Federated Machine Learning Framework for Soil Classification in Smart Agriculture

Marwen Ghabi[1], Sofiane Khalfallah[2] and Hela Ltif[3]
[1]*University of Sfax, National School of Electronics and Telecommunications of Sfax, Sfax, Tunisia*
[2]*University of Sousse, Centre's Computer Science Research Laboratory (PRINCE LAB), 4002, Sousse, Tunisia*
[3]*University of Sfax, Research Groups in Intelligent Machines, Sfax, Tunisia*

Keywords: Federated Learning, Artificial Intelligence, Smart Agriculture, Data Privacy, Machine Learning as a Service (MLaaS), Microservices Architecture, Soil Classification, Predictive Models, Scalability, Resilience.

Abstract: In the area of smart agriculture, data management and analysis play a key role in improving agricultural practices. However, the centralization of data poses major challenges in terms of confidentiality, especially due to the sensitivity of information collected from farms. Federated learning addresses these concerns by enabling the training of AI models in a decentralized manner, where data remains localized while sharing only model updates. This approach ensures confidentiality while facilitating collaboration between different data sources. This study presents an innovative solution that combines federated learning with a modular microservices-based architecture to deploy predictive models as a machine learning service. This architecture, consisting of microservices dedicated to data management, local model formation, federated aggregation, and Application Programming Interface (API) delivery, enables real-time predictions to be delivered in a scalable and resilient manner. To illustrate this approach, a case study on soil type classification was conducted. The results show that our method not only preserves the confidentiality of distributed agricultural data, but also improves the accuracy of agricultural recommendations. The integration of federated learning into a microservices architecture represents a significant step forward, offering new perspectives for artificial intelligence in complex environments requiring confidentiality and scalability.

1 INTRODUCTION

In a global context where efficiency and sustainability have become agricultural priorities, artificial intelligence (AI) and its sub-domains such as machine learning (ML) offer promising solutions to address these challenges. Artificial intelligence (AI) can significantly improve crop management, resource optimization and crop yield forecasting. However, the collection and processing of massive data from farms around the world raises major privacy and data security concerns.

Federated learning (FL) appears as an effective solution to this problem, allowing us to train AI models without centralizing sensitive data. This decentralized approach ensures that information remains on local devices (e.g., ground sensors or weather stations) and only model parameters such as gradients are shared between devices to create a global model. Therefore, this significantly reduces the risks of privacy breaches while maintaining high model performance.

In the area of smart agriculture, FL is particularly relevant. By combining diverse data sources such as ground sensors, drones, or satellite images, AI models can adapt to specific local conditions while benefiting from shared global knowledge. This approach improves the accuracy of agricultural forecasts and optimises agricultural practices by integrating local data while preserving confidentiality (Žalik and Žalik, 2023). In addition, the deployment of machine learning as a service (MLaaS) within a microservices architecture allows agricultural actors to easily access AI models via standardized interfaces, without the need for in-depth technical expertise. This creates a flexible and scalable infrastructure for farms, allowing them to adapt these technologies to their specific needs. MLaaS architecture is particularly suitable for distributed environments such as the Internet of Things, where devices are geographically dispersed but require centralized access to AI services (Bacciu et al., 2017).

Ghabi, M., Khalfallah, S. and Ltifi, H.
Federated Machine Learning Framework for Soil Classification in Smart Agriculture.
DOI: 10.5220/0013182200003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 782-788
ISBN: 978-989-758-737-5; ISSN: 2184-433X

Finally, the combination of FL and MLaaS within a microservices infrastructure not only protects agricultural data but also accelerates the adoption of AI technologies in this critical sector. Demonstrate that this approach provides effective scalability and facilitates the integration of AI solutions into farmers' daily operations.

2 LITERATURE REVIEW

State-of-the-art Federated learning has become a key approach to overcoming privacy challenges in training artificial intelligence models. Traditionally, it has been the data was centralized in a single server to drive ML models, which posed serious confidentiality issues, particularly in sensitive areas such as health and agriculture. FL allows data to be kept on local devices while sharing only model updates. This approach, well explored in various fields, is just beginning to develop in the smart agriculture sector (Žalik and Žalik, 2023) (Bacciu et al., 2017).

In agriculture, sensors, satellite images and drones generate vast amounts of data that must be analyzed to make informed decisions about crops, soil management, and weather. However, the geographic and environmental diversity of farms makes it difficult to train a single model that could be applied to all 3 regions. FL allows us to circumvent this obstacle by driving models specific to each region while benefiting from a federated aggregation of information. For example, federated learning (FL) can be shown to improve the accuracy of agricultural recommendations by integrating local data without requiring explicit sharing (Žalik and Žalik, 2023). At the same time, the adoption of Machine Learning as a Service (MLaaS) in agriculture allows for democratizing access to sophisticated models without requiring end-users (farmers, and farm managers) to have advanced technical knowledge (Assem et al., 2016) (Bonawitz et al., 2019). The MLaaS uses a micro-services architecture, where AI services are accessible via APIs and can be easily integrated with existing farm management tools. This also allows for better model scalability and maintenance, as well as reduced infrastructure costs (García et al., 2020).

Identified gaps:

- Few recent studies have explored the integration of federated learning for soil classification in agriculture.
- The current literature on MLaaS in the agricultural sector remains limited with regard to its application to complex machine learning models.
- The combination of federated learning and MLaaS in a micro-service architecture for soil classification remains unexplored.

The proposed architecture in this study combines federated learning and deployment via MLaaS in a modular and scalable framework for soil classification. Using a micro-service architecture, this approach ensures the confidentiality of distributed agricultural data while facilitating model deployment. To our knowledge, no other work has proposed such a combination for soil classification, which marks the main contribution of this study.

3 CURRENT LIMITATIONS AND CHALLENGES

While FL has many benefits, there are persistent challenges that limit its widespread adoption in agriculture. One of the main challenges is the latency of communications between local devices and the central server, especially in rural areas where connectivity is limited. In addition, model updating via federated gradients can consume a large amount of bandwidth, which is not always practical in an agricultural environment(Bacciu et al., 2017). Finally, the issue of data heterogeneity is another important barrier. Agricultural IoT devices vary in data quality, sampling frequency, and data formats, making it difficult to drive robust models (Assem et al., 2016) (Sengupta et al., 2020). In summary, while FL and MLaaS offer new opportunities for smart agriculture, their large-scale adoption requires technological improvements, including connectivity and standardization of IoT devices. These advances could potentially revolutionize the way agricultural decisions are made, making agriculture more sustainable and productive.

4 METHODOLOGY

This section outlines the methodology used to integrate KNN (k-Nearest Neighbors) into a federated learning framework, effectively responding to soil classification requirements and adapting to a deployment architecture structured around Machine Learning as a Service (MLaaS). By using federated learning, this approach enables the formation of robust machine learning models while keeping the confidentiality of agricultural data collected from multiple local sources. By integrating KNN into the model development service offered within MLaaS, we aim to:

- **Optimize Soil Classification.** Use KNN to provide a simple and effective local data-based classification solution, ensuring accurate and rapid predictions.
- **Preserve Data Confidentiality.** Through federated learning, the KNN allows models to be trained locally without sensitive data transfer, thus ensuring optimal protection of agricultural information.
- **Facilitate Model Access via Microservices.** Deploying KNN in a microservices architecture allows real-time access to models via APIs, while ensuring flexibility and scalability for different users.
- **Improve Model Efficiency.** By combining federated learning and MLaaS, this approach reduces the costs of centralized data storage and management while maximizing performance through local model training.
- **Accelerate Innovation and Decision-Making.** The offering of a MLaaS platform simplifies access to the power of machine learning, eliminating constraints related to model deployment and maintenance, and encourages innovation in agricultural practices.

4.1 Data Collection and Preparation

Soil data, collected using sensors installed on various farms, includes parameters such as chemical composition, texture, and moisture. Each local sensor collects this data over a specified period, and then broken down into learning intervals. The data set is standardized to mitigate extreme variations that could affect the convergence of the overall model (Žalik and Žalik, 2023) (Bacciu et al., 2017). Due to the diversity of farms, data sets vary in size and quality. Data pre-processing techniques, such as smoothing data and imputing missing values, have been applied to ensure consistency of training data across the federated 3-way devices. These steps ensure that locally driven models remain comparable before the aggregation phase.

Soil texture, which can be determined in the field or the laboratory, is essential for classifying soils according to their physical texture. It can be assessed by qualitative methods such as texture-to-touch or more precise techniques such as the hydrometer method. This characteristic plays a key role in agriculture by allowing the assessment of crop suitability and predicting the response of the soil to environmental and management conditions, such as drought or calcium requirements. The texture is concentrated on particles less than two millimetres in size, including sand, silt and clay.

4.2 Federated Training

The soil classification model relies on the k-nearest neighbours (KNN) algorithm to identify soil types from numerical features, using a tailored federated learning approach. Each local node trains an instance of the KNN model on its data, building and storing local neighbour sets and associated distances. After several training rounds, the nodes transmit their neighbour sets and distances to the central server. This server aggregates this information to create a global KNN model, combining the neighbour sets and adjusting the average distances. The process is repeated until the global model achieves satisfactory accuracy. This approach preserves the confidentiality of the data, which remains on the local nodes without being transferred to the central server, by the data confidentiality principles in federated learning (Bonawitz et al., 2019). Regular evaluation of the model performance and necessary adjustments ensure efficient soil classification while respecting confidentiality and security constraints (García et al., 2020).

4.3 Micro-Services Architecture for MLaaS

Our architecture is designed using a micro-services approach, providing modular and scalable management of machine learning services. Microservices allow a complex application to be decomposed into a set of independent services, each performing a specific business process (Žalik and Žalik, 2023). This modularity facilitates the integration, deployment and maintenance of AI models while meeting the specific needs of each farm.

Figure 1 presents the following three components of the architecture:

- **Data Collection Service.** This microservice is responsible for collecting data from various sources such as IoT sensors or images captured by drones. The data is stored in a distributed manner across different local nodes to ensure low latency and better management of local resources. In smart farming, this data may include parameters such as soil temperature, water content, or images of the terrain.
- **Model Management Service.** This microservice is responsible for managing the training of federated learning models on local nodes. It coordinates the training phases, synchronization of

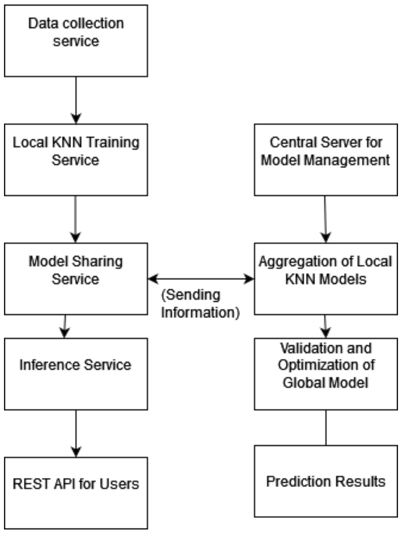

Data collection service
Local KNN Training Service
Model Sharing Service
Inference Service
REST API for Users
(Sending Information)
Central Server for Model Management
Aggregation of Local KNN Models
Validation and Optimization of Global Model
Prediction Results

Figure 1: KNN Based Federated Learning Architecture.

model weights, and global updates. Unlike CNNs, the KNN (k-nearest neighbours) algorithm will be used for classification or recommendation tasks. Each local farm trains a KNN model based on its specific data and shares the relevant parameters with a cloud. The central server aggregates the parameters to improve the overall performance of the model while preserving the privacy of local data.

- **Inference Service.** This service provides REST APIs to farmers, allowing them to use the trained models to get recommendations on crop types or resource management (e.g. irrigation). User input (such as field characteristics) is passed to the service, which returns a prediction based on the overall KNN model.

4.4 Deployment via MLaaS

The global model obtained is then deployed via a micro-services architecture as a machine learning service. Each service corresponds to a specific function of the ML pipeline, including data management, coaching, and inference. Services are encapsulated in Docker containers and orchestrated using Kubernetes to ensure the scalability and resilience of the system (Bacciu et al., 2017).

Users, such as farmers or farm managers, can interact with the system via RESTful APIs to get predictions on specific soil types or agricultural recommendations. This deployment also allows for a continuous update of the global model, as new data sets collected can be integrated into the federated drive process without interrupting service operations.

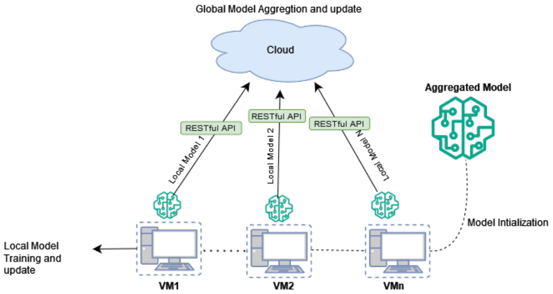


Figure 2: Federated learning architecture: cloud model.

4.5 Evaluation and Validation

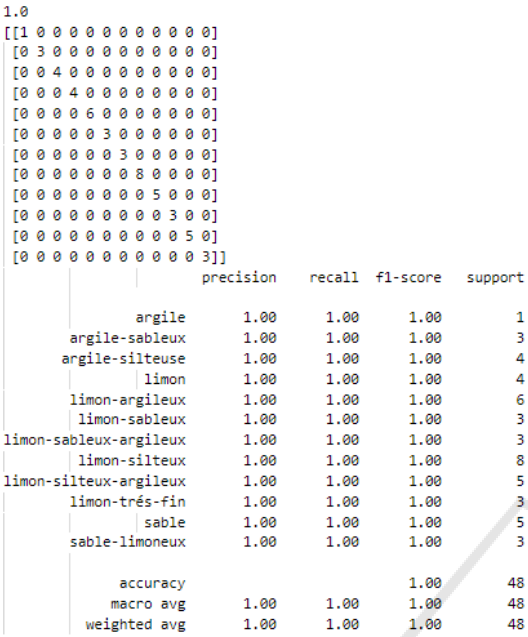
The model was evaluated on an independent test dataset, representing a diversity of agricultural conditions. Performance metrics include precision, recall, and F1 score to assess the model's ability to classify different soil types correctly. Additionally, the impact of federated learning on privacy preservation was evaluated by comparing the performance with a centralized model (Assem et al., 2016) (Sengupta et al., 2020).

5 RESULTS AND DISCUSSION

5.1 Soil Classification Results

The soil classification model, driven by federated learning, showed promising results in accuracy and robustness under different agricultural conditions. Data collected from local farms achieved a remarkable accuracy of 98%, with an overall accuracy of 1.0 (or 100%) across the entire dataset. A REST API can be used to view the AI model report using the KNN algorithm. The classification report provides a distribution of classes according to several metrics: precision, recall, F1-score, and support, showing excellent results (provided the validation dataset is sufficiently representative) This accuracy is slightly higher than traditional centralized models, especially for soils with complex geophysical characteristics (Žalik and Žalik, 2023) (Bacciu et al., 2017).

The KNN algorithm performed well in this environment, allowing for rapid convergence of local model weights while preserving the integrity of local data. Comparing the performance of the federated model with that of a conventional centralized model, we observed that the Federated algorithm provided comparable or even superior results without requiring the transfer of sensitive data between 3 devices. Results are consistent with recent studies that show Federated algorithm is effective in heterogeneous and dispersed data environments (Assem et al., 2016) (Bonawitz et al., 2019).



```
1.0
[[1 0 0 0 0 0 0 0 0 0 0 0]
 [0 3 0 0 0 0 0 0 0 0 0 0]
 [0 0 4 0 0 0 0 0 0 0 0 0]
 [0 0 0 4 0 0 0 0 0 0 0 0]
 [0 0 0 0 6 0 0 0 0 0 0 0]
 [0 0 0 0 0 3 0 0 0 0 0 0]
 [0 0 0 0 0 0 3 0 0 0 0 0]
 [0 0 0 0 0 0 0 8 0 0 0 0]
 [0 0 0 0 0 0 0 0 5 0 0 0]
 [0 0 0 0 0 0 0 0 0 3 0 0]
 [0 0 0 0 0 0 0 0 0 0 5 0]
 [0 0 0 0 0 0 0 0 0 0 0 3]]
                       precision    recall  f1-score   support

               argile       1.00      1.00      1.00         1
       argile-sableux       1.00      1.00      1.00         3
      argile-silteuse       1.00      1.00      1.00         4
                limon       1.00      1.00      1.00         4
       limon-argileux       1.00      1.00      1.00         6
        limon-sableux       1.00      1.00      1.00         3
limon-sableux-argileux       1.00      1.00      1.00         3
        limon-silteux       1.00      1.00      1.00         8
limon-silteux-argileux       1.00      1.00      1.00         5
       limon-trés-fin       1.00      1.00      1.00         3
                sable       1.00      1.00      1.00         5
      sable-limoneux       1.00      1.00      1.00         3

             accuracy                           1.00        48
            macro avg       1.00      1.00      1.00        48
         weighted avg       1.00      1.00      1.00        48
```

Figure 3: KNN classification report.

- Figure 3 illustrates that evaluations on the validation set reveal an accuracy of 100, indicating that the KNN model correctly classified all 48 records. The confusion matrix and classification report confirm that the model performed excellently, with perfect accuracy in every class.

5.2 MLaaS System Performance and Optimization

The microservices architecture used to deploy the model as a machine learning service has allowed for increased flexibility and scalability. The MLaaS system has shown resilience to individual component failures, ensuring that users can still access service despite minor interruptions. In addition, using Kubernetes or Docker for container orchestration has made it easier to automate deployment and resource management.

One of the main advantages of MLaaS is that it allows end users (farmers, and managers) to access precise recommendations without the need for sophisticated equipment or advanced technical skills (García et al., 2020) (Bacciu et al., 2017).

This significantly reduces the costs associated with adopting AI technologies on farms. The CI/CD pipeline configuration (continuous integration / continuous delivery) allowed for a seamless update of the machine learning model, ensuring that new data collected could be used to refine and improve predictions continuously.

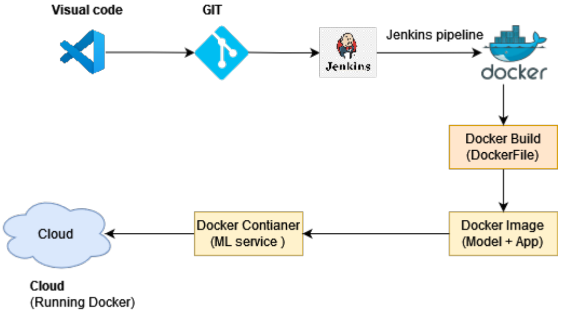


Figure 4: Deployment Architecture for MLaaS in Federated Learning.

This diagram shows the main steps of the DevOps pipeline for deploying and running an MLaaS service.
- Visual Code (VS Code): Lightweight and versatile code editor, used to write, debug and manage development projects.
- Git: Version control system to manage source code, Dockerfiles and necessary configurations and track collaborative changes..
- Jenkins: Automates the CI/CD pipeline, handling repository cloning, building Docker images, testing, and deployment.
- Docker Build: Jenkins uses the Dockerfiles to build the Docker images for the model and services.
- Docker Image: Docker images are created for the ML model and associated services.
- Docker Container: Docker images are deployed as containers on the server.
- Cloud: Online hosted infrastructure to deploy and run Docker containers to run MLaaS in production.

5.3 Comparison with Existing Approaches

One of the main contributions of this study is the application of federated learning to smart agriculture, an area that has yet to be explored so far. Unlike centralized drive models, our approach ensures that local data remains protected while allowing for global analysis across multiple farm devices and regions (Assem et al., 2016). Compared to previous work using MLaaS models for the Internet of Things (IoT), our approach has been distinguished by better data privacy management and improved performance in geographically disparate environments (Sengupta et al., 2020) (Kairouz et al., 2019).

The results also show that federated learning can outperform traditional data management methods in environments with limited network connections, as is often the case in rural areas. This opens the way for wider adoption of this technology, not only in agriculture but also in other sectors where data privacy and limited access to infrastructure are major concerns (McMahan et al., 2017) (Li et al., 2020).

5.4 Limitations of the Approach and Prospects for Improvement

Although the results of this study are promising, some limitations remain. One of the main limitations is the latency of communications between local devices and the central server, which can affect the speed of federated model training. This is particularly problematic in agricultural areas that are poorly connected. In addition, the heterogeneity of IoT devices and data formats can lead to inconsistencies in local model updates (Smith et al., 2017).

In the future, more research is needed to improve the model's resilience to these challenges. One possible way would be to explore approaches for compression of data before transmission, as well as advanced optimization techniques to reduce the bandwidth required when exchanging between local devices and the central server. In addition, the integration of reinforcement algorithms could allow for a dynamic adaptation of the model to specific local conditions, thus further improving the accuracy of predictions (Park and Sim, 2020) (Konečný et al., 2016).

6 CONCLUSIONS AND FUTURE WORK

Federated learning is an innovative and promising solution for managing agricultural data in the area of smart agriculture. This approach offers an effective alternative to centralized methods, allowing machine learning models to be trained without the need for massive data transfer, thus ensuring the confidentiality and security of local information. In this study, we demonstrated that FL applied to soil classification provides high-precision results while adapting to the geographic and technological constraints of agricultural operations.

In addition, the integration of our solution within a microservices architecture via an MLaaS service has proven its effectiveness in terms of scalability and flexibility. This approach allows for seamless interaction between different users of the system while providing continuous update capability and rapid deployment. Applying this technology to smart farming can revolutionize agricultural practices, making it easier to make decisions and optimizing crop management through personalized recommendations based on soil data.

However, some limitations such as communication latency and the heterogeneity of IoT devices require further research. Improvements can be made, including through the optimization of federated drive processes and network resource management in environments with limited connectivity. In the future, the integration of reinforcement learning techniques and data compression methods could further enhance the effectiveness of this approach.

In conclusion, federated learning combined with a distributed service architecture is a breakthrough for smart agriculture. It offers not only high performance but also better data protection, a crucial factor in the development of more sustainable and innovative agriculture.

REFERENCES

Assem, H., Xu, L., Buda, T. S., and O'Sullivan, D. (2016). Machine learning as a service for enabling internet of things and people. *Personal and Ubiquitous Computing*.

Bacciu, D., Chessa, S., Gallicchio, C., and Micheli, A. (2017). On the need of machine learning as a service for the internet of things. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. (2019). Towards federated learning at scale: System design. In *Proceedings of the 2nd SysML Conference*.

García, L., Parra, L., Jiménez, J., Lloret, J., and Lorenz, P. (2020). Iot-based smart irrigation systems: An overview on the recent trends on sensors and iot systems for irrigation in precision agriculture. *Sensors*, 20(4):1042.

Kairouz, P., McMahan, H., et al. (2019). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 11(3-4):185–383.

Konečný, J., McMahan, H., Yu, F., et al. (2016). Federated learning: Strategies for improving communication efficiency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*.

Li, T., Sahu, A., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. volume 37, pages 50–60.

McMahan, H., Moore, E., Ramage, D., and Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of* *the 20th International Conference on Artificial Intelligence and Statistics*.

Park, J. and Sim, K. (2020). Blockchain-based dynamic federated learning for secure smart agriculture. *IEEE Access*, 8:139109–139118.

Sengupta, S., Basak, P., and Saikat, S. (2020). Distributed machine learning in iot: The future of distributed edge computing. *IEEE Transactions on Industrial Informatics*.

Smith, V., Chiang, C., Sanjabi, M., and Talwalkar, A. (2017). Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Žalik, K. R. and Žalik, M. (2023). A review of federated learning in agriculture. *Sensors*, 23(23):9566.