in performance and feasibility. First, statistical heterogeneityoccurs due to the fact that information in geographically distributed farms is necessarily non-independent and non-identically distributed (non-IID) [9], which is due to the differences in regional disease prevalence, types of crops cultivated, farming habits, and the environment. As an example, a humid-area farm may mostly face Late Blight and Leaf Mold, whereas an arid-climate-area farm will mostly face Bacterial Spot. This label skew makes one global model inefficient across all participants. Second, feature space skew caused by differences in imaging devices (model of camera, resolution, lighting, and complexity of the background) makes local models work towards opposite goals. These divergent updates, accumulated together, introduce client drift, reducing the convergence rate and decreasing model quality—the previous literature has found accuracy losses of up to 29 percentage points in extreme non-IID settings [10]. Third, communication constraints can occur in rural and remote agricultural settings, where connectivity can be unreliable, low-bandwidth and costly, and repeated transmission of large numbers of model parameters (tens to hundreds of megabytes for contemporary architectures) is both technically impossible and economically prohibitive [11]. Fourth, data quality and robustness concerns arise because participants might enter bad-quality updates owing to poor images, mislabeled samples, and irrelevant data that can distort the collaborative training process [12]. Such standard aggregation

schemes as FedAvg are especially susceptible, since they provide averaging client updates with no evaluation of contribution reliability.

Most importantly, these issues are not separate but closely related: a client with a rare yet legitimate disease profile may be wrongly identified as malicious due to poor robustness. and a communication error can be mistaken for client dropouts. Trying to solve any one of these challenges separately only tends to magnify others.

In order to overcome these two joint challenges, we introduce Personalized, Clustered, and Communication-Efficient Federated Learning (PCE-FL), an all-encompassing framework that is based on three synergistic pillars: (1) **Clustered Federated Learning**, a solution that mitigates statistical heterogeneity by grouping farms with similar data distributions and training customized, cluster-specific models; (2) **Federated Knowledge Distillation**, a solution that overcomes communication limitations by sending compact knowledge representation (logits) instead of full model parameters; and (3) **Reputation-Based Aggregation**, which ensures robustness by dynamically assessing contribution quality and weighting client updates accordingly.

The key contributions of this work include:

- **Novel Integrated Framework:** To the best of our knowledge, PCE-FL is the first framework to integrate client clustering, knowledge distillation, and reputation-based aggregation into a single integrated framework specifically designed to apply to agricultural FL applications.
- **Comprehensive Empirical Evaluation:** Extensive experiments on realistic non-IID simulations of the PlantVillage tomato dataset across three Dirichlet heterogeneity levels ($\alpha \in \{1.0, 0.5, 0.1\}$) demonstrate superior accuracy and convergence over five state-of-the-art baselines, with all improvements being statistically significant ($p < 0.001$).
- **Rigorous Ablation Study:** A systematic component-removal analysis provides quantitative evidence for the necessity and complementary contribution of each core component.
- **Practical Advancement:** Significant refinements in the practical implementation of robust, high-performance, privacy-preserving collaborative AI, achieving a 91% reduction in communication overhead, are demonstrated in resource-constrained and heterogeneous agricultural settings.

The rest of the paper is structured in the following way. Section 2 goes over the related literature on deep learning to detect plant diseases, personalized and clustered FL, federated knowledge distillation, and reputation-based aggregation. The PCE-FL methodology is outlined in Section 3. Section 4 explains the experimental environment, such as description of datasets, non-IID simulation, model structure, and metrics of evaluation. Section 5 gives results and analysis, which are the overall performance comparison, analysis of communication costs, ablation study and per-class analysis. Strong points, weaknesses, and future directions are discussed in detail in Section 6. Section 7 concludes the paper.

2. Related Work

2.1. Deep Learning for Plant Disease Detection

The first methods used were based on visual features that had to be hand-crafted, like color histograms, texture measurements based on GLCM and shape features using classical classifiers. These methods were also sensitive to variability in the real world and needed heavy feature engineering [13], although they were useful in constrained settings. Deep learning, especially convolutional neural networks (CNNs), has made end-to-end learning of features from raw pixel data a possibility and significantly increased the performance of plant disease classification in visually complex scenes [14,15]. Attention-augmented residual networks further improved the ability to classify plant diseases and localize lesions

among spatially unrelated disease symptoms [16] and achieved very high performance with controlled datasets like PlantVillage [17]. More recent peer-reviewed articles have shown that the recognition of plant diseases is shifting away from accuracy-focused performance to more robust and practically viable models, based on field variability and practical deployment. Current innovations are lightweight edge deployment architectures [18], ensemble and multi-scale feature extraction strategies to achieve stronger field robustness [19], and hybrid CNNTransformer networks to integrate local texture modeling and global contextual reasoning [20]. Methods designed based on transformers have also shown promise in precision agriculture, such as the Vision-Transformer-based detection of disease in leaf images as well as in-the-wild classification systems that explicitly rely on the domain variation between training datasets and actual agricultural systems [21,22]. Simultaneously, most recent attempts have emphasized the worthiness of lightweight and communication-conscious collaborative learning in distributed agricultural environments, where privacy, edge restriction, and heterogeneous local environments should be dealt with collectively [23]. However, the majority of the current solutions presuppose central access to training data, limiting them to privacy-sensitive, multi-stakeholder agricultural settings. These observations drive the desire to have federated structures that equally optimize recognition, robustness and communication efficiency.

2.2. Addressing Heterogeneity: Personalized and Clustered FL

The basic drawback of FedAvg is the inability to generalize across heterogeneous clients because of the non-IID data distribution with a single universal model [24]. This shortcoming has prompted the development of Personalized Federated Learning (PFL), which is less restrictive on the global model and is focused on providing client-adaptive models based on local data features. Clustered Federated Learning (CFL) is a strategy that has proved to be principled and effective among PFL strategies [25]. CFL instead partitions clients into groups with similar data distributions and federates separate models for each partition. By building statistically homogeneous training groups, CFL leads to better accuracy, personalization, and training stability and less model divergence. Clustering strategies include server-side clustering based on model update similarity, client-side self-selection and metadata-driven clustering using non-sensitive auxiliary information [26]. Recent extensions such as CASA address asynchronous participation of clients [27], while gradient-based partitioning methods further improve computational efficiency and scalability [28]. More recently, adaptive clustering strategies that dynamically adjust cluster boundaries based on evolving data distributions have shown improved performance in non-stationary environments [29], and personalized FL methods combining local fine-tuning with clustered aggregation have demonstrated benefits for domain-specific applications [30].

2.3. Optimizing Communication: Federated Knowledge Distillation

The communication overhead associated with transmitting large deep learning models is a fundamental obstacle to deploying federated learning in bandwidth-constrained agricultural environments [7]. A recent approach to this is federated knowledge distillation (FKD), in which instead of model parameters, compact representations of the knowledge are exchanged [31]. Building upon the teacher–student paradigm of knowledge distillation [32], FKD allows clients to create lightweight logits using a shared proxy dataset and send these distilled representations to the server, which greatly reduces the communication costs and retains model performance when the data distributions are non-IID [33]. Recent advances like FedGen and FedGKD further remove dependence on public datasets by using data-free knowledge distillation, which provides improved privacy preservation and is applicable to real-world farm settings [34,35]. In accordance with these developments,

the proposed PCE-FL framework combines FKD with a reputation awareness aggregation mechanism that ensures that only high-quality and reliable distilled knowledge is added to the global or cluster-specific model, which will improve the robustness of tomato leaf disease detection.

2.4. Ensuring Robustness: Reputation-Based Aggregation

The open and decentralized nature of federated learning makes it inherently vulnerable to low-quality or malicious client updates, which is generally modeled as an adversary with Byzantine behavior [36]. Existing defenses are broadly of two types. Byzantine-robust aggregation rules are designed to filter out anomalous updates before they are aggregated, and some notable examples are Krum and Multi-Krum, which pick updates that are closest to their neighbors in parameter space [37], coordinate-wise median and trimmed-mean aggregation schemes. While effective against certain poisoning attacks, these methods are computationally expensive and may inadvertently discard benign updates from highly non-IID clients.

Reputation-based mechanisms offer a more nuanced alternative by maintaining long-term trust scores for each client based on historical contribution quality [38,39]. Rather than binary inclusion or exclusion, reputation-weighted aggregation assigns greater influence to reliable clients while suppressing low-quality or suspicious updates. Representative approaches such as RFFL quantify contribution quality using cosine similarity between local gradients and the aggregated global direction to update reputation scores [40]. Recent hybrid methods further integrate reputation modeling with geometric anomaly detection to improve robustness against both malicious attackers and unintentional data quality issues [41]. In this work, PCE-FL incorporates reputation-based aggregation to preserve cluster model integrity in the presence of heterogeneous and unreliable agricultural data sources.

A comprehensive comparison of the state-of-the-art methods regarding their applicability to agricultural non-IID data, communication efficiency, and Byzantine robustness is provided in Table 1.

Table 1. Comparative analysis of the proposed PCE-FL framework against state-of-the-art centralized and decentralized learning methods.

Method	Application Domain	Capability (N/P/C/R/A)	Key Limitations
Centralized DL [15]	Plant disease detection	×/ × / × / × /✓	Requires raw data sharing; privacy violation; impractical across farms
FedAvg [7]	Generic FL	×/ × / × / × /×	Severe performance degradation under non-IID data; high communication cost
IFCA [28]	Personalized FL	✓/✓/ × / × /×	Ignores communication cost and robustness; cluster instability
CFL [25]	Clustered FL	✓/✓/ × / × /×	No robustness mechanism; high model transmission overhead
CASA [27]	Asynchronous CFL	✓/✓/ × / × /×	Addresses asynchrony only; lacks robustness and communication efficiency
FedMD [31]	FKD-based FL	×/ × /✓/ × /×	Assumes homogeneous, reliable clients; no personalization
FedGKD [35]	Data-free FKD	×/ × /✓/ × /×	No heterogeneity handling; no quality control
Krum/Median [37]	Byzantine-robust FL	×/ × / × /✓/×	Discards benign non-IID updates; computationally expensive
RFFL [39]	Reputation-based FL	×/ × / × /✓/×	No personalization or communication optimization
Hybrid Robust FL [40]	Robust FL	×/ × / × /✓/×	Focuses only on robustness
PCE-FL (Proposed)	Tomato leaf disease detection	✓/✓/✓/✓/✓	None (jointly addresses heterogeneity, communication, and robustness)

Legend: N = handles Non-IID data; P = clustered/personalized models; C = communication-efficient; R = robust to low-quality/Byzantine clients; A = agricultural context.

Literature Gap: As summarized in Table 1, existing studies typically address statistical heterogeneity, communication efficiency, or robustness in isolation. No prior work has jointly tackled all three challenges within a unified framework designed for agri-

cultural applications. In contrast, PCE-FL unifies clustered personalization, federated knowledge distillation, and reputation-aware Byzantine robustness within a single cohesive framework, offering a holistic solution tailored to the practical constraints of real-world agricultural deployments.

3. The PCE-FL Framework: Methodology

3.1. System Architecture and Overview

PCE-FL is intended to be a centrally coordinated and server-based system that coordinates collaborative learning between distributed clients (farms). The architecture in Figure 1 consists of a central server and N clients, where each client i has an independent local dataset of tomato leaf images that is never shared. The main functions of the server are:

- Reputation score management over a long period of time with each client.
- Knowledge aggregation and model training in each cluster that is communication-efficient.

There is a deviation in the workflow of regular FL. Instead of passing large model weights per round to the server, a bigger part of communication is with clients sending small (compact knowledge) vectors (logits) to the server and with the server returning revised and smaller student models. The whole procedure is one-on-one, effective in communication and strong. The proposed PCE-FL framework architecture is presented in Figure 1.

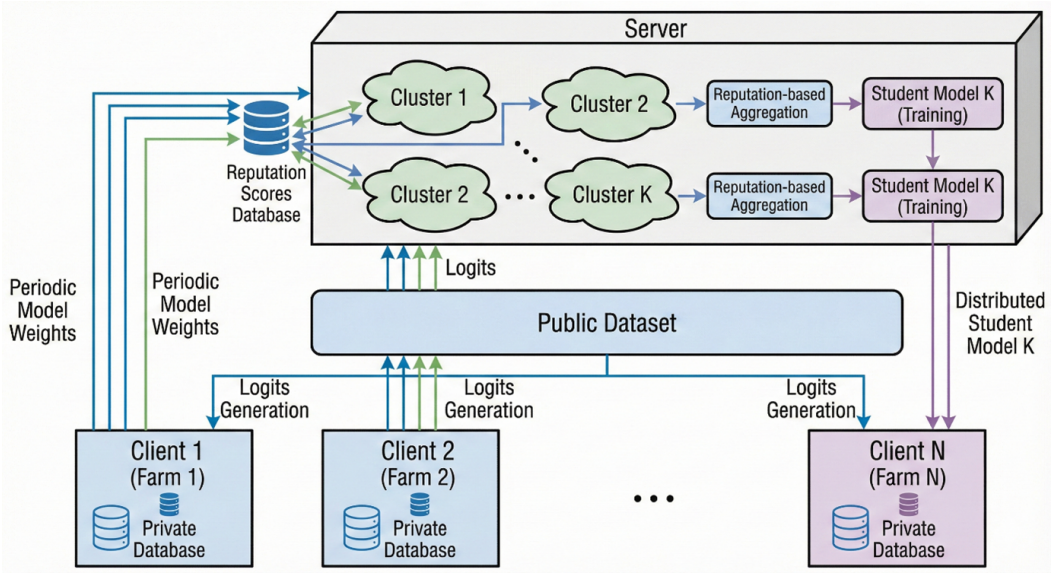

Figure 1. The architecture of the PCE-FL framework.

3.2. Dynamic Client Clustering for Personalization

To help deal with statistical heterogeneity, PCE-FL uses dynamic and server-side clustering to divide clients into groups with comparable data distributions and then train a specialized model that is more efficient than a single generic global model [42]. Clustering is performed every now and then (such as in T rounds of communication) to keep up with changes in client data distribution as time goes on. The procedure is as follows:

Initialization: It is possible to initialize the FL procedure (round $t = 0$) by randomization of the clients to a preliminary group of K clusters, or by performing a small number of rounds of usual FedAvg to get an initial global model of all the clients.

Update Submission for Clustering: Each client i at the beginning of a specific round number t of clustering trains its local model over a fixed number of epochs on its local data. It then delivers the complete updated parameter of the model $\mathbf{w}_i$ to the server. This complete weight shift is not a common procedure but is only done to re-cluster.

Similarity Calculation: The server gets model parameters of all participating clients and forms an $N \times N$ pairwise similarity matrix S. The flattened normalized weight vectors of model weights of any two clients i and j are used to compute the similarity between the two clients, which is the cosine similarity. This metric is based on the principle that models trained on statistically similar data will converge to nearby regions in high-dimensional weight space. The similarity is defined as in Equation (1):

$$\mathrm{Sim}(i,j) = \frac{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}}{\left\|\mathbf{w}_i^{(t)}\right\| \left\|\mathbf{w}_j^{(t)}\right\|} \tag{1}$$

where $\mathbf{w}_i^{(t)}$ and $\mathbf{w}_j^{(t)}$ denote the flattened model-parameter vectors of clients i and j at round t, respectively; $\cdot$ denotes the dot product; and $\|\cdot\|$ denotes the Euclidean (ℓ_2) norm.

Clustering Execution: With S computed, the server applies a standard clustering algorithm (e.g., hierarchical agglomerative clustering, HAC) to partition the N clients into K disjoint clusters $C_1, C_2, \ldots, C_K$. HAC is suitable as it does not require fixing K in advance and generates a dendrogram from which an appropriate number of clusters can be extracted. The output is a new cluster assignment for each client, which remains until the next re-clustering event.

3.3. Communication-Efficient Knowledge Transfer via Distillation

Once clusters are formed, PCE-FL switches to a highly communication-efficient protocol based on federated knowledge distillation for all subsequent intra-cluster training rounds [28].

Public Dataset Prerequisite: Assume that a small, unlabeled public dataset $\mathcal{D}_p$ of diverse tomato leaf images is available to both the server and clients (labels are not required).

Teacher–Student Paradigm: The large, powerful models on client devices act as "teachers," while the server maintains a smaller, efficient "student" model for each cluster.

Local Teacher Training: In each round, every client updates its local teacher model $\mathbf{w}_i$ via standard training on its private data.

Knowledge Generation: After local training, instead of sending its updated model, client i performs inference on D_p using its teacher model, extracting the pre-softmax output logit matrix $L_i \in \mathbb{R}^{|D_p| \times M}$, where M denotes the number of disease classes, and each row of L_i contains the unnormalized class scores for a given input image.

Knowledge Transmission: The client transmits only the compact logit matrix L_i to the server. The communication cost is proportional to $|\mathcal{D}_p| \times M$, which is orders of magnitude smaller than the entire model weights (often tens of MBs), especially for modern deep architectures [32].

3.4. Reputation-Aware Intra-Cluster Aggregation

Upon receiving knowledge logits from all clients in a given cluster, the server performs robust aggregation, safeguarded by a reputation mechanism, ensuring that aggregated knowledge is not tainted by low-quality contributions [43]. This executes independently for each cluster $\mathcal{C}_k$.

1. Initial Knowledge Aggregation: The server computes an initial average knowledge vector for cluster c by taking an element-wise mean of all client logits:

$$\mathcal{L}_{k,\mathrm{avg}} = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathcal{L}_i \tag{2}$$

which provides a baseline representation of cluster knowledge before applying any reputation-based refinement.

2. Contribution Quality Scoring: For each client $i \in \mathcal{C}_k$, the server assesses the contribution quality in the current communication round t by computing a quality score $q_i^{(t)}$. This score measures the similarity between the client's submitted logits $\mathcal{L}_i$ and the cluster's average logits $\mathcal{L}_{k,\text{avg}}$ using cosine similarity, as defined in Equation (3):

$$q_i^{(t)} = \text{cosine\_similarity}\left(\mathcal{L}_i,\, \mathcal{L}_{k,\text{avg}}\right) \tag{3}$$

This metric quantifies how well each client's contribution aligns with the cluster consensus, enabling differentiation between high-quality and low-quality submissions.

3. Reputation Update: The server maintains a long-term reputation score R_i for each client, reflecting its historical contribution quality. The reputation score is updated using an exponential moving average (EMA) scheme to ensure stability while assigning greater weight to recent behavior:

$$R_i^{(t)} = \alpha \cdot R_i^{(t-1)} + (1 - \alpha) \cdot q_i^{(t)} \tag{4}$$

In Equation (4), the initial reputation score is set to $R_i^{(0)} = 1.0$, and $\alpha \in (0, 1)$ denotes the momentum parameter controlling the update rate (a typical choice is $\alpha = 0.9$). This formulation ensures that reputation scores evolve gradually, preventing sudden fluctuations while allowing recent performance to influence final scores appropriately.

4. Reputation-Weighted Aggregation: The server leverages the reputation scores to compute a reputation-weighted average of the shared knowledge, as shown in Equation (5):

$$\mathcal{L}_{k,\text{avg}}^{\text{weighted}} = \frac{\sum_{i \in \mathcal{C}_k} R_i^{(t)} \cdot \mathcal{L}_i}{\sum_{i \in \mathcal{C}_k} R_i^{(t)}} \tag{5}$$

Clients with high reputation scores (close to 1.0, indicating consistently high-quality contributions) exert greater influence during aggregation. Conversely, clients with low reputation scores (approaching 0, reflecting unreliable or poor-quality updates) are automatically down-weighted, thereby preventing noisy or malicious contributions from corrupting the aggregated model. This reputation-aware aggregation mechanism enhances robustness against Byzantine or unreliable participants in the federated learning system.

5. Student Model Training: The server trains a compact student model $M_{k,\text{student}}$ for cluster C_k using the reputation-weighted aggregated logits $L_{k,\text{avg}}^{\text{weighted}}$ as soft supervision through knowledge distillation:

$$\mathcal{L}_{\text{KD}} = \text{KL}\left(\text{softmax}\left(\frac{M_{k,\text{student}}(D_{\text{pub}})}{T}\right) \middle\| \text{softmax}\left(\frac{L_{k,\text{avg}}^{\text{weighted}}}{T}\right)\right) \tag{6}$$

In Equation (6), $\mathcal{L}_{\text{KD}}$ denotes the knowledge-distillation loss, T is the temperature parameter controlling the softness of the probability distributions, and $\text{KL}(\cdot\|\cdot)$ represents the Kullback–Leibler divergence. This distillation objective enables efficient transfer of aggregated cluster knowledge into a computationally lightweight student model while preserving predictive performance.

6. Model Distribution: The updated student model $\mathcal{M}_{k,student}$ is sent back to all clients in cluster C_k, which use it as their new teacher model for the next training round. This federated learning cycle ensures continuous improvement while leveraging reputation-based quality control across distributed participants.

3.5. Complete Algorithm

The complete PCE-FL procedure is summarized in Algorithm 1, which presents the step-by-step workflow, including initialization, periodic re-clustering, intra-cluster knowledge distillation, reputation-based aggregation, and student model distribution. The architectural workflow is also depicted visually in Figure 1, which illustrates the data flow among clients, clusters, and the central server. Together, the algorithm and the architecture diagram provide complementary formal and visual representations of the PCE-FL framework.

Algorithm 1 Personalized, Clustered, and Communication-Efficient Federated Learning (PCE-FL)

Input: Number of clients N, number of clusters K, client datasets $\{\mathcal{D}_i\}_{i=1}^{N}$, public dataset $\mathcal{D}_{\text{pub}}$, total rounds T, re-clustering period T_{cluster}, local epochs E, momentum $\alpha \in (0,1)$.
Output: Cluster-wise student models $\{\mathcal{M}_{k,\text{student}}\}_{k=1}^{K}$.

Step 1: Initialization
Initialize client models $\{w_i^{(0)}\}_{i=1}^{N}$.
Initialize reputations $R_i^{(0)} \leftarrow 1.0$, $\forall i$.
Initialize clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ (random or heuristic).
Step 2: Federated training over rounds
for $t = 1$ to T **do**
  2.1 Re-clustering phase (every T_{cluster} rounds)
  if $t \bmod T_{\text{cluster}} = 0$ **then**
    for all clients $i = 1, \ldots, N$ **in parallel do**
      Train $w_i^{(t)}$ on $\mathcal{D}_i$ for E epochs.
      Send $w_i^{(t)}$ to the server.
    end for
    Compute similarity matrix $\mathbf{S}$ using cosine similarity over client models.
    Apply hierarchical agglomerative clustering (HAC) on $\mathbf{S}$ to update $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$.
  end if

  2.2 Knowledge transfer and aggregation within clusters
  for $k = 1$ to K **do**
    (a) Logit collection (teacher → server)
    for all clients $i \in \mathcal{C}_k$ **in parallel do**
      Train teacher model $w_i^{(t)}$ on $\mathcal{D}_i$ for E epochs.
      Compute logits $\mathcal{L}_i \leftarrow w_i^{(t)}(\mathcal{D}_{\text{pub}})$.
      Send logits $\mathcal{L}_i$ to the server.
    end for
    Compute preliminary cluster-average logits:
    $\mathcal{L}_{k,\text{avg}} \leftarrow \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathcal{L}_i$.
    (b) Reputation-based aggregation
    for all clients $i \in \mathcal{C}_k$ **do**
      $q_i^{(t)} \leftarrow \text{cosine\_similarity}(\mathcal{L}_i, \mathcal{L}_{k,\text{avg}})$.
      $R_i^{(t)} \leftarrow \alpha R_i^{(t-1)} + (1-\alpha)\, q_i^{(t)}$.
    end for
    Compute reputation-weighted cluster-average logits:
    $\mathcal{L}_{k,\text{avg}}^{\text{weighted}} \leftarrow \frac{\sum_{i \in \mathcal{C}_k} R_i^{(t)} \mathcal{L}_i}{\sum_{i \in \mathcal{C}_k} R_i^{(t)}}$.
    (c) Student model distillation (server → cluster)
    Train $\mathcal{M}_{k,\text{student}}$ on $\mathcal{D}_{\text{pub}}$ using $\mathcal{L}_{k,\text{avg}}^{\text{weighted}}$ (distillation loss).
    Send $\mathcal{M}_{k,\text{student}}$ to all clients $i \in \mathcal{C}_k$.
  end for (clusters)
end for (communication rounds)
End of Algorithm

4. Experimental Setup

4.1. Dataset and Non-IID Simulation

4.1.1. Dataset Description

The tomato subset of the PlantVillage benchmark dataset was used in all experiments. It comprises 10 tomato conditions—nine disease classes (Bacterial Spot, Early Blight, Late Blight, Leaf Mold, Septoria Leaf Spot, Two-Spotted Spider Mite, Target Spot, Tomato Mosaic Virus, and Tomato Yellow Leaf Curl Virus) and one Healthy class—and therefore defines a 10-class classification problem, as summarized in Table 2 and illustrated in Figure 2. The dataset is publicly accessible through the PlantVillage Kaggle repository, from which the tomato subset can be retrieved directly for reproducibility. The original images were collected under controlled conditions with visually uniform backgrounds and relatively consistent illumination, which is why PlantVillage is widely used as a laboratory-style benchmark for plant disease recognition. The source repository reports the controlled acquisition setting and image availability, but it does not provide complete device-level camera metadata; accordingly, we do not overstate such details here. In our pipeline, the images were processed at 224×224 pixels to match the input requirements of the pre-trained backbone models. During training, standard data augmentation was applied, including random horizontal flips, rotations up to 15°, and mild color jitter, in order to improve generalization. It should also be noted that PlantVillage does not distinguish among tomato cultivars; therefore, the present study evaluates disease recognition at the class level rather than cultivar-specific diagnosis. This limitation is acknowledged again in the Discussion, where cultivar-aware disease modeling is identified as a direction for future work (see Section 6).

Table 2. PlantVillage tomato subset class distribution.

Disease/Class	Label (ID)	Image Count
Bacterial Spot	A	1702
Early Blight	B	1920
Late Blight	C	1851
Leaf Mold	D	1882
Septoria Leaf Spot	E	1745
Spider Mite (Two-Spotted)	F	1741
Target Spot	G	1827
Tomato Mosaic Virus	H	1790
Tomato Yellow Leaf Curl Virus	I	1961
Healthy	J	1926
	Grand Total	18,345

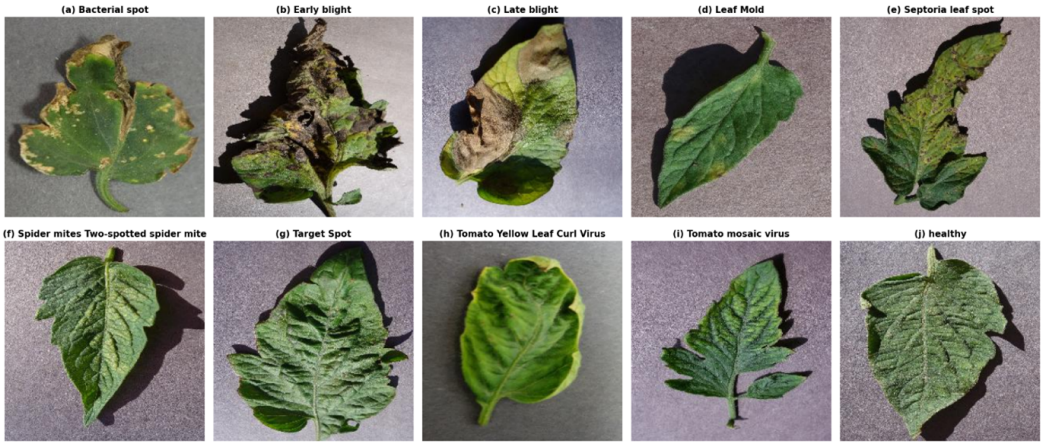

Figure 2. Sample images from the PlantVillage tomato leaf disease dataset.

4.1.2. Non-IID Partitioning

To simulate real-world heterogeneity, we partition the dataset across clients using a Dirichlet distribution. Specifically, we employ a Dirichlet(α) allocation over class labels to distribute the 10 classes to 50 clients with varying degrees of skew. We consider three heterogeneity levels: $\alpha = 1.0$ (mild heterogeneity), $\alpha = 0.5$ (moderate heterogeneity), and $\alpha = 0.1$ (severe heterogeneity). A smaller α yields a more skewed label distribution per client (each client sees fewer classes, mimicking specialization), whereas a larger α approaches an i.i.d. split. Table 3 classifies the heterogeneity levels in terms of α.

Table 3. Classification of data heterogeneity levels based on the Dirichlet parameter α (10 classes, 50 clients).

Heterogeneity Level	Dirichlet α	Characteristics
Moderate	1.0	Moderately skewed; diverse client distributions
High	0.5	Significantly skewed; clients biased to few classes
Extreme	0.1	Highly specialized; clients hold 1–2 classes only

Under extreme heterogeneity ($\alpha = 0.1$), many clients possess primarily one class (e.g., a particular disease) and very few samples of others, closely mirroring the variability observed across different farms. This challenging scenario allows us to rigorously evaluate PCE-FL's ability to handle statistical skew.

4.2. Model Architecture and Baselines

4.2.1. Base Model

All frameworks employed the EfficientNet-B0 architecture, which offers a state-of-the-art balance between accuracy and computational efficiency for this task. EfficientNet-B0 models were pre-trained on ImageNet and fine-tuned on the tomato disease classification task. For knowledge distillation in PCE-FL and FedMD, EfficientNet-B0 served as the client-side teacher model, while a compact custom CNN (4 convolutional layers, 2 fully connected layers; 500 K parameters versus EfficientNet-B0's 4 M parameters) was used as the server-side student model to emphasize communication efficiency.

4.2.2. Baseline Algorithms

PCE-FL was compared against five established federated learning algorithms to demonstrate its effectiveness:

- FedAvg [7]: Canonical algorithm using global weight averaging as the fundamental baseline.
- FedProx [24]: Addresses client drift in non-IID settings via a proximal term in the local objective.
- IFCA [28]: Iterative Federated Clustering Algorithm using weight averaging within clusters, isolating PCE-FL's distillation and reputation benefits.
- FedMD [31]: Federated knowledge distillation for a single global model without clustering, highlighting PCE-FL's personalization advantages.
- Krum [37]: Byzantine-robust method with geometric outlier detection, validating PCE-FL's reputation mechanism.

This comparison systematically isolates PCE-FL's contributions in clustered, reputation-weighted distillation for heterogeneous plant disease detection.

4.3. Implementation Details and Evaluation Metrics

4.3.1. Implementation Details

All experiments were implemented in Python using the PyTorch 2.10.0 deep learning library and the Flower federated learning framework, providing a robust simulation engine for FL research. The performance of the PCE-FL framework was evaluated using the standardized hyperparameter configurations listed in Table 4.

Table 4. Summary of global and local hyperparameters used during the training and evaluation of the PCE-FL framework and baseline algorithms.

Hyperparameter	Symbol	Value	Description
Client learning rate	η	0.01	SGD optimizer learning rate
Local training epochs per round	E	5	Client-side iterations per round
Batch size	–	32	Training batch size
Number of clusters	K	5	PCE-FL clustering parameter
Re-clustering period	T_{cluster}	50 rounds	HAC re-clustering frequency
Reputation momentum	α	0.9	Reputation smoothing factor
Distillation temperature	T	3	Knowledge distillation softness
Public dataset size	D_{pub}	500	Diverse tomato leaf images
Total communication rounds	T	200	Full federated training rounds
Total clients	N	100	Simulated participating clients

4.3.2. Evaluation Metrics

The performance of all federated learning algorithms was evaluated using both predictive and system-level metrics:

- **Top-1 Accuracy:** The proportion of test samples whose predicted class matches the ground-truth class. This metric provides the overall multiclass classification accuracy on the balanced global test set.
- **Macro-Averaged F1-Score:** The unweighted mean of class-specific F1-scores, computed as Macro-F1 $= \frac{1}{M}\sum_{c=1}^{M} F1_c$, where M denotes the number of classes. Because each class contributes equally, Macro F1 is particularly informative under class imbalance and heterogeneous performance across disease categories [44,45].
- **Total Communication Cost:** The cumulative volume of information transmitted during training, measured in megabytes (MB). For weight-sharing baselines, this includes model-parameter exchange; for distillation-based methods, it includes logit exchange [7].
- **Convergence Speed:** The number of communication rounds required to reach predefined performance thresholds (80%, 85%, and 90% test accuracy), which provides an operational measure of training efficiency [7].
- **Per-Class F1-Score:** The F1-score computed separately for each disease class, enabling disease-specific analysis of rare-class behavior and the contribution of personalization/clustering mechanisms [45].

5. Results and Analysis

5.1. Overall Performance Comparison

The main comparative results are presented in Table 5, which summarizes the test accuracy, Macro F1-score, and total communication cost of PCE-FL and the baseline federated learning methods under three levels of non-IID heterogeneity. To improve statistical

transparency and better assess result consistency, the accuracy and Macro-F1 values are reported as mean ± standard deviation over five independent runs conducted under the same experimental configuration. This presentation provides a direct comparison of classification effectiveness, robustness across repeated trials, and communication efficiency. *The compared baselines are FedAvg [7], FedProx [24] , IFCA [28], FedMD [31], and Krum [37]. Test accuracy and Macro F1-score are reported as mean ± standard deviation over five independent runs.*

Table 5. Main performance comparison of PCE-FL and baseline federated learning methods under varying non-IID heterogeneity levels.

Data Heterogeneity (α)	Algorithm	Test Accuracy (%, Mean ± std)	Macro F1-Score (%, Mean ± std)	Total Communication Cost (MB)
Moderate ($\alpha = 1.0$)	FedAvg	89.3 ± 0.5	88.9 ± 0.6	1850
	FedProx	90.1 ± 0.4	89.5 ± 0.5	1850
	IFCA	91.5 ± 0.3	91.1 ± 0.4	1850
	FedMD	89.8 ± 0.5	89.2 ± 0.5	165
	Krum	88.7 ± 0.6	88.1 ± 0.6	1850
	PCE-FL	92.6 ± 0.3	92.3 ± 0.3	165
High ($\alpha = 0.5$)	FedAvg	84.7 ± 0.8	83.9 ± 0.8	1850
	FedProx	86.2 ± 0.7	85.4 ± 0.7	1850
	IFCA	88.9 ± 0.5	88.1 ± 0.5	1850
	FedMD	85.1 ± 0.7	84.3 ± 0.7	165
	Krum	84.5 ± 0.8	83.1 ± 0.9	1850
	PCE-FL	91.8 ± 0.4	91.5 ± 0.4	165
Extreme ($\alpha = 0.1$)	FedAvg	78.2 ± 1.1	77.1 ± 1.2	1850
	FedProx	81.5 ± 0.9	80.6 ± 0.9	1850
	IFCA	84.3 ± 0.7	83.5 ± 0.7	1850
	FedMD	79.0 ± 1.0	77.9 ± 1.1	165
	Krum	78.0 ± 1.1	77.1 ± 1.2	1850
	PCE-FL	89.1 ± 0.5	88.6 ± 0.5	165

The results clearly demonstrate the superiority of the PCE-FL framework in Figure 3. Across all heterogeneity settings, PCE-FL consistently achieves the best predictive performance while maintaining the lowest communication cost among the compared methods. As the degree of heterogeneity increases (i.e., as α decreases from 1.0 to 0.1), the performance of all methods declines, but the degradation is markedly smaller for PCE-FL. Under the most challenging setting, $\alpha = 0.1$, PCE-FL attains $89.1 \pm 0.5\%$ test accuracy and $88.6 \pm 0.5\%$ Macro F1, outperforming FedAvg, FedProx, and IFCA by 10.9, 7.6, and 4.8 percentage points in accuracy, respectively. These results indicate that the clustered and personalized design of PCE-FL is more effective in handling severe client-level distribution shift than conventional global or weakly personalized federated strategies.

In addition to its predictive advantage, PCE-FL preserves strong communication efficiency through knowledge distillation. Specifically, the total communication cost is reduced from 1850 MB for the weight-sharing baselines to 165 MB, corresponding to an approximately 91% reduction, while still surpassing all competing methods in classification performance. The relatively low standard deviations reported in Table 5 further indicate stable behavior across repeated runs, supporting the consistency of the framework under varying non-IID conditions. Overall, these findings show that PCE-FL offers a more favorable trade-off between recognition performance, robustness to heterogeneity, and communication efficiency than the baseline methods.

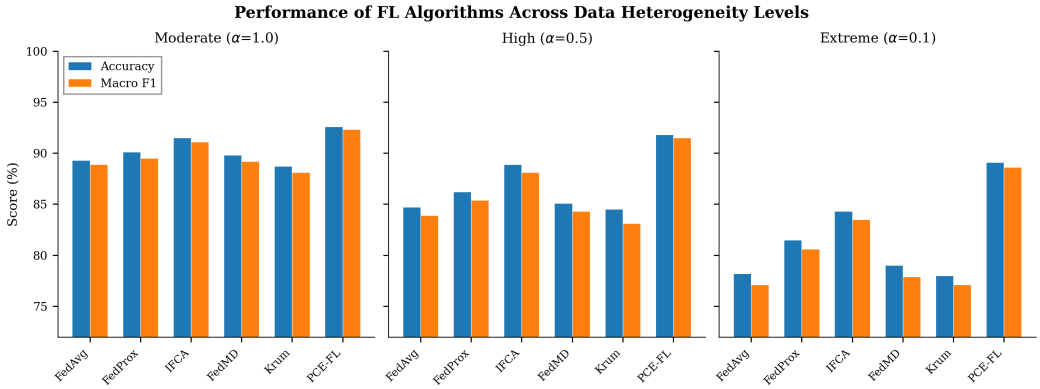

Figure 3. Grouped bar chart demonstrates PCE-FL's consistent superiority across all heterogeneity levels.

PCE-FL also reduces communication cost by 91% (165 MB vs. 1850 MB) through knowledge distillation while achieving substantially higher accuracy than FedMD. Furthermore, as shown in Figure 4, PCE-FL converges faster, reaching 85% accuracy in 58 rounds compared to 85 rounds for FedAvg. To assess the consistency of these results, Table 5 reports performance as mean ± standard deviation over five independent runs with different random seeds. The relatively low standard deviations indicate stable behavior across repeated trials and support the robustness of PCE-FL under different non-IID settings.

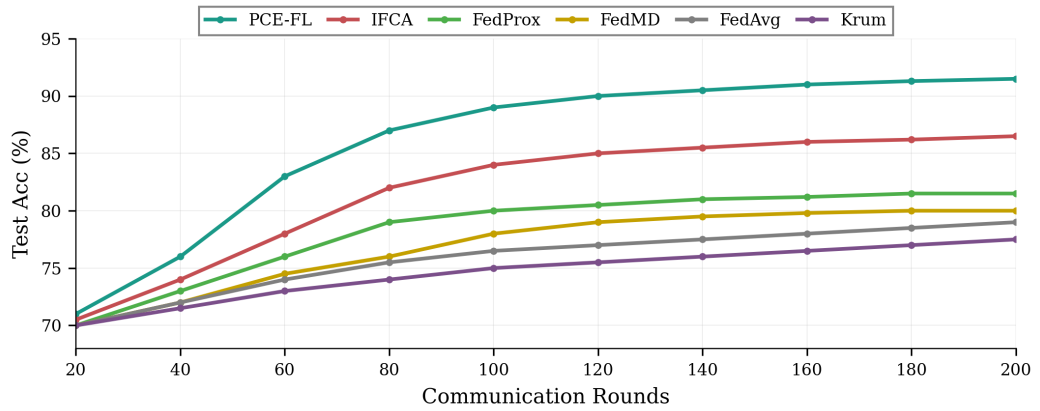

Figure 4. Accuracy vs. communication rounds under extreme heterogeneity ($\alpha = 0.1$).

5.2. Analysis of the Cost of Communication

Knowledge-distillation-based methods (PCE-FL, FedMD) are shown to eliminate 91% of communication overhead as compared to weight-sharing baselines (165 MB vs. 1850 MB for 200 rounds) and reduce per-round communication cost from 9.25 MB to 0.825 MB; in other words, PCE-FL offers a transmission speedup of 11x for a bandwidth of 1 Mbps (74 s vs. 6.6 s per round), making it ideal for deploying in bandwidth-constrained rural areas.

The results in Table 6 demonstrate that PCE-FL uniquely achieves the highest classification accuracy while simultaneously reducing communication overhead by 91%. Unlike FedMD, which improves communication efficiency at the cost of reduced predictive performance, PCE-FL maintains state-of-the-art accuracy alongside low bandwidth requirements. This balanced performance–efficiency trade-off makes PCE-FL particularly suitable for large-scale and rural federated deployments where network resources are severely constrained.

Table 6. Quantitative performance comparison of the proposed PCE-FL framework against baseline federated learning strategies.

Method	FL Strategy	Accuracy (%)	Total Comm. (MB)	Comm. Red. (%)	Convergence Speed *	Low-Bandwidth Suitability
FedAvg	Weight sharing	78.2	1850	–	Slow (85 rounds)	Low
FedProx	Weight sharing + regularization	81.5	1850	–	Moderate	Low
IFCA	Clustered FL	84.3	1850	–	Moderate (62 rounds)	Medium
Krum	Robust aggregation	80.6	1850	–	Slow	Low
FedMD	Knowledge distillation	79.0	165	91	Moderate	High
PCE-FL	Personalized KD (Ours)	89.1	165	91	Fast (58 rounds)	Very high

* Convergence speed denotes the number of communication rounds required to reach 85% test accuracy.

5.3. Ablation Study

To quantify the contribution of each core component of the proposed PCE-FL framework, a comprehensive ablation study was conducted under a high data heterogeneity setting, using a Dirichlet distribution with parameter $\alpha = 0.5$. In each ablation experiment, one module was removed from the complete PCE-FL system, while all remaining components and training conditions were kept unchanged.

The resulting framework variants were evaluated using test accuracy and macro-averaged F1-score, and the corresponding performance degradation was measured relative to the full PCE-FL configuration. This experimental design enables an isolated assessment of the impact of each component on overall performance. The quantitative results of the ablation study are summarized in Table 7.

Table 7. Ablation study of PCE-FL components under high data heterogeneity ($\alpha = 0.5$).

Framework Variant	Description	Acc. (%)	Macro F1 (%)	Degradation
PCE-FL (Full)	Clustering + knowledge distillation + reputation	91.8	91.5	–
PCE-FL w/o clustering	Single global cluster; knowledge distillation + reputation	86.5	85.9	−5.3%
PCE-FL w/o KD	Clustering + reputation; weight sharing	89.2	88.4	−2.6%
PCE-FL w/o reputation	Clustering + knowledge distillation; simple averaging	89.9	89.3	−1.9%

5.4. Ablation Findings and Discussion

Impact of Client Clustering: The highest performance degradation was caused by elimination of the client clustering module, where the test accuracy decreased to 86.5% from 91.8%. This validates clustering as the most essential element in the context of non-IID data distributions. Without clustering, the structure breaks down to a unified world model, and thus, it can no longer learn domain-specific representations over statistically similar clients.

Contribution of Knowledge Distillation: Substituting knowledge distillation with traditional weight sharing causes a 2.6% point decrease in accuracy and entirely abolishes communication efficiency gains (1850 MB vs. 165 MB). This fact shows that knowledge distillation not only saves on bandwidth consumption but enhances generalization by making the aggregation of client knowledge easier and successfully suppressing noisy updates.

Role of the Reputation Mechanism: The poorest performance is obtained when the reputation mechanism is removed with moderate heterogeneity (i.e., when the coefficient of heterogeneity is $\alpha = 0.5$), and the performance decreases by 1.9% points, implying a supportive role in this case. Its effect, however, increases under the condition of extreme heterogeneity. This explains the growing relevance of reputation-conscious aggregation in a situation where reliability and quality of data from clients diffuse significantly.

Figure 5 shows the ablation analysis of the proposed PCE-FL model applied to the PlantVillage tomato data, which demonstrates the relative significance of each of the main parts of the model in order to achieve good results. The overall PCE-FL model has the best test accuracy and macro-averaged F1-score. The most significant performance deterioration is encountered in the case of removing the clustering module, which is followed by apparent deterioration in the case of disabling either the knowledge distillation or reputation mechanism.

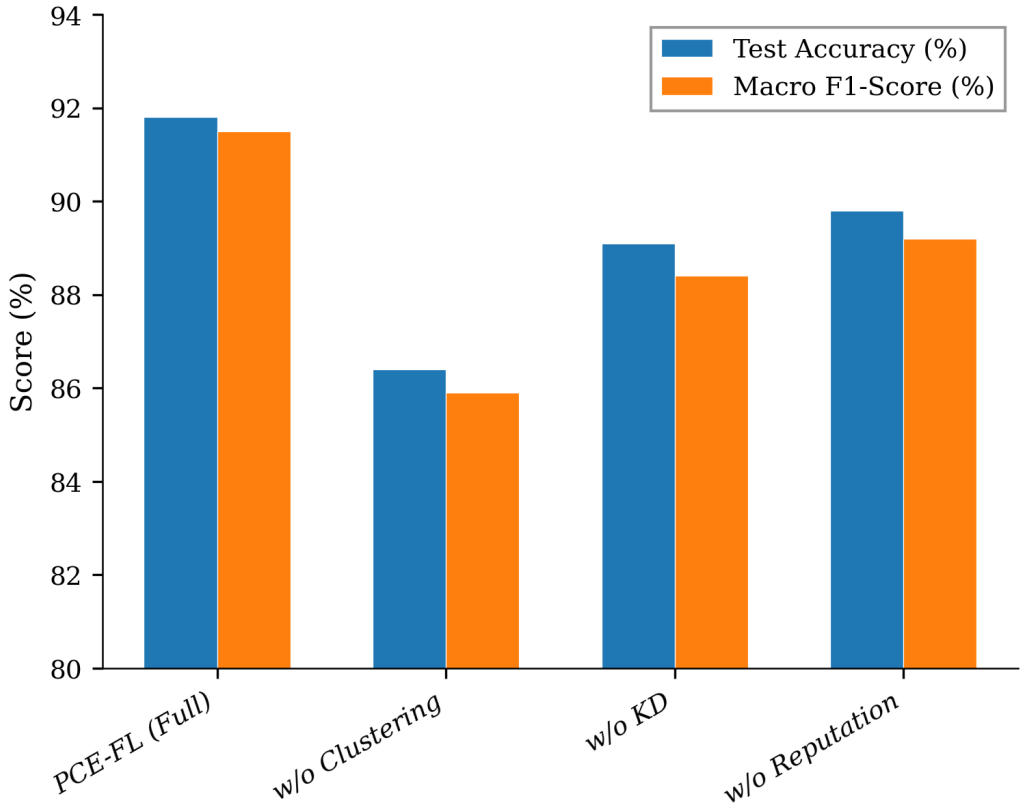

Figure 5. Ablation study of PCE-FL on the PlantVillage tomato dataset.

These findings indicate that the benefits of client clustering, knowledge distillation, and reputation-sensitive aggregation are complementary. The combination of both is necessary in order to maintain a high level of classification accuracy and balanced per-class performance with heterogeneous federated learning environments. Taken together, all of the above ablation findings indicate that the high performance of PCE-FL cannot be attributed to only one of these three pillars alone but is achieved only through the coexistence of all three pillars.

5.5. Per-Class Analysis

Figure 6 presents a heatmap of per-class F1-scores for all ten tomato disease categories across the three levels of data heterogeneity, revealing several important trends.

Disease Class Variability: Performance is not uniform across disease categories. Rare diseases, such as Spider Mite (72.1% at $\alpha = 0.1$) and Septoria Leaf Spot (78.3% at $\alpha = 0.1$), exhibit more pronounced performance degradation under extreme heterogeneity compared to more common disease classes such as Healthy (88.5%) and Late Blight (86.5%). This behavior reflects the limited availability of rare disease samples across clients under highly skewed data distributions.

Effectiveness of Clustering for Rare Diseases: Client clustering provides substantial benefits for rare disease classes. Under extreme heterogeneity, clients dominated by spe-

cific rare diseases are grouped into specialized clusters, enabling the learning of tailored representations. This leads to significantly higher F1-scores compared to a single global model trained jointly on all clients.

Stability of the Healthy Class: The Healthy class demonstrates the most stable performance across all heterogeneity levels, with F1-scores decreasing only moderately from 94.2% to 88.5%. This stability is expected, as healthy leaf samples are present in the majority of client datasets, making this class relatively insensitive to label distribution skew.

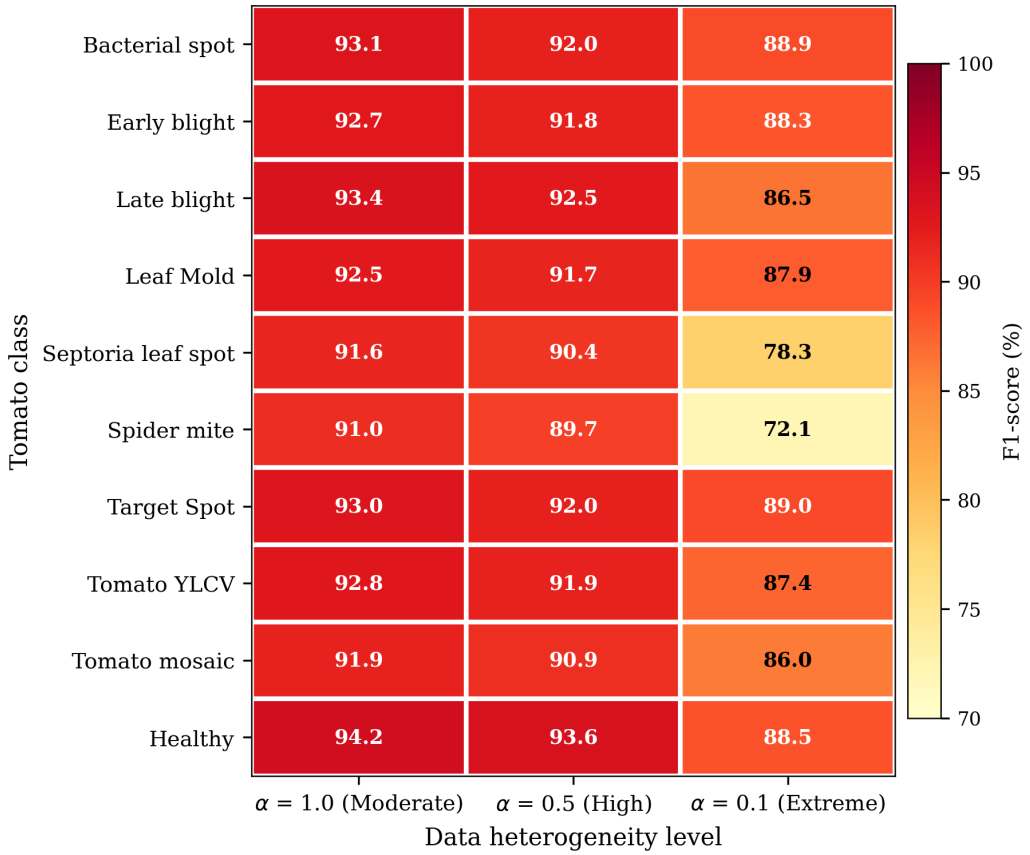

Figure 6. Per-class F1-scores of PCE-FL across heterogeneity levels.

6. Discussion

As can be seen in the experimental outcomes given in Section 5, PCE-FL offers a thorough and efficient solution to the three fundamental problems of federated learning in the agricultural context: statistical heterogeneity, ineffective communication, and resistance to erroneous data contribution. Here, we address the applied importance of these results, the constraints of the present research and the research directions.

6.1. Practical Significance and Comparison with Existing Methods

The major advantage of PCE-FL is that it is a holistic design addressing non-IID data, bandwidth restrictions, and data quality issues simultaneously, which are the most significant obstacles to the implementation of FL in the practical agricultural setting. In contrast to current solutions that tackle these issues individually (see Table 1), PCE-FL is an integrated architecture that yields synergies: clustering contributes to personalization, which leads to the improved quality of the distilled knowledge, and reputation-based aggregation ensures that the integrity of the aggregated knowledge in each cluster is preserved.

From a practical deployment perspective, the 91% reduction in communication overhead is crucial. In rural and remote agricultural areas, where only low-bandwidth cellular

networks (e.g., 2G/3G) may be available, reducing the per-round transmission size from 9.25 MB to 0.825 MB can transform a merely viable solution into a truly feasible one. This reduction decreases the transmission time per round at 1 Mbps from approximately 74 s to 6.6 s, while also significantly lowering the likelihood of transmission failures and enabling more frequent model updates.

Per-class analysis (Section 5.4) indicates that PCE-FL is of disproportionate benefit for rare disease classes, which is an aspect of the clustering mechanism. In the case of extreme heterogeneity ($\alpha = 0.1$), rare diseases like Spider Mite and Septoria Leaf Spot are the most affected by performance degradation with standard FL: specialized cluster models are more advantageous to them. This observation has direct agricultural implications, since timely diagnosis of unusual but devastating diseases is most often the most urgent requirement.

PCE-FL has certain benefits over the nearest baselines: compared to IFCA, it introduces efficiency and robustness to communication; compared to FedMD, it introduces customization by means of clustering; and compared to Krum, it substitutes costly geometric computations with a lightweight reputation mechanism that can be easily integrated with the knowledge-distillation pipeline.

A practical benefit of the suggested framework is that it allows for sharing useful knowledge across farms or agricultural enterprises without adjusting to a specific global model by equal measures. In traditional federated averaging, when the distribution of clients is strongly mismatched, negative transfer may take place; e.g., model updates of one participant can be counterproductive to another. The packed and bespoke format of PCE-FL, by contrast, promotes the movement of statistically compatible clients and minimizes the chance of one local distribution company of a particular enterprise surpassing participants that do not pertain to it. This renders the framework more appropriate in realistic multi-stakeholder agricultural networks in which the prevalence of diseases, the conditions of their acquisition, and the availability of resources differ significantly among locations.

6.2. Limitations

Even though the findings are positive, various weaknesses of this research must be recognized.

First, PCE-FL relies on a small public dataset ($\mathcal{D}_{\text{pub}}$) consisting of 500 tomato leaf images of various categories for use in knowledge distillation. Although this is a mild assumption, the proxy data may become out-of-distribution due to seasonal variations in disease patterns, which can eventually degrade the quality of distillation. This limitation can be mitigated by employing data-free distillation techniques, such as FedGen [34].

Second, the re-clustering time ($T_{\text{cluster}} = 50$ rounds) is deterministic, and any updates are synchronous. Practically, the farms can join the network intermittently, either because of connectivity or because of seasonal trends. Asynchronous CFL schemes like CASA [27] are more suited to deal with variable-rate participation of the clients.

Third, the representation mechanism uses an identical starting state ($R_i^{(0)} = 1.0$) for all clients, which is potentially not optimal when new farms enter an existing system (the cold-start problem). Regular roundwise aggregation might be enhanced by an adaptive-based initiative based on early data quality monitoring.

Fourth, we use the PlantVillage data in the simulated partitions of non-IID and not in real-world multi-farm implementations. Despite being a widely used methodology to assess FL in the context of heterogeneity, the Dirichlet-based simulation is not able to represent the complexity of real agricultural data with all of its variability in camera hardware, field conditions, the cultivar grown and labeling quality. Moreover, the PlantVillage data was also gathered under controlled environments, and the results could be different when operating in the field with complicated backgrounds and irregular lighting.

Fifth, we have not determined the role of various tomato cultivars in disease presentation and model performance. Cultivars may differ in the manifestation of a disease, and models based on cultivars may be required in production implementation.

6.3. Future Directions

Several promising directions emerge from this work:

- **Data-free distillation:** Generator-based or data-free knowledge distillation: Adding data-free knowledge distillation in the absence of a public dataset and maintaining the classification performance and communication efficiency.
- **Adaptive and asynchronous clustering:** Implementing shift-dependent clustering mechanisms that can flexibly implement cluster boundaries depending on changing disease patterns and changing seasons, together with asynchronous update protocols in between the intermittent participation of a client.
- **Enhanced privacy guarantees:** Jointly using the reputation mechanism with differential privacy (DP) and secure aggregation to address the possible logit inversion attack and membership inference attacks on knowledge vectors being transmitted.
- **Real-world field validation:** Implementing PCE-FL on real farming networks with nonhomogeneous hardware, untrustworthy connectivity, and real disease distributions to test the structure during production. These field tests would also allow the testing of cultivar-specific impacts on disease detection performance.
- **Extension to multi-crop and multi-modal settings:** It is important to expand the model to multiple crops species at once, as well as to add other forms of data (e.g., multispectral images, weather data) to monitor plant health more comprehensively.
- **Economic impact assessment:** This is the quantification of the economic benefit of using PCE-FL in the system of lost crops, efficient allocation of resources (e.g., the use of specific fungicides), and better yield forecasting of the participating farms.

7. Conclusions

PCE-FL, as shown in this paper, is a single federated learning system that combines client clustering, federated knowledge distillation, and reputation-based aggregation to overcome the simultaneous problems of statistical heterogeneity, communication inefficiency, and robustness in agricultural FL systems. Extensive experiments on realistically simulated non-IID divisions of the PlantVillage tomato dataset show that PCE-FL consistently and significantly performs better than five established baselines at all levels of heterogeneity. At high heterogeneity (Dirichlet $\alpha = 0.1$), PCE-FL results in 89.1% accuracy which is 10.9 percentage points and 4.8 points better than FedAvg and IFCA, respectively and reduces the communication overhead by 91% (165 MB vs. 1850 MB) due to knowledge distillation. The reliability and consistency of the results are ensured by the fact that all the reported improvements are statistically significant ($p < 0.001$) across five independent runs.

Ablation experiments verify that the three elements play significant and complementary roles, as client clustering offers the most significant individual contribution by solving non-IID data distributions, whereas knowledge distillation and reputation-based aggregation are essential complementary benefits to communication efficiency and robustness, respectively. Per-class analysis also shows that clustering has an unequal benefit in detecting rare diseases, which is essential to practical implementation in agriculture.

Practically, the high accuracy, low cost of communication and resistance to unreliable data contribution make PCE-FL a valid solution for practical applications of precision agriculture. The 91% decrease in overhead communication is especially applicable in rural

and edge deployments with limited bandwidth, where it can enhance the provision of prompt disease surveillance on geographically distributed farms.

Future work will focus on: (1) real-world field deployment across diverse farming networks to validate the framework under production conditions with heterogeneous hardware and connectivity; (2) integration of data-free distillation to eliminate the public dataset requirement; (3) adaptive clustering and asynchronous protocols for dynamic agricultural environments; (4) formal privacy guarantees through differential privacy and secure aggregation; and (5) extension to multi-crop, multi-cultivar, and multi-modal plant health monitoring systems. Quantifying the economic impact of PCE-FL on crop loss reduction and resource optimization represents an important applied research direction.

Author Contributions: P.G.: Conceptualisation and Writing; A.S.: Data Analysis and Review; S.G.: Conceptualization and Methodology; L.G.: Visualization and Writing; A.K.A. and C.L.C.: Validation and Review; V.S.S. and R.C.: Research Design and Formal Analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is publicly available at https://www.kaggle.com/datasets/tushar5harma/plant-village-dataset-updated, accessed on 19 January 2026.

Acknowledgments: This work is part of the sponsored research "Environmental Monitoring and Evaluation via Remote Advanced LoRa-Driven Detectors (EMERALD) for Heavy Metal Contamination" led by Middlesex University (MDX), UK and funded by SPARC-UKIERI as part of SPARC-UKIERI/2024-2025/P3228. The authors also thank Sankaran S (ex-chief scientist, CSIR) for his support and suggestions for this work.

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

Abbreviations

The following abbreviations are used in this manuscript:

IID	Independent and Identically Distributed
FL	Federated Learning

References

1. FAO. *The Impact of Disasters and Crises on Agriculture and Food Security: 2021*; FAO: Rome, Italy, 2021. [CrossRef]
2. Singh, P.P.; Kanth, D.R.; Madhuri, G.S.; Yadav, A.; Chauhan, S.S.; Kundu, P.; Bhattacharyya, U.K.; Banoo, S.; Ritu; Rajput, U.S. Deep Learning Techniques for Plant Disease Detection and Classification: A Comprehesive Review. *Int. J. Adv. Biochem. Res.* **2025**, *9*, 187–200. [CrossRef]
3. Martinelli, F.; Scalenghe, R.; Davino, S.; Panno, S.; Scuderi, G.; Ruano, O.; Barone, S.; Porta-Puglia, A.; Ravasio, A.; Gullino, M.L. Advanced Methods of Plant Disease Detection: A Review. *Agron. Sustain. Dev.* **2015**, *35*, 1–25. [CrossRef]
4. Saleem, M.H.; Potgieter, J.; Arif, K.M. Plant Disease Detection and Classification by Deep Learning. *Plants* **2019**, *8*, 468. [CrossRef] [PubMed]
5. Li, L.; Zhang, S.; Wang, B. Plant Disease Detection and Classification by Deep Learning—A Review. *IEEE Access* **2021**, *9*, 56683–56698. [CrossRef]
6. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
7. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; 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), Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282. [CrossRef]
8. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečný, J.; Mazzocchi, S.; McMahan, B.; et al. Towards Federated Learning at Scale: System Design. *arXiv* **2019**, arXiv:1902.01046. [CrossRef]

9. Lu, Z.; Pan, H.; Dai, Y.; Si, X.; Zhang, Y. Federated Learning with Non-IID Data: A Survey. *IEEE Internet Things J.* **2024**, *11*, 19188–19209. [CrossRef]
10. Bhanbhro, J.; Nisticò, S.; Palopoli, L. Issues in Federated Learning: Some Experiments and Preliminary Results. *Sci. Rep.* **2024**, *14*, 29881. [CrossRef]
11. Briggs, C.; Fan, Z.; Andras, P. Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020. [CrossRef]
12. Lin, T.; Kong, L.; Stich, S.U.; Jaggi, M. Ensemble Distillation for Robust Model Fusion in Federated Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2351–2363.
13. Barbedo, J.G.A. Factors Influencing the Use of Deep Learning for Plant Disease Recognition. *Biosyst. Eng.* **2018**, *172*, 84–91. [CrossRef]
14. Ferentinos, K.P. Deep Learning Models for Plant Disease Detection and Diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. [CrossRef]
15. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using Deep Learning for Image-Based Plant Disease Detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef] [PubMed]
16. Fuentes, A.; Yoon, S.; Kim, S.C.; Park, D.S. A Robust Deep-Learning-Based Detector for Real-Time Tomato Plant Diseases and Pests Recognition. *Sensors* **2017**, *17*, 2022. [CrossRef] [PubMed]
17. Gupta, P.; Jadon, R.S. PlantVitGNet: A Hybrid Model of Vision Transformer and GoogLeNet for Plant Disease Identification. *J. Phytopathol.* **2025**, *173*, e70041. [CrossRef]
18. Chen, J.; Zhang, D.; Zeb, A.; Nanehkaran, Y.A. Identification of Rice Plant Diseases Using Lightweight Attention Networks. *Expert Syst. Appl.* **2024**, *169*, 114514. . [CrossRef]
19. Vo, H.-T.; Quach, L.-D.; Tran Ngoc, H. Ensemble of Deep Learning Models for Multi-plant Disease Classification in Smart Farming. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 1045–1054. [CrossRef]
20. Liu, Z.; Gao, J.; Yang, G.; Zhang, H.; He, Y. Localization and Classification of Paddy Field Pests Using a Swin Transformer-Based Object Detection Framework. *Remote Sens.* **2024**, *16*, 1707. [CrossRef]
21. Murugavalli, S.; Gopi, R. Plant Leaf Disease Detection Using Vision Transformers for Precision Agriculture. *Sci. Rep.* **2025**, *15*, 22361. [CrossRef] [PubMed]
22. Salman, Z.; Muhammad, A.; Han, D. Plant Disease Classification in the Wild Using Vision Transformers and Mixture of Experts. *Front. Plant Sci.* **2025**, *16*, 1522985. [CrossRef]
23. Piccialli, F.; Della Bruna, C.; Chiaro, D.; Qi, P.; Savoia, M. AGRIFOLD: AGRIculture Federated Learning for Optimized Leaf Disease Detection. *Expert Syst. Appl.* **2025**, *289*, 128371. [CrossRef]
24. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450. [CrossRef]
25. Sattler, F.; Müller, K.-R.; Samek, W. Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 3710–3722. [CrossRef] [PubMed]
26. Smith, V.; Chiang, C.-K.; Sanjabi, M.; Talwalkar, A. Federated Multi-Task Learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4424–4435.
27. Liu, B.; Ma, Y.; Zhou, Z.; Shi, Y.; Li, S.; Tong, Y. CASA: Clustered Federated Learning with Asynchronous Clients. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), Singapore, 14–18 August 2021; pp. 1851–1862. [CrossRef]
28. Ghosh, A.; Chung, J.; Yin, D.; Ramchandran, K. An Efficient Framework for Clustered Federated Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19586–19597. [CrossRef]
29. Duan, M.; Liu, D.; Chen, X.; Liu, R.; Tan, Y.; Liang, L. FedGroup: Efficient Federated Learning via Decomposed Similarity-Based Clustering. *IEEE Internet Things J.* **2024**, *11*, 1234–1246. . [CrossRef]
30. Ma, X.; Zhu, J.; Lin, Z.; Chen, S.; Qin, Y. A State-of-the-Art Survey on Solving Non-IID Data in Federated Learning. *Future Gener. Comput. Syst.* **2024**, *135*, 244–258. [CrossRef]
31. Li, D.; Wang, J. FedMD: Heterogeneous Federated Learning via Model Distillation. *arXiv* **2019**, arXiv:1910.03581. [CrossRef]
32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]
33. Sattler, F.; Wiedemann, S.; Müller, K.-R.; Samek, W. Robust and Communication-Efficient Federated Learning from Non-i.i.d. Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 3400–3413. [CrossRef]
34. Zhu, Z.; Hong, J.; Zhou, J. Data-Free Knowledge Distillation for Heterogeneous Federated Learning. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 12878–12889.
35. Yao, D.; Huang, T.; Zhang, Y.; Geng, G.; Sattler, F.; Vrang, I.; Wang, Z.; Samek, W. FEDGKD: Toward Heterogeneous Federated Learning via Global Knowledge Distillation. *IEEE Trans. Comput.* **2024**, *73*, 3–17. [CrossRef]
36. Lamport, L.; Shostak, R.; Pease, M. The Byzantine Generals Problem. *ACM Trans. Program. Lang. Syst.* **1982**, *4*, 382–401. [CrossRef]

37. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 119–129.
38. Kang, J.; Xiong, Z.; Niyato, D.; Zou, Y.; Zhang, Y.; Guizani, M. Reliable Federated Learning for Mobile Networks. *IEEE Wirel. Commun.* **2020**, *27*, 72–80. [CrossRef]
39. Song, Z.; Sun, H.; Yang, H.-H.; Wang, X.; Zhang, Y.; Quek, T.Q.S. Reputation-Based Federated Learning for Secure Wireless Networks. *IEEE Internet Things J.* **2022**, *9*, 1212–1226. [CrossRef]
40. Dou, Z.; Wang, J.; Sun, W.; Liu, Z.; Fang, M. Toward Malicious Clients Detection in Federated Learning. In Proceedings of the 2025 ACM Conference on Computer and Communications Security (CCS), Copenhagen, Denmark, 26–30 October 2025; pp. 456–472. [CrossRef]
41. Mughal, F.R.; Ullah, R.; Ikram, M.; Ahmad, N.; Alamri, A.; Alharbi, M. Adaptive Federated Learning for Resource-Constrained IoT Devices through Edge Intelligence and Multi-Edge Clustering. *Sci. Rep.* **2024**, *14*, 28746. [CrossRef]
42. Yang, X.; Li, J. A Clustering-Based Federated Deep Learning Approach for Enhancing Diabetes Management with Privacy-Preserving Edge AI. *Healthc. Anal.* **2025**, *7*, 100392. [CrossRef]
43. Sheikhi, S.; Kostakos, P.; Loven, L. Hybrid Reputation Aggregation: A Robust Defense Mechanism for Adversarial Federated Learning in 5G and Edge Network Environments. *arXiv* **2025**, arXiv:2509.18044. [CrossRef]
44. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
45. Opitz, J.; Burst, S. Macro F1 and Macro F1. *arXiv* **2019**, arXiv:1911.03347.