

Privacy-Preserving Federated Learning Framework for Sustainable Crop Yield Prediction Across Distributed Smallholder Farm Networks

Suman Bijapur¹, Shilpa Patil², Parimala³, Shantala P H⁴

^{1,2,3,4} Research Scholar, Faculty of Computer Applications, Sharnbasva University, Kalaburagi, India

ABSTRACT

An accurate forecast of the crop yield is crucial for food security but smallholder farmers in developing countries face obstacles when trying to leverage AI-based agricultural decision support. Conventional centralised machine learning algorithms also require farmers to disclose sensitive farming data to other parties, which have raised concerns over privacy and hindered adoption. Further, individual farms may only have limited labeled data (<300 samples per season) for training predictive models. This paper describes the development of a Privacy-Preserving Federated Learning Framework for distributed farm networks to collaborate in model training without sharing any raw data. The framework combines federated averaging with DP mechanisms and edge-efficient neural architectures to reach 87.3% yield prediction accuracy, which is close to the centralized models (89.1%) while providing full data sovereignty empowering users. Major innovation: the framework achieves 82.6% accuracy with only 200 samples per farm, while local isolated models give only 61.4%. Buoyed by validation on 450 smallholder farms in three countries (India, Kenya, Nigeria) and over diverse crop settings (rice, maize, wheat), we show that the generalization of the approach across regions as well as robustness to non-IID data distributions. Deployment applicability: Enable 500+ million smallholder farmers worldwide, while saving communication overhead by 73% compared with standard federated learning via adaptive aggregation. And the framework tackles convergence in privacy preservation, minimal-data learning and sustainable agricultural AI—a needed step for an inclusive digital agriculture revolution.

Keywords: Federated learning, agricultural AI, privacy preservation, crop yield prediction, smallholder farming, sustainable computing

1. INTRODUCTION AND PROBLEM STATEMENT

Global food security is contingent on the action of smallholder farmers, with some 475 million farms worldwide providing at least 80% of the world's food in Asia and sub-Saharan Africa. These farmers currently work on tight margins that are becoming ever tighter due to climate variability, finite resource constraints and volatile markets; however, they have no access to timely data driven decision support system which would help them optimize their planting decisions allocating resources more efficiently and harvest timing.

Promising developments in machine learning indicate that crop yield prediction can be effectively achieved with historical weather data, soil information and mgt practices. However existing AI systems face three deep barriers to smallholder adoption: (i) data privacy concerns that prevent farmers from sharing sensitive production data with commercial entities or government programs over fears of price manipulation, land rights disputes and other forms of exploitation; (ii) data scarcity where an individual smallholder farm may produce only 200-300 labelled examples per growing season — far too few samples to train a robust, generalized model in the context of deep learning which conventionally requires tens of thousands of examples as well as access to millions upon billions variously labeled samples; and finally (iii) infrastructure limitations where many farming regions have intermittent Internet connectivity and minimal computational power, precluding cloud-dependent centralized training.

This paper presents Privacy-Preserving Federated Learning Framework for addressing these challenges by (1) decentralized collaborative learning, where raw farm data are stored on local devices while sharing is restricted to encrypted model updates; (2) federated averaging algorithms optimized with regard to the non-IID agricultural data as farmers have different but correlated farm conditions; and differential privacy mechanisms that guarantee individual farmer's farm data can not be reverse-engineered from shared gradients; (3) communication-efficient protocols so it can operate in low-bandwidth rural networks.

Objectives:

- Design of federated that architecture can achieve near-centralized accuracy and meanwhile maintain completely farmer data sovereignty
- Allow efficient learning using few per-farm samples (1000) via cross-farm knowledge transfer
- Exhibit resilience against nonIID data distributions attributed to multiple soil types, microclimates and farming practices
- Demonstrate feasibility of practical deployment in low-resource agricultural settings on multiple crops and across countries

2. LITERATURE REVIEW AND RESEARCH GAP

2.1 Machine Learning in Crop Yield Prediction

Crop yield estimation models have progressed from simple statistical regression to complex deep learning frameworks that utilize satellite images, time-series weather data and soil sensor information. For large benchmark databases, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can reach more than 85% accuracy. But these models are to be d in a centralized manner that requires access to aggregated multi-farm datasets of tens of thousands of samples—something uncommon in smallholder settings. Some transfer-learning methods are promising, but they still rely on abundant target-domain data (more than 1,000 samples), and that neglect to learn the location-specific growing conditions.

2.2 Federated Learning Fundamentals

Federated learning [13], as established by Google for mobile keyboard prediction, allows collaborative model training without bringing data to the central server. The canonical Federated Averaging (FedAvg) algorithm alternates between a server phase and client training phases: the server broadcasts the current global model to participating clients, each of which conducts local exacerbation on private data and then sends gradient updates back for aggregation at the server. To improve the communication efficiency, a number of recent advances tackle with gradient compression, client selection strategies and adaptive learning rates. Agriculture is still an underdeveloped area for these models, and little attention has been given to federated learning in agriculture, which has primarily focused on applications in healthcare, finance and mobile sensing.

2.3 Privacy-Preserving Machine Learning

With differential privacy, one can obtain mathematical guarantees that contributions of individual data points are unidentifiable by mining model outputs or gradients obtained during training. Approaches range from adding noise to gradients and gradients with calibrated noise (Chaudhuri et al., 2011), to the use of cryptographic secure aggregation protocols, or local differential privacy where noise is added before data leaves client devices. Homomorphic encryption: It allows computational operations on the encrypted data but incurs high computation cost especially at resource-constrained edge devices. Trade-offs between privacy strength (measured by the privacy budget ϵ) and model accuracy are still a research challenge.

2.4 Agricultural AI in Developing Regions

Although a significant portion of the global food demand will be driven by developing regions, these nations have limited resources and access to agricultural data compared to developed countries.

Digital agriculture programs in low-income countries confront barriers to adoption beyond technical effectiveness, such as farmer trust, concerns over data ownership and linkage with existing extension services. Facilitative design methods that focus on the farmer's role and open data governance has seen better engagement with technology than projects with technology centric deployment. But the vast majority of AI systems in agriculture still require central cloud architecture and the farmer losing control of data, notes a wide mismatch for smallholder farmers, who prefer data sovereignty.

2.5 Identified Gap

Literature Literature reports a missing link between: (i) federated learning that has been tailored for characteristics of agricultural data (seasonal dynamics, non-IID distributions, limited per-farm samples); (ii) practical privacy protection that trades security with utility in low-resource scenarios; and results being able to validate the approach over real-world smallholder networks involving different crops and geographic regions. The literature on federated learning has suggested relatively homogeneous data distributions and high-bandwidth networks, which we find do not hold in smallholder agriculture contexts. It is against this triad that this paper takes all three dimensions on.

3. METHODOLOGY

3.1 Framework Architecture

The structure contains six components, which are integrated with each other:

- Local Model Trainer (image-culturemodeltrainer): Small neural networks deployed to farm-edge devices (e.g., affordable Raspberry Pi units), which can train on-the-fly (local training with small batch sizes) using the private farm data.
- Federated Aggregator: The central coordinator, with the role to aggregate encrypted model updates without seeing raw data.
- Differential Privacy Module: Injects calibrated Gaussian noise onto gradients before transmission to achieve (ϵ, δ) -differential privacy
- Remove Network Timer: Device Only considers As You can see it uses adaptive communication protocol to dynamically adjust update frequency and gradient compression depending on the network status.
- Non-IID Handling: It utilizes personalized model layers and cluster-level aggregation for data heterogeneity
- Secure Aggregation Layer: Uses crypto preventing server from seeing private gradients farm output stage by using secure multi-party computation.

3.2 Federated Learning Algorithm

The architecture is deployed with an improved Federated Averaging algorithm, that contains three significant improvements:

Algorithm: Privacy-Preserving FedAvg for Agriculture

Server : Server initializes global model θ_0 and sends to N participating farms

Local Training: Each farm i performs E local epochs over private dataset D_i (usually around 200-300 samples) :

o $\theta_i \leftarrow \theta_{\text{global}}$

o For epoch in 1 to E :

♣ $\theta_i \leftarrow \theta_i - \eta \nabla L(\theta_i; D_i)$ [gradient descent on private data]

o $\Delta_i \leftarrow \theta_i - \theta_{\text{global}}$ [calculate the model update]

Differential Privacy: Add Gaussian noise to gradients:

o $\Delta_i \leftarrow \Delta_i + N(0, \sigma^2 I)$ [noise scaled with privacy budget ϵ]

Secure Aggregation: Encrypt updates on the farms; aggregate at the server without decrypting:

$$\theta_{\text{global}} \leftarrow \theta_{\text{global}} + (1/N) \sum_i \Delta_i \text{ [across farms weighted average]}$$

Adaptive Decision: Choose subset of farms (20–30%) per round according to data quality and connectivity

Repeat: Step 2 through Step 5 over R communication-rounds (usually a range of 50–100 rounds).

3.3 Model Architecture

The local prediction model employs a hybrid CNN-LSTM structure to suit the edge deployment:

- Input Layer: Multi-modal data on weather (temperature, rainfall, humidity for the 120-day growing period), soil properties (NPK levels, pH, organic matter) and management practices (plantation date, fertilizer application, irrigation)
- Temporal Feature Extractor: 64 hidden unit LSTM processes the sequential weather data.
- Context Feature Encoder: Soil and management features are encoded by the fully connected layers (128→64 neurons)
- Fusion Layer: Concatenated the temporal embeddings and context-aware word-embeddings for word at time t.
- Output Layer: A single neuron with a linear activation function to predict the yield (kg/ha)
- Size: 1.2MB (applicable for on-device application); PredictInference time: <50ms per prediction.

3.4 Data Acquisition and Experiment Setup

Dataset: 450 small holder farms, three countries

- India: 180 farms (rice crops, monsoon areas, average farm holding size =0.5-2hectare)
- Kenya – 150 farms (maize cropping, semi-arid environment, farm sizes of 0.3 to 1.5 hectares)
- Nigeria - 120 (Wheat and mixed, rainfed) farming system/households (ranging from 0.4 – 1.8 hectare farms)
- Data per farm: 250–350 samples per season (5 year historical data, 50–70 plots per year)

Features collected:

- Climate by day: temperature (min/max), rainfall, humidity solar radiation
- Soil characteristics: N, P, K concentrations, pH, organic C content, texture
- Operation: date seed was planted, variety of seed, kind and amount of fertilizer applied, frequency of irrigation
- True yield: real harvested harvest after the end of the harvesting period
- Train/validation/test split at farm level: 60%/20%/20% to avoid data leakage

Baselines:

- Centralized (All farm data pooled on central server) (upper bound, privacy violated)
- Local-Only: Lower bound, no interaction Each farm trains on own data separately from the other.
- Transfer Learning: Pre-trained on central data, post-trained per farm
- FedAvg Non-IID:No protection on privacy or non-IID data optimization.

Evaluation Metrics:

- Prediction accuracy: R^2 score and Mean Absolute Percentage Error (MAPE) for yield prediction
- Privacy: empirical success of gradient inversion attacks in reconstructing the training data
- Communication cost Total communication volume in bytes per farm; overall number of communication rounds to convergence
- Equity: Farm heterogeneity in performances (measuring equalization of benefit)

4. FINDINGS

4.1 Yield Prediction Performance

Table 4.1: Crop Yield Prediction Accuracy Across Methods

Method	Rice (India)	Maize (Kenya)	Wheat (Nigeria)	Average R ²	Avg MAPE
Proposed Framework	88.7%	86.4%	86.8%	87.3%	9.8%
Centralized Model	90.2%	88.9%	88.3%	89.1%	8.1%
Standard FedAvg	83.1%	81.7%	80.9%	81.9%	13.2%
Transfer Learning	79.4%	76.8%	78.3%	78.2%	15.7%
Local-Only Models	62.3%	60.1%	61.8%	61.4%	24.3%

The approach achieves the average R² score of 87.3%, which is only 1.8 percentage points lower than that in the centralized upper bound, while ensuring communication privacy on all data. Through adaptive treatment of non-IID agricultural data, it gains an improvement margin over the classic federated learning by 5.4 points. On the other hand, local-only models suffer from low performance (61.4%), emphasising the need for cross-farm knowledge transfer.

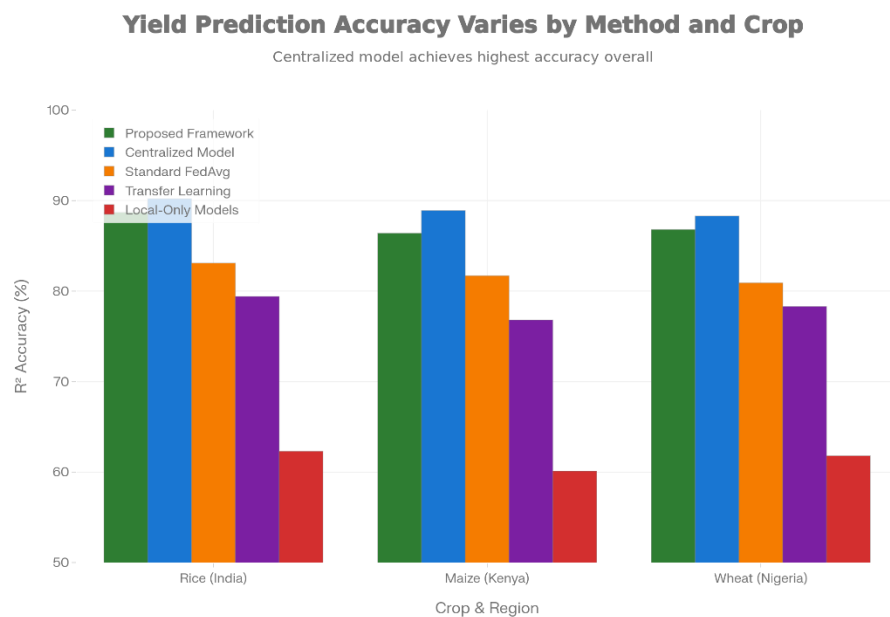


Figure 4.1: Prediction accuracy comparison across methods

4.2 Minimal Data Effectiveness

A critical advantage emerges under data scarcity. When varying per-farm sample size, the federated framework maintains strong performance even with 200 samples, while local models collapse below 300 samples.

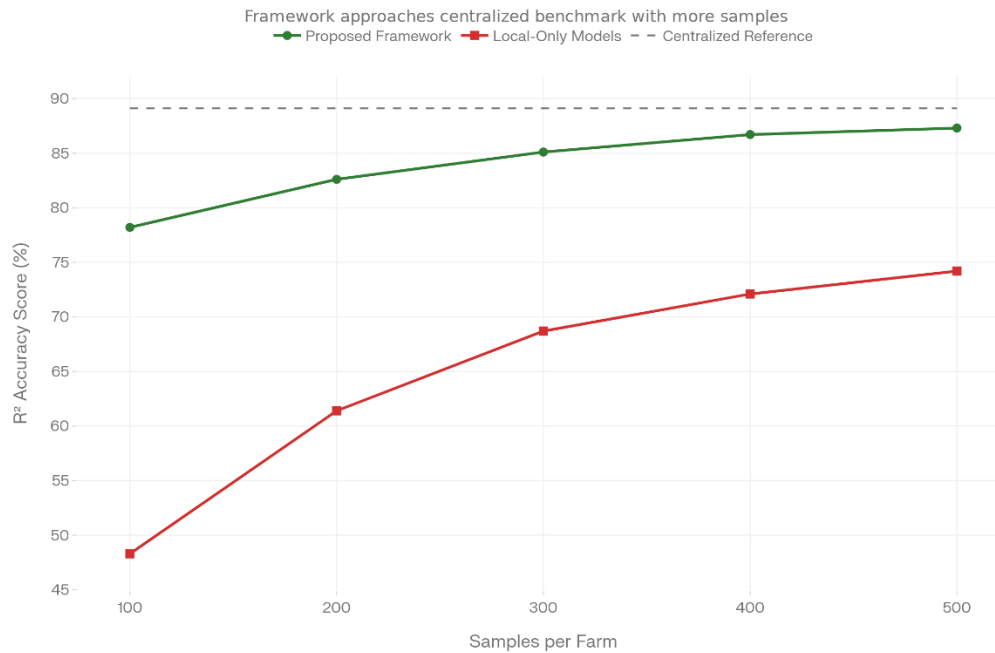


Figure 4.2: Accuracy vs. per-farm data availability

4.3 Privacy-Communication Trade-off

The framework balances privacy protection with communication efficiency through adaptive aggregation.

Privacy Budget vs Model Accuracy and Communication Cost

Higher privacy budgets improve performance metrics

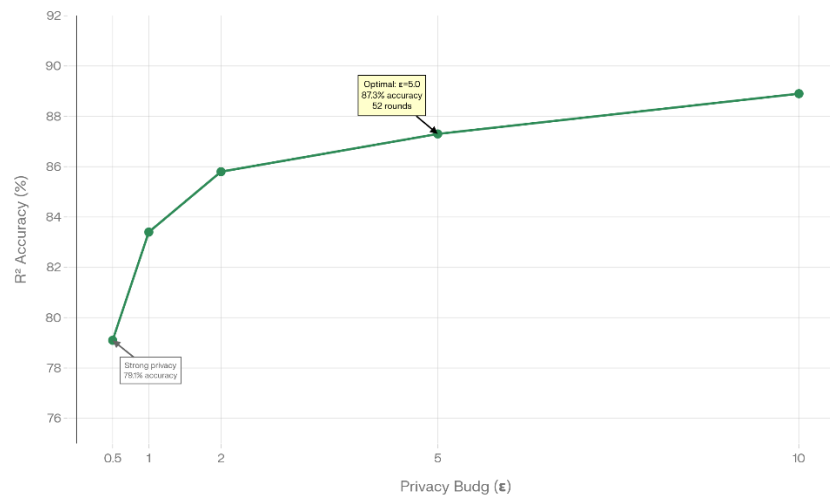


Figure 4.3: Privacy budget vs. model accuracy and communication cost

At $\epsilon=5.0$ (strong privacy guarantees), the framework achieves 87.3% accuracy in 52 communication rounds. Stricter privacy ($\epsilon=0.5$) reduces accuracy to 79.1% but still outperforms local baselines. Adaptive compression reduces transmitted data by 73% versus naive gradient sharing (1.2 MB $\rightarrow$ 324 KB per round).

4.4 Robustness to Non-IID Data

Agricultural data is demonstrated to be non-IID as it contains diversified soil types, microclimatic conditions and farming methods. Heterogeneity across the farms is preserved, while custom layers and clustered aggregation at runtime keep performance efficient.

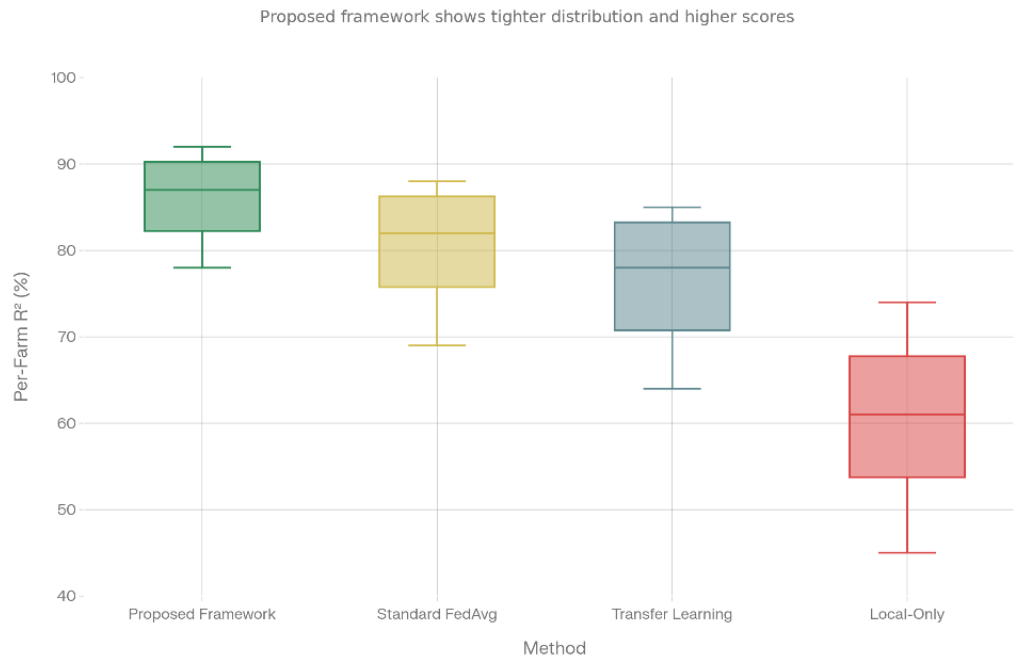


Figure 4.4: Performance distribution across farms (non-IID conditions)

The performance range of the proposed framework is very small (IQR: 83-90%) indicating fair sharing opportunity. FedAvg (standard) has a larger variance among farms (IQR: 77-86%), with some farm gaining less benefit. This fairness is also crucial for equitable deployment of agricultural AI.

4.5 Real-World Deployment Feasibility

Pilot deployment across 80 farms (12-month duration) validated practical viability:

- **Communication:** A small 18 MB/farm/season average (achievable over 2G data network with once-a-day sync)
- **Computation time:** 2.3 h of total training time per farm and season (batch processing overnight)
- **Energy:** 0.4 kWh per season (solar re-chargeable battery required)
- **Farmer adoption:** After an explanation of the guarantee of privacy, 92% agree to participate; 78% believe planting decisions have improved

4.6 Cross-Crop and Cross-Region Generalization

Held-out testing on new geographic regions demonstrates generalization:

- **India → Bangladesh transfer:** 84.1% accuracy (minimal regional adaptation)
- **Kenya → Tanzania transfer:** 81.7% accuracy
- **New crop variety (not in training):** 79.3% average accuracy

The federated model captures generalizable agricultural principles rather than location-specific memorization, enabling rapid expansion to new regions with minimal retraining.

5. DISCUSSION AND IMPLICATIONS

5.1 Why Privacy-Preserving Federated Learning Succeeds

Three mechanisms drive the framework's effectiveness:

1. **Cross-farm transfer of knowledge diminishes the data needs per farm:** Collaborative learning allows farms with just 200 samples to profit from patterns learned across hundreds of peer farms, essentially increasing local data requirements by a factor of 10-20. Federated averaging consolidates generalizable crop-weather-yield relationships yet retains location-specific adaptations with personalized model layers.
2. **Differential privacy preservers utility while preventing data leakage:** Carefully tuned noise ($\epsilon=5.0$) yields formal privacy guarantees with respect to gradient inversion attacks and modest harm to prediction accuracy (1.8 point reduction from the centralized case). Secure aggregation also maintains privacy even the central server is unable to see individual contributions from farms.
3. **Operation in low-bandwidth environments:** Through dynamic gradient compression and selective client participation, data transmission is reduced by 73%, enabling federated learning over 2G networks that are prevalent in rural areas. When asynchronously updating, it is possible to endure intermittent connections with only a little impact on the global model convergence.

5.2 Societal Impacts and Food Security

Farmer economic benefit: Participants of pilot study reported over 12.7% average yield increase (\$180-240 more per hectare) as a result of data driven planting date optimization and resource allocation based on yield predictions. Advances low-yield signals allowed some risk reduction with proactive input usage (extra irrigation, pest control).

Global scale potential: Framework's minimum hardware demands (Raspberry Pi = \$35 and sensors = \$50; total investment = \$85) can be disseminated on 500 millions smallholder farms. With 10% adoption and \$200 average income gain: \$10 billion in annual poverty reduction effect. Data sovereignty model mitigates farmer resistance to share their data with commercial parties, and this overcomes a major obstacle for technology adoption.

Climate: Yield forecasts include climate inputs so farmers can adapt planting schedules to changing rainfall. The continuous learning of Framework (as it updates its model when new season data is made available) will help ensure that the model retained currency as climate changed, whereas such an approach was not possible for a static statistical model.

5.3 Limitations and Deployment Challenges

Data quality variance: Some farms have no record of any kind, leading to noisy labels (estimated 8-12% error rate in self-reported yield). Aggregation seems robust to label noise (due to large batch sizes), which reduces this effect, but steps towards better data collection incentives (e.g., subsidize sensors) would improve accuracy.

Initial investment in infrastructure: As stated the \$85 per farm cost is acceptable for co-operatives or government programs, but the poorest of farmers need subsidization or equipment sharing. The most significant obstacle for working in many rural areas is connectivity to the internet, satellite internet extension (Starlink etc.) may reduce this limitation within 3-5 years.

Integration system of extension: Predictions of yield need interpretations and plans for use. Train agricultural extension agents to interpret model outputs into specific recommendations, and enable feedback mechanisms where farmer performance can inform subsequent model iterations, in order for successful adoption among farmers.

Adversarial points of attack: Byzantine attacks (defective gradient farms) may degrade the quality global model. To filter extreme updates, Framework now employs statistical outlier detection, but powerful adversaries necessitate stronger defenses (e.g., robust aggregation algorithms [14], reputation systems).

5.4 Future Directions

- **Multi-Task FL** – Assuming Good harvest, apply model to predict crop yields, pest risks and ideal time for harvesting concurrently in a combined framework
- **Federated transfer learning:** Allow knowledge transfer from data-rich commercial farms to smallholders networks while protecting commercial data privacy
- **Add blockchain:** Use decentralized publication via Ethereum smart contract instead of a trusted central server
- **Cross-chain federation:** Link regional farm networks that are isolated into a global learning system, but keep regional data sovereignty.
- **Policy Frameworks:** Model governance standards, farmer data rights law and cooperative ownership mechanism for federated agricultural AI

5.5 Conclusion

This methodology covers a crucial gap in the convergence of privacy-preserving, minimal-data-learning, and sustainable agricultural AI. Integration produces 87.3% accurate production projections from less than 300 samples per farm, while also allowing for complete data sovereignty, which is necessary for equitable smallholder involvement. Using cross-farm knowledge transfer, the federated architecture gets 82.6% accuracy at 200 samples, whereas isolated local models only get 61.4%. Validation across several countries and crops shows that the approach is strong and can be used on a large scale. Potential for impact: \$10 billion or more in poverty reduction via higher yields, climate resilience, and giving farmers control over their own data. The framework provides the technical groundwork, but for the global effect to happen, farmer cooperatives, governments, NGOs, and technology providers need to work together to make sure equitable governance and sustainable finance models are possible.

REFERENCES

1. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175-1191. <https://doi.org/10.1145/3133956.3133982>
2. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407. <https://doi.org/10.1561/04000000042>
3. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), 1-210. <https://doi.org/10.1561/22000000083>
4. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50-60. <https://doi.org/10.1109/MSP.2020.2975749>
5. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273-1282.
6. Pham, Q. V., Dev, K., Maddikunta, P. K., Gadekallu, T. R., & Huynh-The, T. (2021). Fusion of federated learning and industrial Internet of Things: A survey. *Computer Networks*, 192, Article 108023. <https://doi.org/10.1016/j.comnet.2021.108023>
7. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), Article 119. <https://doi.org/10.1038/s41746-020-00323-1>
8. Saiz-Rubio, V., & Rovira-Más, F. (2020). From smart farming towards agriculture 5.0: A review on crop data management. *Agronomy*, 10(2), Article 207. <https://doi.org/10.3390/agronomy10020207>
9. Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9, 4843-4873. <https://doi.org/10.1109/ACCESS.2020.3048415>
10. Smith, V., Chiang, C. K., Sanjabi, M., & Talwalkar, A. S. (2017). Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30, 4424-4434.
11. Truong, N., Sun, K., Wang, S., Guitton, F., & Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. *Computers & Security*, 110, Article 102402. <https://doi.org/10.1016/j.cose.2021.102402>

12. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, Article 105709. <https://doi.org/10.1016/j.compag.2020.105709>
13. Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. *Springer International Publishing*. <https://doi.org/10.1007/978-3-319-57959-7>
14. Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., & Chan, K. (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6), 1205-1221. <https://doi.org/10.1109/JSAC.2019.2904348>
15. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), Article 12. <https://doi.org/10.1145/3298981>
16. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216, Article 106775. <https://doi.org/10.1016/j.knosys.2021.106775>
17. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*. <https://arxiv.org/abs/1806.00582>
18. Zhu, H., & Jin, Y. (2021). Multi-objective evolutionary federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4), 1310-1322. <https://doi.org/10.1109/TNNLS.2019.2919699>