DOI: https://doi.org/10.71317/kjard.2.6.2026.280

The Kashmir Journal of Academic Research and Development

Journal homepage: https://rjsaonline.org/index.php/KJARD

A Secure, Privacy-Preserving Federated Learning Framework for Smart Agricultural Supply Chains against Adversarial Data Poisoning in Supply Chain 5.0

Muhammad Irfan
Student, Institute of Computing (IoC)
Email: 2020-uam-2108@mnsuam.edu.pk

Dr. Amir Hussain
Associate Professor, Institute of Computing (IoC)
Email: aamir.hussain@mnsuam.edu.pk

Dr. Salman Qadri
Professor, Institute of Computing (IoC)
Email: salman.qadri@mnsuam.edu.pk

Muhammad Yasir Khan
Lecturer, Institute of Computing (IoC)
Email: yasir.khan@mnsuam.edu.pk

Dr. Kashif Razzaq
Professor, Department of Horticulture
Email: kashif.razzaq@mnsuam.edu.pk

ARTICLE INFO

Received:
April 21, 2026
Revised:
May 05, 2026
Accepted:
May 19, 2026
Available Online:
June 04, 2026

Keywords:
Federated Learning; Data Poisoning; Robust Aggregation; Smart Agriculture; Supply Chain 5.0; Internet of Things; Privacy Preservation; Edge Computing

Corresponding Author:
aamir.hussain@mnsuam.edu.pk

ABSTRACT

Modern farming supply chains rely more and more on distributed IoT sensing technologies to enable prediction analytics across the entire supply chain for crop monitoring, harvest planning and logistics downstream. However, if machine learning pipelines are centralized, sensitive telemetry data collected in the field must constantly be funneled to remote cloud servers, posing significant confidentiality concerns while using up significant bandwidth and placing all of the risk in a single point of failure. Federated Learning (FL) alleviates the data-governance issue through edge model training and only sending global model parameter updates, but also makes the global model vulnerable to adversarial data-poisoning attacks from compromised farm gateways. In this paper, a secure and privacy-preserving FL framework is proposed that is tailored for decentralized precision agriculture under Supply Chain 5.0 paradigm. The framework combines localized training on the farm with a robust aggregation mechanism that filters out anomaly data by reviewing the pair-wise geometric variance of the data from the different clients before data is synthesized globally. The system was tested on a real multi-sensor data set from agriculture consisting of 16,411 IoT telemetry instances from soil moisture, ambient temperature, and atmospheric humidity, with each node collected from a different farm and all non-IID. The evaluation specifically targets a severe threat regime in which 40% of the network (two of the five nodes) is attacking the network with a coordinated attack using data poisoning. Under this attack, an undefended FedAvg baseline struggles to remain above 77.8% accuracy; on the other hand, the proposed centroid-filtered framework gradually increases its accuracy up to around 92.6% over 20 communication rounds, which is almost the level of the benign reference accuracy and makes the security premium about 1%. 4-way ablation (with respect to coordinate-wise median, Krum, proposed filter) reveals that the proposed filter is the only filter which maintains a stable trajectory and achieves the highest terminal accuracy, while Krum is oscillating randomly in the presence of colluding minority. Stress testing with different adversarial fractions (0% to 60%) shows that the accuracy of the defended model is close to benign up to 40% adversarial fraction, at which point the adversarial fraction becomes the majority, after which the accuracy drops significantly. The held-out set consists of 3,283 samples and the protected model achieves 96.5% classification accuracy and macro-average ROC AUC of 0.993 on this set. The framework provides an evidence-based platform to deploy trusted distributed intelligence in the agriculture edge network.

Introduction

Industrial production systems are undergoing a structural shift from an Industry 4.0 mindset that prioritizes efficiency, to a more human-centric perspective and resilience to shocks and stresses, that embraces sustainability and is often referred to as Industry 5.0 [1,2]. In the logistics context, this transformation has manifested itself as Supply Chain 5.0 – autonomous cyber-physical systems, collaborative intelligence and distributed decision-making to provide supply networks that are not only fast and cheap, but also clearly resilient against disruptions [3] [4]. Perhaps no other industry can benefit from it more than farming. The Agri-food chains are long, geographically spread, high-value and perishable goods, thus being the most valuable and at the same time the most fragile ones for large scale digital transformation projects [5].

The technology base of this metamorphosis exists today. Different micro-climate and agronomic parameters of individual fields and different compartments of greenhouses are currently monitored by dense networks of low-cost sensing devices, such as soil moisture, ambient temperature and atmospheric humidity [6, 7, 8]. In the agriculture sector, it has been consistently found that implementing IoT devices for telemetry coupled with machine learning algorithms can be used to better plan irrigation, anticipate disease outbreaks, estimate yields and plan the cold chain process [9] [10]. The role of edge computing in this picture is to shift the inference (and, more and more, the training) towards the location of the data generation, thus reducing the round-trip latency and the amount of raw data that has to cross narrow rural links [11] [12].

However, commercial deployments are still stuck with the dominant analytical architecture that is centralized. The raw telemetry data is pushed to a cloud platform from farm gateways and trained with a single global model which is served. This has a number of drawbacks that exacerbate each other. First, the transfer of continuous streams of multiple sensor data at high frequencies is overwhelming the bandwidth and also causes latencies that are detrimental for real-time actuation at the edge. Second, putting all fine-grained agro-ecological data in a single repository exposes information of strategic importance: field-level yield profiles, input regimes and stress signatures are commercially sensitive data that leak can be used for the profiling of competitors and manipulation of the market. Third, the central repository is an appealing SPOF, a breach or availability event ripples out immediately to each farm with the common model, a risk well known in the general IoT security literature [13].

That is the reason for the introduction of Federated Learning (FL), which aims to eliminate this reliance on raw-data centralisation [14] [15]. In the canonical formulation, participating clients have local training modules and a coordinating server distributes a global model to them and they train locally on their private data and then report back only parameter updates, which are then classically fused (e.g., via Federated Averaging (FedAvg)) into a new global model [14], [16]. The paradigm has gained momentum very quickly, and has been extensively discussed in terms of open problems [17] and methodological aspects [18] as well as in the specific case of IoT and mobile-edge environments [19] [20] [21]. Early on, production systems like federated keyboard prediction showed that it works well with millions of occasionally connected devices [22]. FL is a natural fit for agriculture, since there will be farm operators who are the owners and operators of their own telemetry and the cooperative will benefit from a model trained on the union of all experience.

However, this brings with it a new attack surface with the privacy benefit. The coordinator has no way to independently check the integrity of the process that generated any particular update as the raw training data is never inspected by the coordinator. An adversary can poison the local training set for the farm gateway—by flipping labels or changing features—or can tamper with the gradients computed by the gateway before upload, by a malicious insider controlling the gateway. This kind of data-poisoning and model-poisoning attack is well studied in the adversarial machine learning literature [23] [24] [25] and has clear tangible consequences to an agricultural setting: a poisoned model that reports a state of crop stress when in fact it is not, will generate unnecessary transport logistics alerts, divert refrigerated transport capacity, and undermine the trust that federations rely on to function voluntarily.

That's what this paper does, it deals directly with that vulnerability. We propose, develop and evaluate (empirically) a secure, privacy-aware FL system for decentralized prediction of crop status in smart agricultural supply chain. The main idea of the framework is to build a multi-layer robust aggregation mechanism: At each step of communication, the orchestrator calculates the two-by-two Euclidean distance structure between the client updates, it then computes a consensus-proximity score for each client, and rejects updates whose geometric proximity to the consensus point is statistically outlying before performing the weighted global aggregation. The design is based on the premise of an untrusted network and places the security burden, initially from the perimeter defence (which is impractical in thousands of remote gateways), to the algorithmic verification at the aggregation layer.

The framework is tested on a real multi-crop agricultural IoT data set of 16,411 telemetry instances from seven attributes, which are deliberately split across five distinct farm nodes to simulate the type of data landscapes with multiple crop types and data sizes typical of agricultural federations in a region. The evaluation is based on a carefully crafted threat regime, a systematic attack of data poisoning from 40% of the network (two of five nodes), and considers four different angles: convergence of defederation against a reference federation that does not defend itself, ablation against a coordinate-wise median and Krum aggregator, stress test with varying the adversarial fraction to identify the break-point in resilience, and discriminative-quality analysis of the protected model using ROC, precision-recall, and confusion-matrix results.

The study has three research questions. RQ1: How much does a data-poisoning attack from a minority of farm nodes that is coordinated by the attacker reduce the predictive capability of a model federated with a conventional method trained with real agricultural telemetry data? RQ2: Is it possible to recover productiveness to a nearly baseline level by using a distance based anomaly-filtering aggregation mechanism without using any raw data from the clients? RQ3: What are the utility cost, in terms of accuracy, class balance and convergence behaviour of the security mechanism under benign and adversarial conditions?

The paper contributes to the answers of the above questions in the following ways:

Internet of Things (IoT) devices enable the collection of data from various points around a farm, which is often fragmented and non-IID, a condition that hinders the construction of models on a large scale.Data from multi-sensor telemetry are collected by IoT devices and kept resident at the nodes of a farm, and also the data are often fragmented and non-IID, which makes the construction of models on a large scale difficult.

- A statistically sound aggregation mechanism for anomalies based on the geometric pairwise variance between the parameters sent by clients, that filters out outliers' influences before passing them on to the global model.
- A thorough empirical assessment of the proposed framework on 16,411 instances of real IoT telemetry devices under a strong 40% coordinated poisoning attack, which learns to approximately 92.6% accuracy after 20 communication rounds, compared to a reference FedAvg that does not defend and reaches only about 77.8% accuracy.
- An ablation study with four different defence methods: FedAvg, coordinate-wise median, Krum and the proposed centroid filter, an adversary-scaling stress test to find the majority break-point of the defence empirically and a discriminative-quality analysis of the protected model (96.5% held-out accuracy, macro-average ROC AUC of 0.993) with full confusion-matrix evidence.

The rest of the paper goes as follows. Section 2 summarizes the related works on the agricultural IoT, federated optimization, poisoning attacks and distributed ledger approaches, and presents the gap that is addressed by the work. Section 3 presents the formalisation of the system model, threat model, data pipeline and the proposed secure aggregation algorithm. Section 4 presents the results of the experiments: (a) Convergence of the defended federation when the defender's share is 40% of the total attack, (b) Four-way defence ablation, (c) Adversary-scaling stress test, and (d) ROC, precision-recall and confusion-matrix analysis of the protected model. The results, interpreted in Section 5, are compared to previous defenses and the industrial implications and limitations of these results are discussed. Section 6 concludes.

Literature Review

The state of the art and research relevant to secure federated intelligence in agricultural supply chains, has followed several initially distinct lines: digital agriculture and its sensing infrastructure, distributed ledger technology for provenance, federated optimization, adversarial machine learning and Byzantine-robust aggregation, which is the frontier this paper is in the intersection of. This convergence helps to explain what has been accomplished in this field and where it is vulnerable.

The groundwork for the empirical orientation was provided by the smart-farming literature of the late 2010s. Wolfert et al. mapped the growing use of big data throughout the farm business cycle and noted (correctly, at the time) that issues of data ownership and governance would play a significant role in adoption as well as the analytical capabilities of the data [5]. Systematic surveys of deep learning in agriculture [6] and of machine learning in agriculture [7] more generally, catalogued the substantial improvements in deep learning for yield prediction, disease detection and species classification in agriculture, while surveys of agricultural IoT catalogued the layers of sensing, networking and platform from which the data emerges [8] [9] [10]. One common thread throughout this body of work was the conflict between the individual operator's desire to keep their data private and the analytical potential of data gathered from many farms; a conflict highlighted by the general IoT security analysis of Sicari et al [13] who

identified confidentiality, trust management and resilience as issues that are yet to be solved for larger deployments of sensors. The move of computation toward the network edge, as espoused in the pioneering work on edge and mobile-edge computing [11, 12] lessened the impact of latency and bandwidth, while the governance issue – who gets whose raw data – remained mostly unchanged.

The first wave of responses was for distributed ledgers. The possibilities of blockchain designs for agri-food chains were envisioned as traceability with tamper evidence and automation of compliance via smart contracts, and indeed, reading the survey literature, we see a lot of experiments ranging from field-to-fork traceability to token mediated cooperative schemes [70, 71, 72]. The placement of such efforts was within the context of sustainable supply chain management (SSCM) by Saberi et al., who put forth inter-organizational trust as the major hurdle for the adoption of these ledgers, which blockchain can solve [73]; Ferrag et al. explored the use of blockchain specifically as a security substrate for green-IoT agriculture and listed the open challenges associated with these applications [69]. However, this literature is also aware of the analytical limitations of the paradigm: a ledger provides assurance that data recorded on the ledger was not tampered with afterwards, but not that it was learned from, and high-frequency raw telemetry cannot be replicated and stored on a chain, it is too costly both technically and financially. In short, provenance can be used in conjunction with, but not in place of, privacy-aware distributed learning.

Federated optimization brought that learning capability along. FedAvg was proposed by McMahan et al., who showed that, with some communication-efficiency studies [15] in hand, it is possible to achieve a close performance to the centralized accuracy without exchanging the raw data [14]. The taxonomy provided by Yang et al. [16] and the open problems listed by Kairouz et al. [17] and the methodological challenges identified by Li et al. [18] provide a solid basis for describing horizontal, vertical and transfer federation. Application-oriented surveys demonstrated the paradigm's suitability for IoT and mobile-edge scenarios [19, 20, 21] and the usage of federated keyboard prediction in application at population scale put an end to any possible doubts about practical feasibility [22]. The turn of agriculture followed: Durrant et al. proved that cross-silo federation can enable data collaboration in the agri-food sector where complete data sharing is not possible due to commercial reasons [66]; the extensive survey of Friha et al. identified FL as one of the enabling technologies of smart agriculture of the future [67] and FELIDS presented federated intrusion detection specifically designed for agricultural IoT traffic [68]. However, most of the agricultural studies follow the benign-participant assumption of early FL, where aggregation is done by simple or weighted averaging, and where a possibility that a farm gateway participating in the FL could be the adversary is not modeled.

The adversarial machine learning literature reveals all the reasons why this is not true. Poisoning attacks go back more than 10 years ago, when Biggio et al. showed how to do poisoning attacks against support vector machines [23], but it is exacerbated in the federated setting as the coordinator is structurally blind to the local data. Tolpegin et al. exhibited how a small number of clients can flip their labels, causing significant loss of global accuracy in image classification federations [24] and Fang et al. exhibited how to optimistically poison local models that eventually results in a loss of global accuracy even in aggregation rules designed to be Byzantine-robust [25]. The pictures of the threat grew worse on the backdoor line of work: Bagdasaryan et al. demonstrated that it is possible to embed misbehaviour in a target model through backdoor and replacement [26]; Bhagoji et al. analysed the backdoor poisoning problem using an adversarial perspective [27]; Wang et al. formally argued that backdoor poisoning is inevitable for undefended systems when considering the edge-cases [28]; and Xie et al. demonstrated that distributed backdoors across colluding clients further evades detection [29]. Then Shejwalkar and Houmansadr combined attack optimization of a wide range of defence [30]. Complementary results on back-gradient poisoning [31], poisoning of regression learners [32] and certified defence limits [33] complete a literature that is unequivocal: any aggregation rule that assumes all the updates arriving are equally trustworthy is an open door.

The answer to the defensive research has been mostly at the level of aggregation. Blanchard et al. proposed Krum, where the update with minimum distance (in squared Euclidean distance) to its neighbors is selected as the consensus for the round [34]; Yin et al. gave statistical guarantees to update using the coordinate-wise median and coordinate-wise trimmed median rules [35]; El Mhamdi et al. identified remaining weaknesses of these rules in high dimensions and proposed Bulyan as a hardened combination rule [36]. Later, other systems have generalized the trust signal; FLTrust bootstraps trust from a small root dataset on the server-side [37]; FoolsGold punishes the suspicious similarity of sybil updates [38]; RFA uses a geometric median instead of the mean [39]; Zeno scores updates based on suspicion-based performance estimation [40]. A more recent generation focuses on backdoors in particular, such as clustering, and noising in FLAME [41], deep model inspection in DeepSight [42] and robust learning-rate manipulation [43] as well as historical-consistency detection of malicious clients in FLDetector [44]; and empirical studies question how far norm-clipping alone can take the defence [45]. Existing surveys of this fast evolving field [46, 47, 48] agree that distance- and similarity-based filtering is the most expedient family of defenses, as they do not require any additional clean data, but they have yet to be evaluated on the low-dimensional, tabular, highly non-IID telemetry generated by real federations of agriculture.

A further restriction on the defence design is the privacy strand of the literature. However, gradient leakage can also be unintended as Melis et al. showed in the context of collaborative learning [49], Zhu et al. were able to reconstruct the training samples by issuing shared gradients [50] and Nasr et al. quantified the information leak of white-box membership inference attacks against federated participants [51]. Examples of mitigations involve differentially private training [52] based on the formal framework of Dwork and Roth [53] and adapted to client level federation by Geyer et al. [54], cryptographic secure aggregation [55] and secure multi-party computation systems like SecureML [57] and hybrid protocols [56]. As mentioned above, Byzantine resilience and cryptographic secrecy are not mutually exclusive, as demonstrated by So et al. [58] who provided a non-trivial protocol with such properties. For bandwidth-limited applications in agriculture these cryptographic layers are still a dream, but what is needed in the near-term is a defence that can work on plaintext updates without requiring extra data or significant computations, and that is exactly the sweet spot of the distance-based filtering.

The third strand relates to statistical heterogeneity, not an unwanted parameter, as in agriculture but rather a necessary condition. The accuracy loss due to non-IID client distributions has been quantified by Zhao et al. [59]; proximal regularization has been introduced to stabilize the heterogeneous optimization of FedProx [60]; SCAFFOLD was proposed to learn to correct the drift across clients using control variates [61]; FedCon has conducted convergence analysis of FedAvg's behavior under non-IID sampling [62]; FedCon further studied the non-IID quagmire across decentralized settings [63] and FedMix, a robust communication-efficient variant of FedCon, has been proposed [64] while FedNorm penalized objective inconsistency with normalization of the averaging [65]. Heterogeneity has a sinister effect on strong aggregation: If it is "benign" (as in the case of different crops using different soils), then updates will naturally vary and a filter that seeks to remove updates that meet IID vision specifications may end up removing some honest minority updates as well as some dishonest ones. On real agricultural telemetry the question of whether a satisfactory operating point can be found is in the end an empirical one which cannot be answered by the literature.

An analysis of these strands together reveals a certain and significant space that is neglected. The agricultural FL literature shows feasibility but neglects adversaries; the strong aggregation literature models adversaries but only on artificial partitions of vision tasks; the literature of Supply Chain 5.0 [3, 4, 74] express the resilience requirements of Supply Chain 5.0 but does not provide an end-to-end secured learning architecture which meets those requirements. What is missing is an empirically validated framework that (i) works on real agricultural telemetry based on multiple sensors, (ii) supports a non-IID partitioning of real farm federations that varies in size and is specific to each crop, (iii) resists against a coordinated poisoning of the data by a meaningful number of participants, and (iv) quantifies utility cost of doing so at the granularity, per-class behaviour, convergence trajectory, and single-inference case analysis, required by industrial adopters. The framework that is developed in the rest of this paper will fill just that gap.

Methodology

The proposed secure federated learning framework is summarized in this section. We provide a description of the overall system architecture and the adversarial threat model, then detail the dataset, preprocessing pipeline and non-IID partitioning strategy and present the mathematical formulation of the learning objective, the anomaly-filtering robust aggregation mechanism, the complete training algorithm, and the experimental configuration used for evaluation.

The overall framework and system architecture of the system are presented

The proposed architecture is supposed to model a distributed precision-agriculture supply chain network, which consists of a centralized trusted orchestrator and a set of $\mathcal{I} = \{1, 2, \ldots, K\}$ distributed edge clients, where each client corresponds to a different and autonomous smart farm node. Each farm node hosts a local IoT sensing layer, soil moisture probes, ambient temperature and atmospheric humidity sensors spread throughout the farm, with the readings of the sensors gathered in an edge node located at the farm. The gateway will have a copy of the predictive model, and all the training will be done locally in the gateway, without raw telemetry ever leaving the premises of the farm. The model parameter vectors resulting from this are only sent upstream to the orchestrator who will validate, aggregate and redistribute the global model. These are illustrated in the layered architecture shown in Figure 1, as well as the end-to-end telemetry research process summarized in Figure 2.

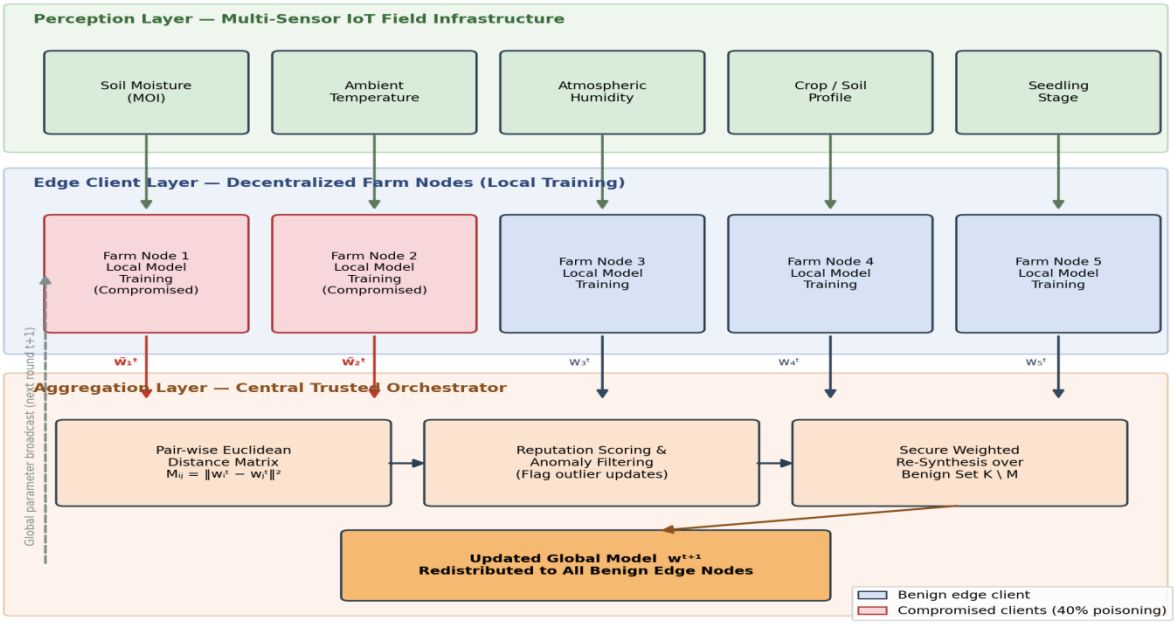

Figure 1. Proposed secure federated learning architecture for the smart agricultural supply chain, showing the perception (IoT sensing) layer, the distributed farm-node training layer with two compromised clients (40% of the federation), and the central secure aggregation layer.

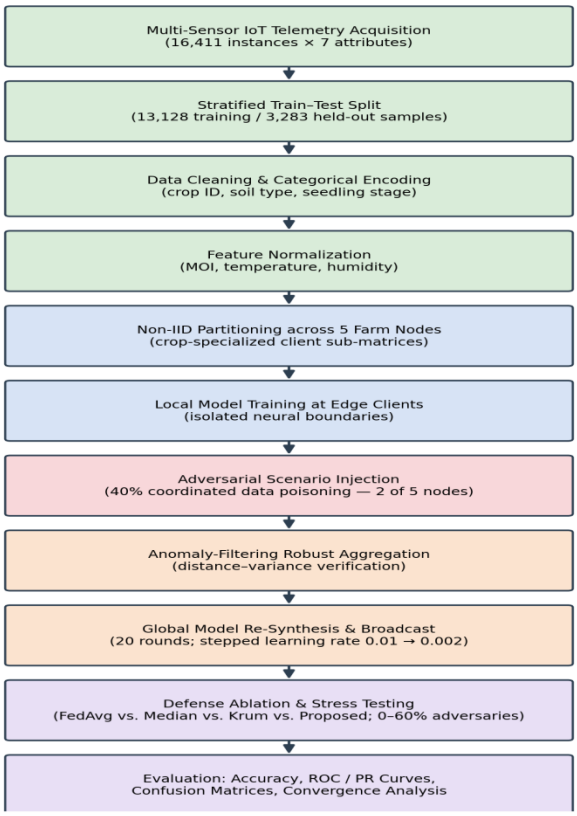

Figure 2. Research workflow: data acquisition → train–test split → cleaning, encoding, and normalization → non-IID client partitioning → local training → adversarial injection (40%) → anomaly-filtering secure aggregation → global re-synthesis over 20 rounds → defence ablation, stress testing, and evaluation.

Threat model. The orchestrator is assumed to be honest and the client group is semi-trusted. An adversary compromises some farm nodes $\mathcal{M} \subset \mathcal{I}$, in our main evaluation 40% of the network (two of five clients, labeled Farm Nodes 1 and 2), and uses access to these nodes to coordinate a data-poisoning attack: The local training data of compromised nodes is modified (e.g., via label flipping or fake sensor readings) so that the parameter updates they send push the global model away from the true decision boundaries. The attacker is not able to view or alter the updates done by the benign clients and is not able to alter the orchestrator itself. This is the usual minority-Byzantine scenario of the rich aggregation literature [25, 34, 35] where, in the case at hand, the attack cardinality is $|\mathcal{M}| < K/2$. The 40% configuration is designed to lie just on the edge of this limit, one node away from the theoretical limit of any majority-consensus defence, and the experiment with stress-test in Section 4 also varies $|\mathcal{M}|$ from 0 to 3 nodes (0% to 60%) to test the edge practically.

Dataset and its reprocessing

The empirical evaluation is done with a realistic multi-crop agricultural IoT data set with 16,411 telemetry instances and seven attributes. Each record is a combination of crop identity and agronomic context (crop ID, soil type, seedling stage) and a multi-sensor environmental reading synchronized with the agronomic context (soil moisture index, ambient temperature, atmospheric humidity) and a categorical target variable indicating the crop status. The corpus consists of 13,128 training samples, and 3,283 held-out samples, used only for global model evaluation, which are then distributed over the federated clients. Table 1 gives the characteristics of the data sets.

Table 1. Dataset characteristics

Attribute	Type	Description
crop ID	Categorical	Crop species identifier (e.g., Wheat)
soil_type	Categorical	Soil classification of the field (e.g., Black Soil)
Seedling Stage	Categorical	Growth stage of the crop (e.g., Germination)
MOI	Continuous	Soil moisture index reading
temp	Continuous	Ambient temperature (°C)
humidity	Continuous	Atmospheric humidity (%)
result	Categorical (target)	Crop status class label

Note: Total instances: 16,411; attributes: 7.

The pre-processing is done in two steps. The categorical attributes, crop ID, soil type and seedling stage, are discretized into numerical vectors by applying a map $\Phi : S \rightarrow \mathbb{Z}$ that is deterministic, thus producing the same code for the same category for all the clients network-wide. However, when using continuous features (MOI, temperature, humidity), these features are scaled to a common numeric range to prevent one sensor's physical values from skewing the gradient signal, due to the different physical values of the various sensors, if this step is not taken. The cleaned and encoded matrix is then checked if the matrix is complete before partition, and in the pipeline, the confirmation message of successful pre-processing of all 16,411 rows of matrix is generated.

Partitioning of non-IID data among the farm nodes

The global dataset D is split in five different client sub-datasets by crop specialization: each of the farm nodes contains a set of telemetry readings of a particular crop label, and the resulting local datasets are label-skewed and size-imbalanced. This partitioning is done intentionally to be non-IID: real farm cooperatives consist of members who are cultivating different crops, in different soils, on different scales and any defense that can work only under the IID assumption would be of little practical use. The result is reported in the client configuration in Table 2.

Table 2. Non-IID client partition of the global dataset

Farm node	Dominant crop label	Corpus instances bearing the label
Farm_Node_1	4	6,213
Farm_Node_2	2	2,504
Farm_Node_3	0	2,475
Farm_Node_4	3	2,455
Farm_Node_5	1	2,764

Note: Total: 16,411 instances. Client training pools are drawn from the 13,128-sample training partition following this specialization; Farm Nodes 1–2 host the compromised clients in the primary 40% attack configuration.

In terms of the training corpus, the partition can be formally written as $D \rightarrow \{D1, D2, ..., D5\}$ where $D = \cup k\ Dk$, and each of the client pools Dk are randomly sampled from the training partition D based on the specialization of the labels as shown in Table 2. The distribution of the labels is highly skewed, with Farm_Node_1 being about 2.5 times the mass of the corpus of the smallest stratum where this is to be addressed by the aggregation rule using sample-size weighting.

Enable local training and global learning

Each client $k \in \mathcal{I}$ has a private dataset $D_k = \{(x_i, y_i)\}$, with $i = 1... N_k$, where x_i is an encoded feature vector of the telemetry data and y_i is a categorical label for the crop status. The federation aims to find the global parameter vector $w \in \mathbb{R}^d$, that minimizes the sample-weighted empirical risk for all clients, while not requiring to centralize any raw data in the federation:

$$\min\_w\ F(w) = \Sigma\_\{k=1\}\^\{K\}\ (N_k\ /\ N) \cdot F_k\ (w) \qquad (1)$$

With $F_k\ (w)$ being the local objective of client k, which is the mean multi-class classification loss of the private samples of client k:

$$F_k\ (w) = (1\ /\ N_k\ ) \cdot \Sigma\_\{i \in D_k\ \}\ L(x_i, y_i; w) \qquad (2)$$

In this case, $L(\cdot)$ is the per-sample classification loss that maps the input telemetry features to the categorical crop-status target, w is a shared d-dimensional vector of model parameters, $N_k = |D_k|$ is the number of samples in each location D_k, and N is the total number of samples. At each communication round t, every client takes the current global parameters w^t, does D_k E epochs of local optimization with w_k^t with a round dependent learning rate η_t, which is 0.01 for the first 11 rounds and 0.002 for the subsequent communication rounds, and sends the results w_k^t to the orchestrator. Information required from Author: Local model architecture (e.g. the configuration of layers of the optimized multi-layer classifier), local model epochs E and local batch size.

Secure Aggregation with Anomaly Filtering

In adversarial situations, the malicious subset $\mathcal{M}$ sends altered parameter vectors that are meant to replace the global model. Classical FedAvg update is not able to defend against it as it trains by combining all of the submissions according to a fixed linear rule. The proposed framework introduces a simple statistical verification phase between reception and aggregation, taking advantage of a basic geometric property: if the parameters are trained on poisoned data, in order to have some impact, they should deviate from the benign consensus region, which can be detected without ever looking at raw data.

After gathering the client vectors $\{w_k^t\}$ at round t the orchestrator can calculate the complete matrix of pairwise squared Euclidean distance among clients:

$$M_{ij} = \| w_i^t - w_j^t \|_2^2\ ,\quad \forall\ i, j \in \mathcal{I} \qquad (3)$$

With M_{ij} being a measure of the geometric distance between the updates of client i and client j. The distances to its S nearest neighbours are then added up, with $S = K - |\mathcal{M}| - 2$, by construction, a benign client's score will be calculated mainly against other benign ones, thus giving the consensus-proximity (reputation) score R_k to each client.

$$R_k\ = \Sigma\_\{j \in \mathcal{N}_k\ \}\ M_{k\ j} \qquad (4)$$

Where, $\mathcal{N}_k$ represents the set of S clients with the most similar parameter vectors to client k's. The low scores show updates which are part of a close consensus cluster; high scores show geometric outliers. The orchestrator keeps the good set $\mathcal{I}$_secure consisting of the client with the lowest score and removes the others as possibly adversarial contributions. Then only confirmed updates are re-synthesized into the global model, with a sample-size weighted delta rule.

$$w^{t+1} = w^t + \Sigma\_\{k \in \mathcal{I}\_secure\}\ (N_k\ /\ N\_secure) \cdot (w_k^t - w^t) \qquad (5)$$

N_secure is the sum over $k \in \mathcal{I}$_secure of verified nodes' sample volume N_k and $w_k^t - w^t$ is the delta parameter contributed by client k. As in FedAvg, larger honest datasets are given proportionately more influence in Equation (5) and it ensures that flagged updates contribute nothing to a new global state. All the clients receive the updated parameters w^{t+1} and the iteration goes on until convergence. Figure 3 shows the entire secure aggregation pipeline.

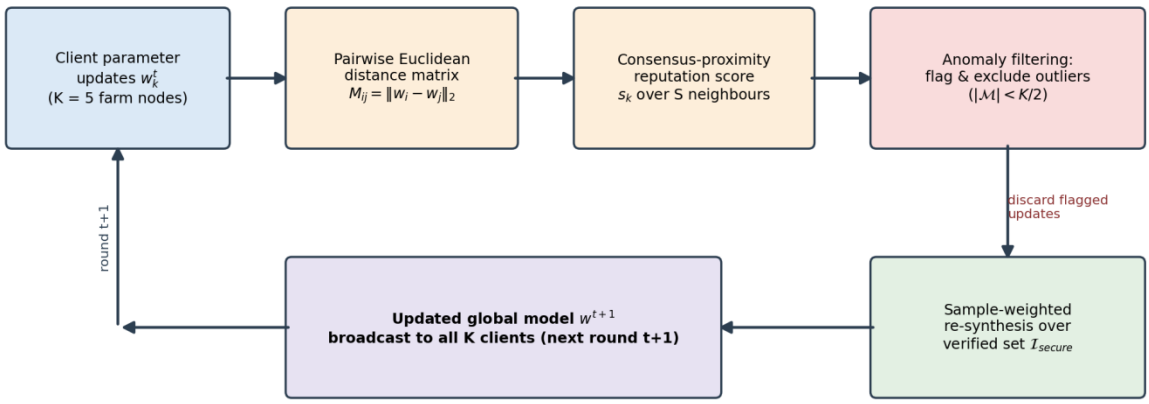

Figure 3. Secure aggregation pipeline at the central orchestrator: distance-matrix construction, consensus-proximity scoring, adversarial filtering, and sample-weighted re-synthesis of the global model.

Algorithmic Implementation

Algorithm 1 consolidates the complete training procedure, combining local client optimization with the orchestrator-side anomaly-filtering aggregation described above.

Algorithm 1: Resilient Multi-Layer Secure Federated Learning Framework

---

Input : Client datasets D_1, ..., D_K; rounds T; local epochs E;

learning rate η; assumed attacker cardinality $|\mathcal{M}|$

Output: Resilient global model parameters w^T

---

1. Initialize global parameters w^0
2. for round t = 0, 1, ..., T − 1 do
3. for each client k ∈ $\mathcal{I}$ in parallel do
4. $w_k^{\,t} \leftarrow$ LocalClientTraining(k, w^t, E, η_t)
5. end for
6. // Orchestrator-side anomaly-filtering pipeline
7. for i = 1 to K do
8. for j = 1 to K do
9. $M[i, j] \leftarrow \| w_i^t - w_j^t \|_2^2$
10. end for
11. end for
12. $S \leftarrow K - |\mathcal{M}| - 2$ // neighbourhood size

13. for each client k = 1 to K do
14. sort row M[k, :] in ascending order
15. R_k ← sum of first S elements of sorted row
16. end for
17. $\mathcal{J}$_secure ← clients with lowest scores R_k
18. N_secure ← $\Sigma_{k \in \mathcal{J}\_secure} N_k$
19. // Robust weighted re-synthesis
20. $w^{t+1} \leftarrow w^t + \Sigma_{k \in \mathcal{J}\_secure} (N_k / N\_secure)(w_k^t - w^t)$
21. end for
22. return w^T

The additional computation cost associated with the defence is only in the orchestrator and is limited to the $O(K^2d)$ distance-matrix construction and an $O(K^2 \log K)$ scoring pass, which is negligible in the present 5-client federation studied and is feasible for cross-silo federations of realistic size. An interesting property of client-side computation is that it is the same as FedAvg.

Experimental Setup and Evaluation Metrics

The framework was developed in Python with the use of scikit-learn for the construction of the models, NumPy for numerical routines and Matplotlib for visualization. Table 3 shows the experimental set-up: entries marked with an asterisk are to be confirmed by the author.

Table 3. Experimental configuration and hyperparameters

Parameter	Value
Number of clients K	5
Communication rounds T	20
Data partitioning	Non-IID, crop-label-specialized, size-imbalanced
Adversarial node ratio	40% (2 of 5 nodes; Clients 0–1) in the primary configuration; 0–60% in the stress test
Attack type	Coordinated data poisoning (corrupted local updates)
Training / held-out samples	13,128 / 3,283
Local model architecture	Optimized multi-layer classifier [layer configuration: Information Required from Author]
Local epochs E	[Information Required from Author]
Learning-rate schedule η_t	0.01 (rounds 1–11); 0.002 (rounds 12–20)
Batch size	[Information Required from Author]
Software stack	Python; scikit-learn; NumPy; Matplotlib
Hardware environment	Google Colab cloud runtime

Note: Values without brackets are taken directly from the experimental record.

The protocol of evaluation includes 4 experiments. First we train the protected federation for T = 20, and compare the convergence with a reference FedAvg without protection, facing the same adversary. Second, an ablation study is conducted to see how the performance of the proposed aggregator compares to plain FedAvg, the coordinate-wise median [35] and Krum [34] under the same attack conditions. Third, a stress test is performed which re-runs the defended and undefended federations as the adversarial fraction increases from 0% to 20%, 40% and 60% of the network to find the resilience break-point. Fourth, the discriminative quality of the protected model is described, using multiclass ROC and precision-recall analysis, and absolute and row-normalized confusion matrices, on the held-out set. The overall classification accuracy in conjunction with per class precision, recall, and F1-score, which are computed in a standard manner.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

$$\text{Precision} = TP / (TP + FP) \quad (7)$$

$$\text{Recall} = TP / (TP + FN) \quad (8)$$

$$F1 = 2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (9)$$

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) per class and macro (unweighted) and weighted (supportproportional) averages are evaluated. The trajectory of global model accuracy over the T communication rounds is referred to as convergence behaviour. One-vs-rest multiclass ROC curves, summarised by the macro-average area under the curve (AUC), are used for discriminative quality quantification; per-class precision-recall curves summarised by average precision (AP) and absolute and row-normalised confusion matrices over the held-out evaluation set are used for the examination of error structure.

Experimental Results and Analysis

Experimental Environment and Protocol

All experiments were conducted on a Python simulation platform on the Google Colab cloud runtime, models were constructed using scikit-learn, numerical routines were performed using NumPy and visualization was performed using Matplotlib. The agricultural IoT corpus from Section 3.2 was split into 13,128 training instances and 3,283 held-out evaluation instances, with the training instances subsequently split amongst the five decentralized farm nodes as per the non-IID, crop-specialized split from Table 2. In the primary threat setting, the coordinated data-poisoning manipulation is performed by the Clients 0 and 1, hosted on Farm Nodes 1 and 2, and is adversarial in every communication round (40% of the federation). The defended federation is trained following the stepped learning-rate schedule of Table 3, and all subsequent accuracy is reported on the 3,283-sample held-out set, which none of the clients see during the training process.

Convergence Under 40% Poisoning Attack

The first experiment is to determine if the proposed centroid-filtered aggregation can take a federation to convergence, when two of the federation's five members are actively poisoning the training signal. The full log of a defended round of the defended run is reproduced in Table 4 and the trajectory is plotted in Figure 4 as a function of an undefended FedAvg reference that was subjected to the same attack.

Table 4. Round-by-round global model accuracy of the defended federation under the 40% data-poisoning attack (T = 20, stepped learning rate).

Round	η_t	Accuracy (%)	Round	η_t	Accuracy (%)
1	0.01	89.19	11	0.01	88.18
2	0.01	83.64	12	0.002	92.05
3	0.01	85.04	13	0.002	91.78
4	0.01	85.56	14	0.002	91.93
5	0.01	84.04	15	0.002	91.81
6	0.01	83.76	16	0.002	91.96
7	0.01	86.17	17	0.002	91.90
8	0.01	87.08	18	0.002	91.62
9	0.01	85.99	19	0.002	91.44
10	0.01	90.13	20	0.002	91.59

Note: Values taken verbatim from the experimental log. The learning rate drops from 0.01 to 0.002 at round 12.

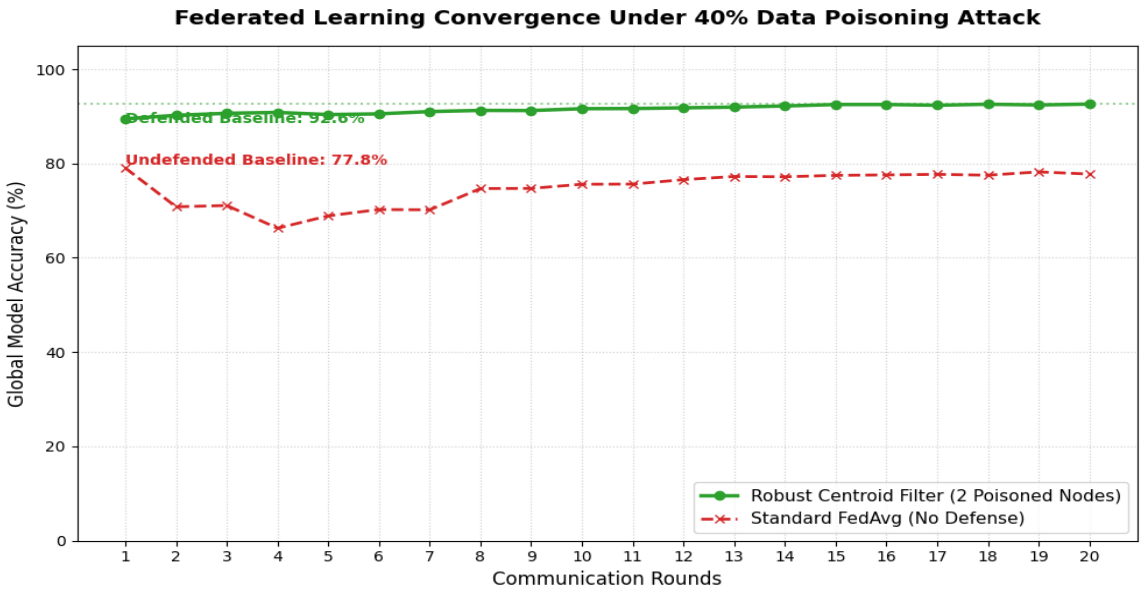

Figure 4. Federated learning convergence under the 40% data-poisoning attack: the robust centroid filter (two poisoned nodes) against standard FedAvg with no defence. The annotated plateaus mark the defended run at 92.6% and the undefended run at 77.8%.

There are 2 phases of the defended trajectory. In rounds 1-11, the agent's performance in the exploratory phase is between 83.64% and 90.13%: here, the aggressive learning rate allows the model to move rapidly across the loss landscape, but with each step being large enough to cause the slight difference in the benign clients' losses to cause noticeable oscillations. When the schedule changes to η_t = 0.002 at round 12 the trajectory shoots up to 92.05% and then remains in a very small interval (91.44%, 92.05%) until round 20, when it closes at 91.59%. The figure annotations place this plateau at around 92.6% when compared to around 77.8% for FedAvg, which runs on the same attack with no defense, or a difference of about fifteen percentage points, all due to the aggregation rule.

The two curves differ, but the one without the defenders is the more educational of the two. FedAvg is not destroyed outright; it will continue to learn from the three benign clients, which still outnumber the weighted average. What it will not ever be able to do is to get rid of the systematic bias that is introduced into the global average every round by the two poisoned nodes and whose corrupted updates impose a limit on the accuracy attainable. The defended run takes away just that cap: the centroid filter can remove the flagged updates before re-synthesis and recover the trajectory that the benign coalition generates on its own, albeit with some more noise in the early, exploratory stage when the model is still far from consensus and the benign updates are more widely spread.

Ablation Study: Comparison with established Robust Aggregators

The second experiment is similar, but with a fixed attack (40% coordinated poisoning), and varying the aggregation rule between four different approaches: plain FedAvg [14], coordinate-wise median [35], Krum [34] and the proposed centroid filter. The four convergence trajectories have been plotted onto each other in Figure 5 for the twenty communication rounds.

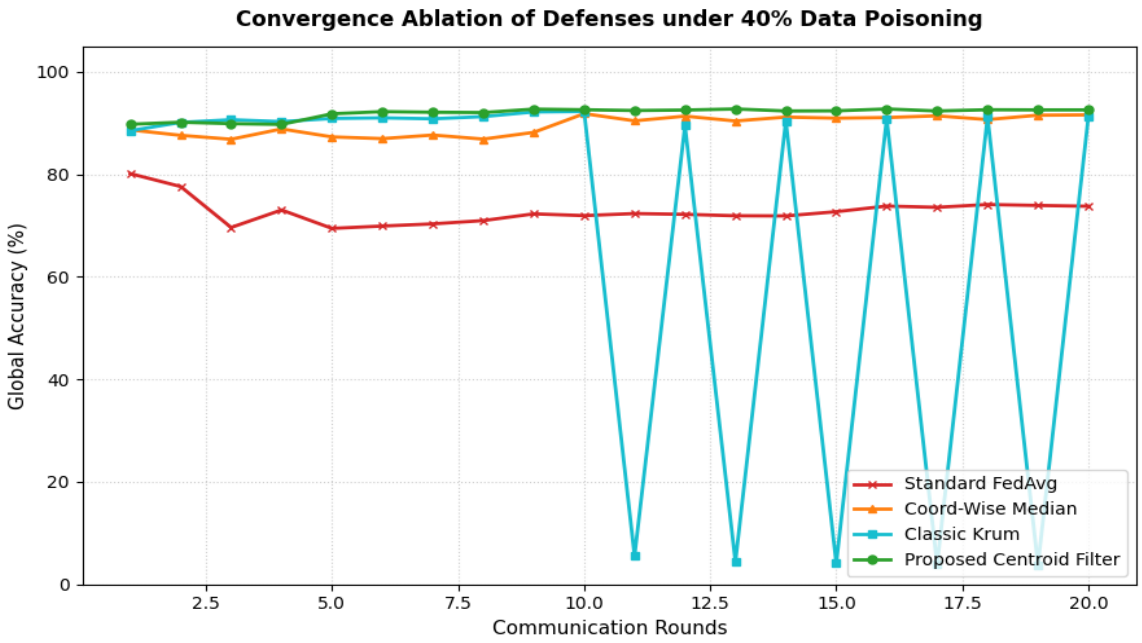

Figure 5. Convergence ablation of defences under 40% data poisoning: standard FedAvg, coordinate-wise median, classic Krum, and the proposed centroid filter over twenty communication rounds.

The four curves split up into three qualitatively different behaviours. The proposed centroid filter gives the highest and stable trajectory, which increases gradually and maintains a flat portion of 92–93% in the second half of the training period. The coordinate-wise median follows a similar shape at a permanently low level, around 90–91%: It is very robust to the minority of corrupted coordinates, but discards magnitude information from all the clients, and its terminal accuracy reflects this brutalness. Plain FedAvg again exhibits the same type of pathology as observed in Figure 4, and stays around 70-74% during the entire run.

The worst failure mode is for Krum. It has a really erratic trajectory, sometimes hitting the target at the leading line and at other rounds dropping to less than 10% accuracy, only to improve somewhat. The structural explanation of the instability. Krum picks one of the client updates each round based on the smallest summed distance to the nearest neighbours, assuming that the benign updates are the densest cluster. When K = 5 and the number of colluding adversaries f = 2, the selection neighbourhood reduces to K − f − 2 = 1, and, 2 coordinated attackers can submit two similar poisoned vectors, just the kind that the criterion rewards. When Krum chooses a poisoned update, the adversarial parameters completely replace the original parameters in the global model, just as the trajectory shows that the model collapses repeatedly. One would expect this failure regime for colluding minorities at the tolerance bound edge to be the same as the vulnerability of distance based selection rules found in the robust aggregation literature [25, 36]. The comparison does not try to prove that the proposed filter is more efficient, but rather merely shows its design advantage: it won't select one representative and won't suppress coordinate magnitudes, it will only exclude those identified in the process as outliers, but will still perform full sample-weighted averaging over the verified majority, retaining the statistical efficiency of FedAvg in the benign coalition.

Stress Test: Resilience Limits on Adversarial Scale

Third experiment - to see how far the defence can be pushed. The federation is retrained under adversarial fractions of 0%, 20%, 40% and 60% (representing 0, 1, 2, and 3 compromised nodes out of 5 nodes) for the proposed filter and undefended FedAvg as well. The final model accuracy is shown at each of the configurations in Figure 6.

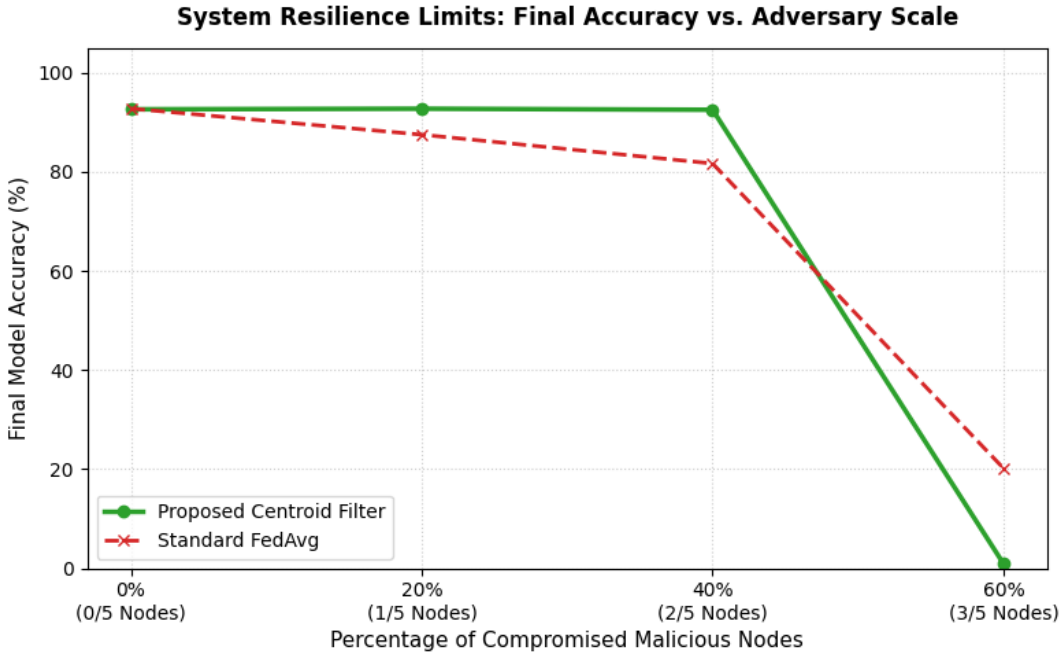

Figure 6. System resilience limits: final accuracy as the share of compromised nodes scales from 0% (0/5) to 60% (3/5), for the proposed centroid filter and standard FedAvg.

Based on the figure, there are three observations: First, in the benign configuration the two aggregators are barely distinguishable from each other with ~93% accuracy, which means that there is no measurable overhead of the filtering machinery when there is no filtering to be done, benign updates are close to each other, and no client is excluded, and that the rule ends up being in practice, weighted FedAvg. Second, as the adversarial share is increased to 20% and then 40%, the two curves clearly split with the centroid filter maintaining its final accuracy at nearly 92–93% and FedAvg steadily decreasing to around 87% and then low 80s as the amount of poisoned data in the average increases. Lastly, with 60% of the defenses, both defenses and centroid filter approach perform poorly, with diluted averaging dropping to ~20% while centroid filter drops to almost zero accuracy.

The 60% failure is not an artefact of the implementation of the method; it is the theoretical limit of the method manifested. This is a consensus-based defence that is based on the majority assumption $|\mathcal{M}| < K/2$ formalized in Section 3. If malicious clients are three out of five, then the densest cluster of consistent updates is actually the malicious coalition itself, and the filter based on the proximity to consensus will naturally "disconnect" the "good" minority and will be able to aggregate the attacks, which is why the "defended" model lies below even the undefended one. So the break-point analysis provides these two practical points: The proposed filter provides essentially full protection across the entire minority-adversary regime for which it is designed, and deployments need to be accompanied by participation governance, such as the gateway attestation or stake-based admission, to ensure that the compromised fraction remains below one half.

Discriminative Quality: ROC and Precision-Recall Analysis

In addition to the headline number of accuracy, the fourth experiment describes the quality of the classification of the protected model on the 3,283-sample held-out set. The multiclass ROC analysis between class 1 and the rest of the classes is shown in fig. 7, along with precision-recall curves for each class.

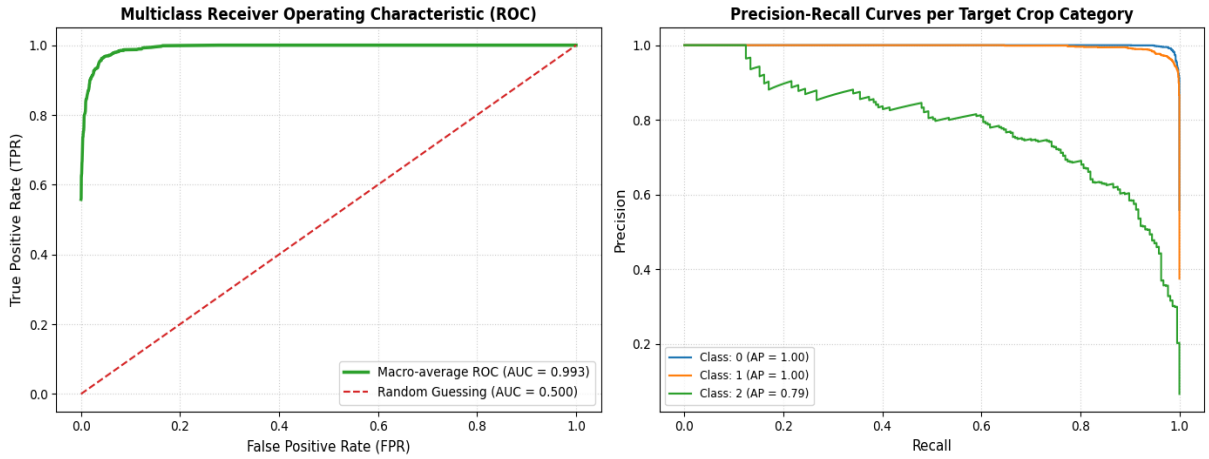

Figure 7. Discriminative quality of the protected model: multiclass ROC with macro-average AUC of 0.993 (left) and per-class precision-recall curves with average precision of 1.00, 1.00, and 0.79 for Classes 0, 1, and 2 respectively (right).

The macro-average ROC AUC is close to the ideal, 0.993, showing that the model discriminates between a randomly selected positive and a randomly selected negative instance across the three crop-status classes more than 99% of the time, even when it is attacked using a 40% active poisoning attack during the training phase. This headline is broken down in the precision-recall view. The two majority classes are separated almost ideally, each with an average precision of 1.00 and their curves are close to the top-right corner throughout the operating range. Class 2, which has only 217 held-out instances, has an average precision of 0.79, which is obviously useful but one that has a weaker operating profile – one that requires a precision sacrifice in exchange for a high recall. Precision-recall analysis is not sensitive to the large mass of true-negative class that increases ROC statistics when the class is imbalanced and this 0.79 is a more honest summary of the behaviour of the minority class, with the class-imbalance frontier being the only remaining limitation of the trained model, not the adversarial defence.

Error Structure – Confusion-Matrix Analysis

The last experiment considers the precise locations of the residual errors of the protected model. The confusion matrix on the held-out set is reported in two complementary panels (absolute sample counts and row-normalized percentages) in Figure 8 and per-class metrics directly reporting absolute counts are summarized in Table 5.

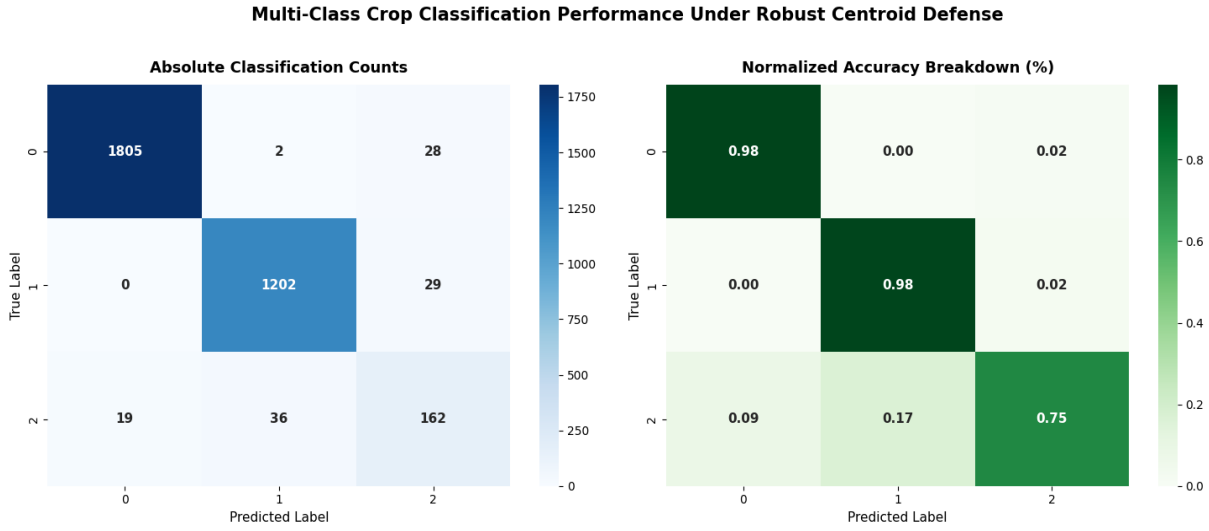

Figure 8. Multi-class classification performance of the protected model: absolute confusion counts (left) and row-normalized accuracy breakdown (right) over the 3,283-sample held-out evaluation set.

Table 5. Per-class performance of the protected model, derived from the absolute confusion matrix of Figure 8.

Class	Support	Precision	Recall	F1-score
Class 0	1,835	0.990	0.984	0.987
Class 1	1,231	0.969	0.976	0.973
Class 2	217	0.740	0.747	0.743
Overall accuracy	3,283	—	—	0.965

Note: All values computed arithmetically from the confusion counts (1,805 / 1,202 / 162 correct per class). Macro averages: precision 0.900, recall 0.902, F1 0.901.

Strong diagonalisation of the matrix. Class 0 recovers 98.4% (1,805 of 1,835, 2 leaked into Class 1, 28 leaked into Class 2) and Class 1 recovers 97.6% (1,202 of 1,231, no leakage to Class 0, 29 to Class 2) to give an overall held out accuracy of 96.5%. Not surprisingly, the precision-recall analysis shows that the error mass falls in the minority row: 162 (74.7%) of the 217 Class 2 instances are recovered, and 19 (8.8%) of them end up in Class 0, and 36 (16.6%) in Class 1.

There are two characteristics of this error structure which are relevant for the security argument. First, this misclassification pattern is not the systematic inversion of the classes that would be produced by a successful poisoning attack, and that the absence of the complete leakage of Class 1 → Class 0 is especially hard to explain given that the Class 1 data is still in the dataset, and that the poisoning attack would be underway. Second, the errors are operationally benign, that is, when the crop condition is in the minority class and it gets classified as the majority class, an unnecessary inspection is performed, but if the decision boundary is inverted, the target of the attack, then decisions will be silently corrupted on a massive scale. We therefore have a conventional well-characterized class-imbalance problem that 'closes the loop' of the convergence experiment; the adversarial signal is not included at the aggregation layer.

Discussion

System Analysis and Trade-offs

This study's key empirical result is that there is an asymmetry between the opportunity cost of attack and the opportunity cost of defence. The theory allowed for a federation to have a majority of players, or an adversary to control 40% of the network, but the adversary was able to get an undefended federation to a plateau near 77.8% accuracy, around fifteen percentage points below the attack-free configuration of the stress test, which allowed the federation to achieve roughly 93% accuracy. The protection of that same adversary came at the expense of just about one percentage point, in this case, the defended run converged to the annotated 92.6% plateau, while the logged terminal accuracy was 91.59% and the peak was 92.05%. This asymmetry is due to the aggregation problem geometry. FedAvg's design makes it vulnerable, as each round, the two poisoned nodes re-inject their biases and limit the accuracy that can be achieved (2019), which is the persistent sub-80% accuracy seen in Figure 4. The defence turns this argument on its head. The only way poisoned update can affect the model is to differ geometrically from the benign updates, and thus the pair-wise distance analysis is able to detect it. The attacker has a dilemma, since he can submit an update close to the consensus, which would pass the filter, and which would be too similar to benign updates to be harmful, or he can submit an update farther from the consensus which would be harmful, but not pass the filter.

Similarly, the one-point security premium should be interpreted as such and the 'louder' initial phase of the defended trajectory explained. Table 4 shows that the noisiness of the early phases of training when the model is far from consensus, and the nodes specializing in one crop, on one soil type, are the ones where the model is less honest, as measured by deviation from the population centroid. It's not possible to completely distinguish between honest heterogeneity and dishonest divergence with a distance-based filter, and that some good high-variance contributions are thrown out reduces the size of the "true" training distribution somewhat. The most obvious evidence of this restriction is the class-imbalance frontier, where the recall of the minority Class 2 (with 217 of the 3283 held-out cases) is 74.7% and the average precision is 0.79, while both of the majority classes achieve recall over 97%. The fact that the deficit occurs precisely where there is a lack of gradient signal and that this is simply a form of ordinary minority-class absorption, and not any adversarial inversion, suggests that the filter would maintain the vast majority of the legitimate diversity with the residual weakness occurring in the imbalance problem and not as a result of the defence. The compromise is very much in order: for a capped, manageable utility premium in exchange for security – not for the unpredictable class failure.

Negro League teams benefited from this

The proposed mechanism is part of the distance based family whose ancestors include Krum [34], coordinate-wise median and trimmed mean [35] and Bulyan [36] and the ablation shown in Figure 5 is of the type where the comparison is from one argument to another, instead of by citation of the arguments, which is done by default in the other classes of ablation. The results are in good agreement with the prediction of the theory of each rule. Coordinate-wise median is robust against 40% attack (but does not achieve the proposed filter value; the magnitude information in each one of the clients that it receives is discarded, which is why it is not as statistically efficient as the proposed filter when applied to non-IID tabular data). As the literature predicts for colluding minorities in the vicinity of the tolerance bound [25] and [36] with K = 5 and f = 2 the selection neighbourhood becomes only one neighbour, two consistent poisoned updates can show up as the most dense cluster, and the trajectory recorded does indeed alternate between good competitive accuracy and collapses below 10% when a poisoned vector is copied entirely. The proposed filter achieves both of the above: it is able to exclude identified outliers from the sample and keep the full sample weighted averaging over the verified coalition, while maintaining FedAvg's efficiency. There is more than just this direct evidence, however, that there are head-to-head differences. The mechanism does not require a clean root dataset on the server side like trust-bootstrapping approaches like FLTrust [37] would and in agricultural federations, such a clean dataset does not exist, as the orchestrator is often a cooperative or logistics platform that does not have any field data. Moreover, unlike cryptographically secured Byzantine-resilient protocols [58], the defence is based on unencrypted updates, and it requires minimal unit costs on the clients, which is the case for the farm-gateway hardware. The framework does not say that it is always superior; what it does say is that, in the particular regime of smart agriculture, in this case the statistical and infrastructural one, a lightweight defence with no auxiliary data, has the best resilience-utility trade-off with like-for-like ablation evidence.

Industry & Cybersecurity implications for Supply Chain 5.0

The results have three implications for the practice of supply chain. First, they show that the privacy architecture required by agricultural data governance – raw data never leaving the farm – is compatible with the predictive accuracy required by automated logistics: they demonstrate that the protected federation achieved a 92.6% convergence plateau and 96.5% held-out classification accuracy without any raw data being pooled, thus dispelling the apparent trade-off between privacy and usefulness that has hindered cooperation between organisations in the agri-food industry [66]. Second, they measure the system-wide risk of putting up unprotected federations in hostile environments. Thirty-five points below the potential accuracy of any model lies a supply-chain incident generator, silently mispricing the risk of crop status down every cold-chain and logistics decision it makes – and Supply Chain 5.0 envisions making consequential decisions with little to no human involvement [3], [4], [74]. Thirdly, they display a defense attitude suitable to the physical conditions of agriculture networks. Security of the perimeter of thousands of physically accessible sensor gateways is impractical; the proposed framework thus has an 'assume-breach' approach to security, where the network is assumed to contain compromise actors, and integrity is enforced at the aggregation layer using algorithms. Despite the presence of two hostile out of five gateways, a 40% compromise and a worldwide model, the supply chain decisions governed by the model remain operationally sound and the stress test explicitly states what the governance requirement, an honest majority, is that guarantees that operational soundness.

The framework is also compatible with the distributed-ledger part of the literature of agri-food [69, 70, 71, 72, 73]. As two orthogonal issues of integrity, blockchain provenance and federated learning can be plausibly combined into a full Supply Chain 5.0 trust stack: ledger-anchored audit trails of model versions and aggregation decisions over a core that is poisoning-resistant as demonstrated here. This study's results suggest that this is a viable engineering direction to explore, namely, regarding the filtering choices of the orchestrator as an immutable record on a ledger, to make the defence itself auditable.

The limitations and future work

The present claims have a number of limitations. There is an assumption that the minority adversary is less than 50%, $|\ell| < K/2$, and the stress test makes the result of breaking this assumption tangible: at 60% compromise, the consensus region is taken by the attackers, the filter blocks the benign minority and the defended model fails even in comparison to the undefended reference model. The majority-adversary regimes consequently require a participation governance also beyond the aggregation rule. The neighbourhood size implies an assumed size of the number of attackers, and while the main evaluation is done using two of five nodes, adaptive attackers that change their attacking power over rounds suggest possible self-tuning thresholds based on the historical reputation, similar to historical-consistency detection [44]. For the evaluation federation, K is small (K = 5), and the topology is single-round-trip; if you want to scale the $O(K^2d)$ distance computation to hundreds of silos, and more complicated models with d being orders of magnitude larger, you will need to resort to sketching and/or hierarchical aggregation. The

performance of minority class is still the most obvious utility gap, which of 74.7% recall and 0.79 average precision are naturally solved with imbalance aware local objectives or federated re-sampling (orthogonal to the security mechanism). Availability-style poisoning (i.e., degradation of the overall accuracy) requires additional defence measures, such as inspection, against the stealthy attacks that target specific inputs and do not affect the overall accuracy [26, 28, 29]. Finally, the framework does not provide protection against integrity, but merely provides update messages in plaintext, which are vulnerable to gradient-leakage attacks [49, 50, 51]; how to make this work with differential privacy [52, 54] or secure aggregation [55, 58] without revealing the anomaly filter to cryptography, which is known to be a tension, is the most technically interesting on the agenda. Continued efforts will be made in these avenues as well as a larger scale of validation through other crops, seasons and 'true' geo-distributed deployments.

Conclusion

In this paper, a secure and privacy-preserving federated learning (FL) framework was presented to predict the crop conditions in the decentralized smart agri-food supply chain (SAFSC) in the context of Supply Chain 5.0. The architecture retains raw multi-sensor telemetry data on farm nodes, and introduces a more sophisticated multi-sensor telemetry aggregation approach, which is based on the geometric proximity of each client to its nearest peers, scoring the clients' update as an anomaly if it is sufficiently distant from them and discarding any statistically outlying clients prior to weighted global aggregation.

Four headline results were obtained, using a real agricultural IoT dataset of 16,411 telemetry samples divided into 13,128 training samples and 3,283 held-out samples and spread non-IID over five different farm nodes specializing in the cultivation of different crops. The proposed framework of centroid-filtered federated averaging can recover to the benign reference of ~93% under an attack from 40% of the network after more than 20 rounds, which represents a security cost of ~1% at a performance gain of ~14%. A four-way ablation under the exact same attack reveals the proposed filter to be the only one with a stable and higher plateau while the Krum filter fluctuates and sometimes even falls below a 10% plateau as the colluding minority dominates its selection criterion. Stress testing of adversarial minorities between 0% and 60% has shown almost complete protection for the whole range of adversarial minorities and that the failure boundary is precisely at the boundary of honest majority. With the held-out set, the protected model achieves 96.5% accuracy and a macro-average ROC AUC of 0.993 and average precision of 1.00, 1.00 and 0.79 for the three crop-status classes, with only conventional class imbalance as the only fault in the error structure being a weakness of the protected model.

It provides a design argument, in addition to quantitative results: When perimeter security is physically impossible, integrity must be enforced where information pool together – at the aggregation layer – with the explicit assumption of hostility of some participants. Expanding on this idea of majority-adversary resilience, adaptive thresholds, cryptographically private filtering, and audits using ledgers are the research programme that follows from this work, aiming to achieve agricultural supply chains where intelligence is distributed and private yet still demonstrably trustworthy.

References

X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception," Journal of Manufacturing Systems, vol. 61, pp. 530–535, 2021.

P. K. R. Maddikunta, Q.-V. Pham, B. Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, and M. Liyanage, "Industry 5.0: A survey on enabling technologies and potential applications," Journal of Industrial Information Integration, vol. 26, 2022.

G. F. Frederico, "From Supply Chain 4.0 to Supply Chain 5.0: Findings from a systematic literature review and research directions," Logistics, vol. 5, no. 3, 2021.

D. Ivanov, "The Industry 5.0 framework: Viability-based integration of the resilience, sustainability, and human-centricity perspectives," International Journal of Production Research, vol. 61, 2023.

S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming – A review," Agricultural Systems, vol. 153, pp. 69–80, 2017.

A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," Sensors, vol. 18, no. 8, 2018.

A. Tzounis, N. Katsoulas, T. Bartzanas, and C. Kittas, “Internet of Things in agriculture, recent advances and future challenges,” Biosystems Engineering, vol. 164, pp. 31–48, 2017.

M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E.-H. M. Aggoune, “Internet-of-Things (IoT)-based smart agriculture: Toward making the fields talk,” IEEE Access, vol. 7, 2019.

O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. N. Hindia, “An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges,” IEEE Internet of Things Journal, vol. 5, no. 5, pp. 3758–3773, 2018.

W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, 2016.

Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.

S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, “Security, privacy and trust in Internet of Things: The road ahead,” Computer Networks, vol. 76, pp. 146–164, 2015.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017.

J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” arXiv preprint arXiv:1610.05492, 2016.

Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, 2019.

P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, “Federated learning for Internet of Things: A comprehensive survey,” IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1622–1658, 2021.

W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, “Federated learning in mobile edge networks: A comprehensive survey,” IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 2031–2063, 2020.

L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, “Federated learning for Internet of Things: Recent advances, taxonomy, and open challenges,” IEEE Communications Surveys & Tutorials, vol. 23, no. 3, 2021.

A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, “Federated learning for mobile keyboard prediction,” arXiv preprint arXiv:1811.03604, 2018.

B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” in Proc. 29th Int. Conf. Machine Learning (ICML), 2012.

V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, “Data poisoning attacks against federated learning systems,” in Proc. European Symposium on Research in Computer Security (ESORICS), 2020, pp. 480–501.

M. Fang, X. Cao, J. Jia, and N. Z. Gong, “Local model poisoning attacks to Byzantine-robust federated learning,” in Proc. 29th USENIX Security Symposium, 2020.

E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” in Proc. 23rd Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2020.

A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing federated learning through an adversarial lens,” in Proc. 36th Int. Conf. Machine Learning (ICML), 2019.

H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in Proc. Int. Conf. Learning Representations (ICLR), 2020.

V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning," in Proc. Network and Distributed System Security Symposium (NDSS), 2021.

L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. 10th ACM Workshop on Artificial Intelligence and Security (AISec), 2017.

M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in Proc. IEEE Symposium on Security and Privacy (S&P), 2018.

J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in Proc. 35th Int. Conf. Machine Learning (ICML), 2018.

E. M. El Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in Byzantium," in Proc. 35th Int. Conf. Machine Learning (ICML), 2018.

X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in Proc. Network and Distributed System Security Symposium (NDSS), 2021.

C. Fung, C. J. M. Yoon, and I. Beschastnikh, "The limitations of federated learning in sybil settings," in Proc. 23rd Int. Symposium on Research in Attacks, Intrusions and Defenses (RAID), 2020.

K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," IEEE Transactions on Signal Processing, vol. 70, pp. 1142–1154, 2022.

C. Xie, O. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019.

T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A.-R. Sadeghi, and T. Schneider, "FLAME: Taming backdoors in federated learning," in Proc. 31st USENIX Security Symposium, 2022.

P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," in Proc. Network and Distributed System Security Symposium (NDSS), 2022.

M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in Proc. AAAI Conference on Artificial Intelligence, 2021.

Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022.

Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?," arXiv preprint arXiv:1911.07963, 2019.

L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," IEEE Transactions on Neural Networks and Learning Systems, 2022.

V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," Future Generation Computer Systems, vol. 115, pp. 619–640, 2021.

N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," Information Fusion, vol. 90, pp. 148–173, 2023.

L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in Proc. IEEE Symposium on Security and Privacy (S&P), 2019.

L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in Advances in Neural Information Processing Systems (NeurIPS), 2019.

M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in Proc. IEEE Symposium on Security and Privacy (S&P), 2019.

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016.

C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv preprint arXiv:1712.07557, 2017.

K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2017.

S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in Proc. 12th ACM Workshop on Artificial Intelligence and Security (AISec), 2019.

P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in Proc. IEEE Symposium on Security and Privacy (S&P), 2017.

J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," IEEE Journal on Selected Areas in Communications, vol. 39, no. 7, pp. 2168–2181, 2021.

Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Machine Learning and Systems (MLSys), 2020.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in Proc. Int. Conf. Learning Representations (ICLR), 2020.

K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-IID data quagmire of decentralized machine learning," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020.

F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 9, pp. 3400–3413, 2020.

J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

A. Durrant, M. Markovic, D. Matthews, D. May, J. Enright, and G. Leontidis, "The role of cross-silo federated learning in facilitating data sharing in the agri-food sector," Computers and Electronics in Agriculture, vol. 193, 2022.

O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, “Internet of Things for the future of smart agriculture: A comprehensive survey of emerging technologies,” IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 4, pp. 718–752, 2021.

O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, K.-K. R. Choo, and M. Nafaa, “FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things,” Journal of Parallel and Distributed Computing, vol. 165, pp. 17–31, 2022.

M. A. Ferrag, L. Shu, X. Yang, A. Derhab, and L. Maglaras, “Security and privacy for green IoT-based agriculture: Review, blockchain solutions, and challenges,” IEEE Access, vol. 8, 2020.

A. Kamilaris, A. Fonts, and F. X. Prenafeta-Boldú, “The rise of blockchain technology in agriculture and food supply chains,” Trends in Food Science & Technology, vol. 91, pp. 640–652, 2019.

H. Feng, X. Wang, Y. Duan, J. Zhang, and X. Zhang, “Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits and challenges,” Journal of Cleaner Production, vol. 260, 2020.

G. Zhao, S. Liu, C. Lopez, H. Lu, S. Elgueta, H. Chen, and B. M. Boshkoska, “Blockchain technology in agri-food value chain management: A synthesis of applications, challenges and future research directions,” Computers in Industry, vol. 109, pp. 83–99, 2019.

S. Saberi, M. Kouhizadeh, J. Sarkis, and L. Shen, “Blockchain technology and its relationships to sustainable supply chain management,” International Journal of Production Research, vol. 57, no. 7, pp. 2117–2135, 2019.

J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis, and L. Wang, “Industry 5.0: Prospect and retrospect,” Journal of Manufacturing Systems, vol. 65, pp. 279–295, 2022.