



OPEN **Balancing centralisation and decentralisation in federated learning for Earth Observation-based agricultural predictions**

Robert Cowlshaw^{1✉}, Nicolas Longép ^{2,3} & Annalisa Riccardi^{1,3}

Crop yield prediction using Earth Observation data presents challenges due to the diverse data modalities and the limited availability of relevant datasets, which are often proprietary or private. Decentralised federated learning has been proposed as a solution to address these privacy concerns as no data labels will have to be distributed to a third party. However, the performance of federated learning is significantly influenced by the number of clients and the distribution of data among them. This study investigates the impact of aggregation levels on federated learning using a proxy model trained on crop type data derived from Copernicus Sentinel-2 images. Interaction of these aggregation levels with other parameters is simulated and studied to aim to generalise the results to different situations. The analysis also includes an examination of the current and future distributions of crop yield datasets to determine the optimal aggregation levels for effective federated learning. The findings highlight that dataset size directly affects the learning outcomes as well as the degree of privacy that can be maintained. Other scenarios and the implications of these results are discussed for a future crop-yield decentralised federated learning architecture.

Keywords Decentralised federated learning, Earth Observation, Agriculture predictions, Machine learning, Crop yield, Crop type

In today's rapidly evolving digital landscape, Earth Observation (EO) data has emerged as a critical tool for understanding and managing our planet's resources. EO data, which includes information collected from satellites and other remote sensing technologies, offers a comprehensive view of the Earth's surface, enabling detailed analysis across various domains. This type of data is invaluable for applications ranging from environmental monitoring to urban planning, providing insights that are both broad in scope and highly specific in detail.

The utility of EO data extends far beyond mere observation. It serves as a powerful resource for addressing some of the most pressing challenges in agriculture, such as crop type detection and yield prediction. By analysing EO data, researchers and practitioners can identify crop types across vast agricultural landscapes with unprecedented accuracy. Previous models such as WorldCereal¹, are centralised machine learning model projects, that have completed this task of ground type and crop type detection and segmentation with high accuracy. This capability not only enhances our understanding of agricultural patterns but also informs strategies for improving food security and resource management. Furthermore, the ability to predict crop yields using EO data is crucial for planning and optimising food production, helping to mitigate the risks associated with climate change, such as famines, and fluctuating market demands.

However, despite its potential, there are significant gaps in the availability of crop yield data, primarily due to issues related to privacy and proprietary restrictions. These challenges hinder the full exploitation of EO data in agriculture, as access to reliable and comprehensive yield data is essential for accurate predictions and informed decision-making.

One promising solution to this dilemma is the adoption of decentralised federated learning. Federated learning is an innovative approach to machine learning that allows models to be trained across multiple decentralised devices or servers while keeping the data localised. This means that sensitive data, such as crop yields, can be utilised for training predictive models without ever leaving the owner's control, thus preserving privacy and ownership using methods such as differential privacy² and homomorphic encryption³. Federated

¹Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow G1 1XQ, UK. ² -lab, European Space Agency, 00044 Frascati, Italy. ³These authors contributed equally: Nicolas Long p  and Annalisa Riccardi. ✉email: robert.cowlshaw.2017@uni.strath.ac.uk

learning has been used in many agricultural tasks already from pest control and disease diagnosis⁴⁵ to crop classification⁶, and many other tasks⁷.

Decentralised federated learning offers unique advantages over traditional centralised approaches, particularly in terms of privacy, security, and scalability. By distributing the learning process across a network of participants, or clients, this method reduces the risks associated with data breaches and enhances the robustness of the learning models. Moreover, it facilitates collaboration among diverse stakeholders, including farmers, local governments, nations and multi-national organisations, without compromising the confidentiality of proprietary data.

Decentralised architecture

In decentralised federated learning, the process of aggregating and coordinating model updates is distributed among clients, rather than relying on a centralised server. In this setup, clients share their model updates (weights) directly with each other, typically through a peer-to-peer network. The decentralised architecture is composed of the following modules:

- *Weight propagation and communication*—Each client exchanges model updates with a subset of other clients, known as peers. This peer-to-peer communication is the foundation of decentralised federated learning, where clients send their locally trained models or updates to neighboring peers. A common method to propagate these updates across the network is through a gossip protocol⁸⁹. In this approach, each client periodically shares its model with randomly selected peers. Those peers then further distribute the model updates, ensuring that the information spreads throughout the network over time.
- *Local aggregation and model updating*—Each client maintains its version of the global model, which it periodically updates based on the models received from its peers. This process is known as local aggregation. Clients aggregate the received models, typically by averaging them, to update their local models.
- *Flexibility and resilience in updates*—Unlike in centralised federated learning, where updates are usually synchronised, decentralised federated learning allows for asynchronous updates. Clients can perform local updates and exchange weights with peers at different times, making the training process more flexible and resilient.
- *Weighted aggregation techniques*—The aggregation process in decentralised federated learning can be enhanced by weighted aggregation. In this method, the received models are weighted based on factors such as the number of local training epochs, the size of the local dataset, or the trustworthiness of the peers. This helps balance the influence of different clients on the global model.
- *Ensuring consensus*—To ensure that all clients eventually converge toward a similar model, decentralised federated learning often employs consensus mechanisms. These algorithms help align the models across the network, despite the decentralised and potentially asynchronous nature of the updates.
- *Privacy preservation*—Privacy remains a critical concern in decentralised federated learning. Since each client retains its local data, privacy is inherently preserved. However, to prevent potential information leakage through model updates, privacy-preserving techniques such as differential privacy² or homomorphic encryption³ are often used. These methods ensure that sensitive information is not revealed during the update exchanges.

The number of clients, the entities training the individual local model on their local dataset, participating in federated learning, also has a significant impact on the learning process and model performance. An increased number of clients generally brings greater data diversity, which can improve the model's generalisation ability across different dataset distributions. However, this also introduces communication overhead and exacerbates issues related to non-identically and non-independently distributed (non-IID) data, where data distributions differ significantly between clients. This can lead to slower convergence and unstable training dynamics, as noted by¹⁰ and¹¹ in their work on addressing non-IID data in federated learning.

On the other hand, involving fewer clients reduces communication costs and can speed up convergence but may result in overfitting, particularly if the data is less diverse or biased¹². Balancing the number of clients is thus essential to optimising both the efficiency and effectiveness of federated learning systems, as discussed in comprehensive reviews by¹³. Current methods to reduce the number of clients include selecting a subset of the full group^{14–18}, but this reduces the total size of the dataset across all clients, a problem if the dataset is not large. Methods for when small datasets are used for training are discussed in¹⁹ by combining data and models together with nearby clients.

A possible method, introduced as a central topic of this paper, for reducing the number of clients while maintaining large-enough balanced datasets, is to change the entity aggregating the data to a different scale. This is particularly relevant for the agricultural application here explored as it can be implemented at the different administrative levels of the region/country.

Definition of data aggregation levels and scenarios

To be consistent with terminology in regards to the level at which the data is being aggregated, the following definitions are given based on the entity participating in the decentralised federated learning protocol:

- *Micro level*—Where the farmers are the clients within the decentralised federated learning protocol. Calculating the number of clients at this level can be estimated by the number of farms in the world. Estimates range from 456.07 million in the 1990s²⁰ to 608 million in 2021²¹.
- *Meso Level*—Where the provinces or counties are the clients within the decentralised federated learning protocol. Again, calculating the number of clients can be achieved by estimating the number of provinces

or counties. For example, in the European Union (EU) the Nomenclature of territorial units for statistics (NUTS) gives 104 regions at NUTS level 1, 283 regions at NUTS level 2 and 1345 regions at NUTS 3²².

- **Macro Level**—Where the countries are the clients within the decentralised federated learning protocol. Considering the nations that are member states of the United Nations, there would be 193 clients²³.
- **Mega Level**—Where the multi-national corporations or international organisations are the clients within the decentralised federated learning protocol. Some examples of these could be EU and/or multi-national agricultural organisations who automatically collect data from farmers over large areas. Estimating the number of entities at this level is challenging, but a reasonable approximation is one entity per continent, resulting in 7 Mega Level entities.

From these data aggregation levels the matching scenarios are envisioned:

- **Micro level scenario**—The labellers of the data, in this case the farmers, select and receive the required EO data from the EO data providers. The EO data is then combined with farmers labels, either crop type or crop yield, and train a model specific to the global task. This is then combined into the global decentralised federated learning model and new model weights are returned to them. They continually repeat the training process with these new model weights and their local dataset and send the model weights into the federated learning network and receive new model weights. This model at any time can be used to infer predictions on any available datasets to the farmers. This scenario can be seen in Fig. 1a with the labelled entities given in Fig. 1c.
- **Meso Level Scenario**—In this case the province or county select and receive the required EO data from the EO data providers. The farmers of the province or county provide their respective labels, either crop type or crop yield statistics to the province or county. This provincial/county EO data and farmers labels are combined and used to train a model specific to the global task. This is then combined into the global decentralised federated learning model and new model weights are returned to the province/county. The training is continually repeated with the new model weights and the local provincial/county dataset, followed by sending the model weights to the global model and receiving new weights. Again this model can be used to infer predictions on dataset available to any individual with access to the model. This scenario can be seen in Fig. 1b with the labelled entities given in Fig. 1c.
- **Macro level scenario**—The scenario is the same as at the meso level however the aggregator of the EO data and labels, and the model trainer are countries. This scenario can be seen in Fig. 1b with the labelled entities given in Fig. 1c.
- **Mega level scenario**—The scenario is the same as at the meso and macro levels however the aggregator of the EO data and labels, and the model trainer are multi-national corporations or international organisations. This scenario can be seen in Fig. 1b with the labelled entities given in Fig. 1c.

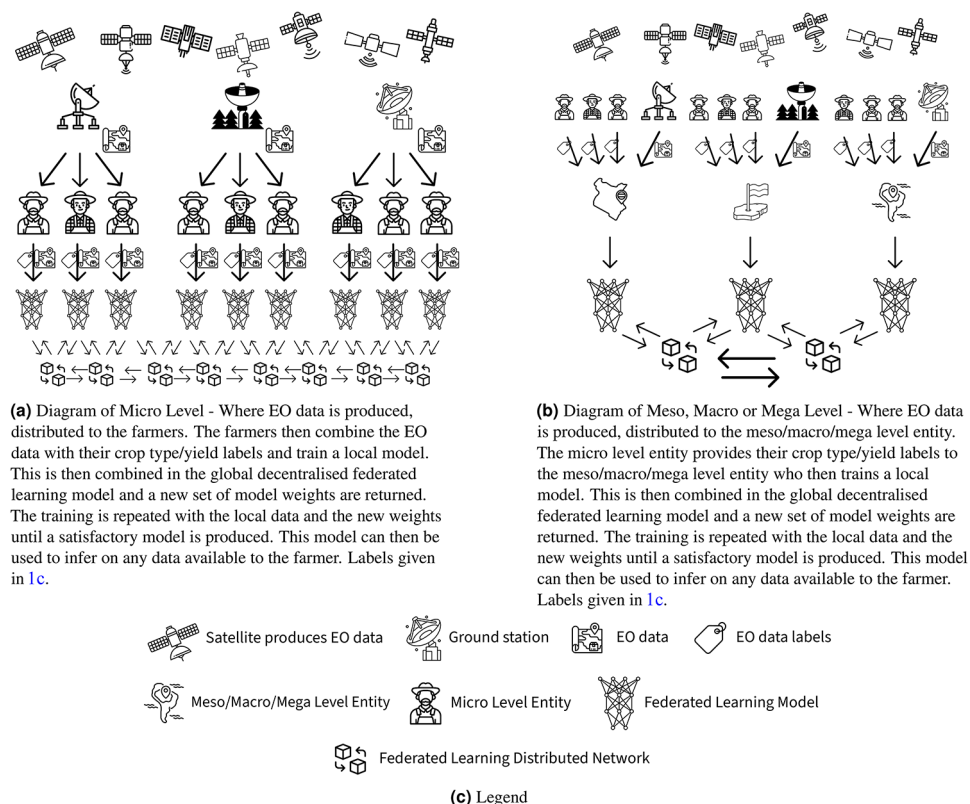


Figure 1. Diagrams of different aggregation levels and scenarios.

This paper tackles the question: How to find the right trade-off between centralisation and decentralisation in federated learning for agricultural predictions? To address this question, several key factors must be examined. First, we need to consider the data that is currently available, as well as what data might become available in the future. Next, to understand how different levels of data aggregation perform in decentralised federated learning, we conduct a simulation for crop type segmentation and classification using real-world data. Additionally, it is crucial to assess how these varying levels of aggregation impact the amount of data that must be transferred across the decentralised network and how privacy is maintained within these datasets. By exploring these factors, this paper aims to clarify the implications of decentralised federated learning in the context of crop yield estimation and to determine the optimal data aggregation levels for crop type detection, ultimately predicting the required level for accurate crop yield prediction.

The main contributions of this paper include demonstrating how the number of clients affect the learning metrics of a federated learning model, presenting the current data labels and Earth Observation (EO) data available for crop type and crop yield prediction, and exploring the different levels at which data aggregation can occur for these agriculture datasets to maximise the effectiveness of model learning while maintaining privacy and security of private and/or proprietary data. Comparison of different parameters to optimise and account for are considered, including simulating federated learning models with scenarios for different data distribution techniques, aggregation levels and number of crop classes in the segmentation task to understand how these parameters interact with each other. Additionally, the paper identifies the specific aggregation levels necessary for developing a decentralised federated learning model aimed at crop yield prediction tasks.

Data availability

EO data

EO data allows for rapid analysis of vast geographical areas, enabling large-scale environmental and agricultural monitoring with unprecedented efficiency. The extensive coverage provided by EO data results in large, diverse datasets, which are ideal for training machine learning models that can improve predictions and insights across a wide range of applications.

Moreover, EO data can capture information from remote and hard-to-reach areas, providing valuable insights where on-the-ground data collection is challenging or impossible. With daily revisit times, EO data can deliver continuous updates, ensuring that decision-makers have access to the most current information available.

To use EO data for crop type classification and yield prediction, a broad range of spectra information are required from the optical satellites in and outwith that of visible light. In previous work, European Space Agency (ESA) WorldCereal¹ uses 9 spectral bands ranging from 490nm to 2190nm from Copernicus Sentinel-2 mission to achieve 97.8% global classification accuracy. The same bands were also used in this work to compare the accuracy achieved with a decentralised federated approach in different scenarios to measure the effect of different aggregation levels. Note that other modalities (or types) of EO data including synthetic aperture radar (SAR) were used in¹. However, only multi-spectral optical S-2 data was used used in this work as sufficient accuracy was obtained in later tests without the use of SAR data.

The federated learning model used in this paper is trained on ESA Copernicus Sentinel-2 surface reflectance products (so called L2A). As the crop type labelled dataset's observation date was on date 01/06/2018²⁴, cloud free images as close in time as possible to this date were selected. The geographic extent of the swath used is between coordinates 50.68°N 2.53°E and 50.89°N and 5.89°E. The full specifications of the data are given in Table 1.

The EO data is then subdivided into smaller patches of 256 × 256 pixels of 10m Ground Sample Distance (GSD), where bands with larger GSD are upsampled to 10m. A representation of the visible light bands of this subdivided EO data where it overlaps the crop yield dataset²⁴, is shown below in Fig. 2. This subdivided dataset can be most easily viewed at²⁵.

The use of EO data for crop yield estimation necessitates the integration of ancillary environmental information. Traditional on-ground measurements, which involve counting the grains per head or pod, provide direct yield assessments²⁷. However, replicating these direct measurements using EO data is challenging. Instead, EO data can be employed to estimate crop yield by assessing the quality and progression of crop growth throughout the growing season. This approach requires temporal EO data that captures phenological changes, complemented by environmental variables such as precipitation, temperature, and sunlight. These environmental factors, which significantly influence crop growth, can be obtained from either modelled data or observational datasets.

Specification	Value
Product name	Sentinel-2 L2A
Date and time	10:54:40 30/06/2018 UTC
Cloud cover	0%
Geographic extent	50.68° N 2.53° E to 50.89° N 5.89° E
GSD	10 m
Sentinel-2 bands	2, 3, 4, 5, 6, 7, 8, 11, 12

Table 1. Crop type EO data specifications.



Figure 2. Visible light bands (B04,B03,B02) of Belgium EO dataset used in this work. Contains modified Sentinel data 2018²⁶. The missing areas are areas with no crops recorded in the crop type dataset²⁴ this includes areas such as cities and water.

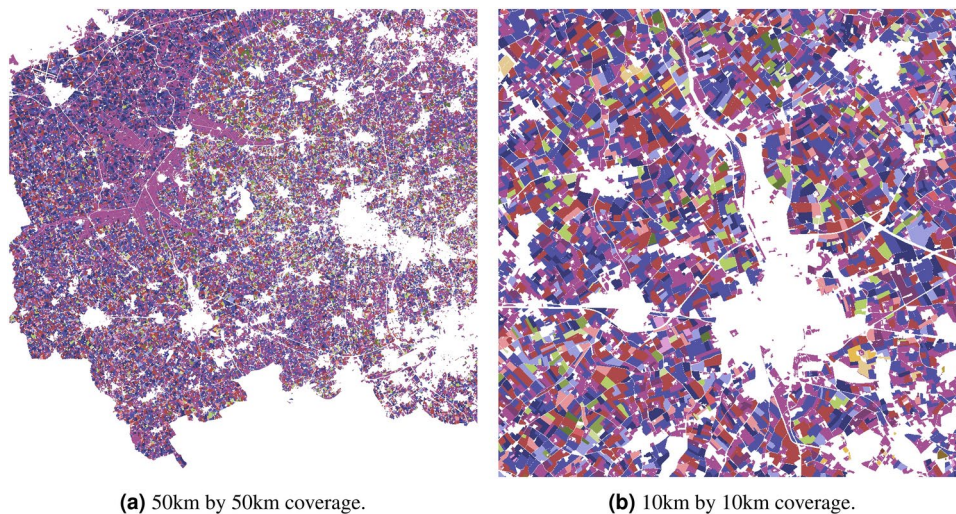


Figure 3. 10m GSD images to show resolution of crop type data available. Each field is separately identified with the crop type. Crop types are randomly coloured for view-ability. (Data from map of Belgium between 50.7 and 51.1511 Degrees North and 2.6 and 3.30915 Degrees East in 2018²⁴).

EO data labels

One challenge with EO data is the need for extensive labelling by experts, who are often geographically dispersed and difficult to coordinate. This can lead to inconsistencies, as the data labelled by different individuals may be non-IID, hindering the training of the machine learning model.

Crop type labels are provided by the farmers and regions across the world. These are often not aligned with each other and stored in different places. Projects such as WorldCereal have aligned many of these datasets, including 74814016 crop fields across the planet as of August 2024²⁸. The 97.8% global classification accuracy means that this model has good generalisation capabilities and can be used to classify new areas. While crop type prediction models are not novel, with several existing models in the literature^{29–31}, the innovation lies in the application of federated learning to these models. This approach contributes to the crop type model's scalability and serves as a proxy to evaluate the scalability of a crop yield prediction model.

The dataset used in the federated learning model here proposed contains 514860 features/fields across Belgium which represent the crop type of each field as of observations on 01/06/2018, an example of which can be seen in Fig. 3. This dataset, referenced in²⁵ was derived from the original dataset²⁴, which contained multiple crop types associated with the same class number. To address this redundancy, similar classes were merged, resulting in a refined dataset with 78 distinct crop types. However, the dataset exhibits an inconsistent distribution across different crop types. To ensure label balance during training, only the top 10 most represented classes were selected. Expanding the dataset would allow for the inclusion of a broader range of labels, enhancing the model's capacity to generalise.

Belgium was selected as the sole country for the crop type dataset in this study due to several reasons. First, the dataset for Belgium is of substantial size, providing sufficient data for training an adequately accurate model for study. The dataset covers the same growing season across all samples, ensuring temporal consistency, and represents a uniform biome, minimising variability due to ecological differences. Since the focus of this study is on exploring different aggregation scales rather than achieving global model generalisation, a global dataset is not required. Additionally, the dataset benefits from low cloud cover, simplifying preprocessing and reducing noise in the data. Because of the origin of the dataset, it is likely that the data has undergone verification procedures, further supporting the presumed accuracy and reliability of the data labels, an essential factor for the study. Lastly, the complexity and resource-intensity involved in creating such a dataset necessitate a practical approach, and limiting the scope to Belgium made this effort feasible without compromising the study's objectives.

Crop yield labels are scarce primarily due to the significant challenges and costs associated with obtaining accurate ground truth data. Collecting reliable yield data often requires extensive fieldwork, specialised equipment, and coordination with local farmers, making it both labor-intensive and expensive. Moreover, the global collection of crop yield data is particularly difficult, as it involves gathering consistent information across diverse agricultural landscapes with varying practices and conditions. Currently, crop yield data can be collected at the mega level by the organisations that produce smart farming equipment. For example, the data is measured by the tractor in the field, and the data is sent back to the manufacturers centralised servers^{32,33}. These datasets are very precise however are proprietary data for the companies. This leads to datasets that are worth a lot of money and are either shared at great expense or not shared at all.

Other datasets that are publicly available are not precise to the level of a few square meters or even to the field. One of the dataset shared by the EU provides data at a resolution of Meso Level per NUTS level 2 region³⁴. This data is patchy with some provinces not existing in the dataset and almost all provinces are missing crop types and/or years of data. The resolution of this data and its extent is displayed in Fig. 4. Without measurements taken at the resolution of a few square meters or per field, this dataset necessitates the aggregation of large volumes of EO data. Such extensive averaging can significantly compromise the capacity to effectively train a machine learning model.

Decentralised federated learning setup

As the crop yield data is not available in high enough resolution to train a model, the crop type segmentation task is used as a proxy to test the different aggregation scales. Initially, the client-side model is established, followed by the implementation of the federated learning aggregation layer. Subsequently, the decentralised protocol layer is configured. The process concludes with the application of a technique for quantifying the results.

To measure the performance of the federated learning on the Micro, Meso, Macro and Mega scenarios, the federated learning model is simulated with the same dataset split between a varying number of clients. These metrics are measured from both distributed validation sets (10% of each client dataset), and the centralised testset, (20% of the overall dataset). There are a total of 5990 labelled patches in the dataset of size 256×256 pixels. Splitting these into the different datasets gives 1198 patches in the centralised testset for all tests, and the 4792 other patches are distributed between the clients.

To measure the outcomes of the studies, five common metrics are calculated from the confusion matrix. These are calculated on the centralised testset and the predictions are compared against the dataset labels. First precision and recall are computed and then averaged across all classes. Secondly, the per class F1-scores are computed and used to calculate the weighted-averaged F1-score with background (F1-W), weighted-average F1-score without background (F1-W-NB) and the non-weighted-averaged F1-score with background (F1-macro).

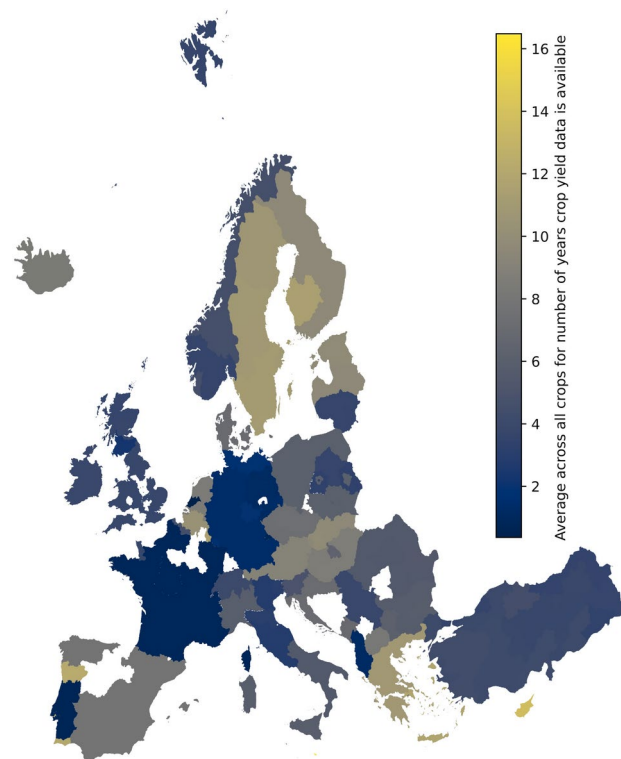


Figure 4. Map of Europe to show crop yield data available from³⁴. Map divided into NUTS level 2 regions. The colouring is the average number of years out of 25 years of possible data that exists across all 79 crop types.

The without background metric is used to analyse when the background is skewing the F1-score due to the existence of class imbalance in the crop type dataset, as the background is encoded as one of the classes.

When determining the parameters and hyperparameters of the local and global model training there are a few things to consider. Firstly, we have the option to conduct a hyperparameter study globally, where optimal hyperparameters are uniformly applied across all clients, or locally, where each client independently optimises their own set of hyperparameters. In this study, we opt for global optimisation of hyperparameters because the clients are simulated and receive a randomised set of labelled data. Given this setup, local optimisation would not be appropriate.

Next we need to determine which parameters to optimise and account for:

- *Model architecture and size*—Different model scales and architectures can train faster and with potentially higher final accuracy.
- *Model aggregation techniques*—Different methods of combining federated individual client models into a single global model exist.
- *Model optimisation functions*—Considerations towards the optimiser, the optimiser function hyperparameters such as learning rate, and potential use of a learning rate scheduler must be considered to find a viable and efficient model training procedure.
- *Loss function*—The type of loss function used as well as the ratio of loss functions if multiple are used.
- *Threshold levels*—A threshold set upon the output soft-maxed neural network layer can add a requirement that the model must acquire a certain level of confidence before it predicts a class.
- *Number of clients in the federation*—This is our primary independent variable to determine how the number of participants effects the learning and final results.
- *Number of classes and class imbalance*—The dataset shows a class imbalance, especially among the most common crop types, as these represent a large share of the total crop classes. There is also an imbalance between the crop classes and the background class, which makes up the largest portion of all classes. When the largest crop classes are included in training, adding more classes further amplifies this imbalance because the sizes of each class tend to decrease sharply, reflecting the distribution of crop types grown in the country represented by the dataset. Consequently, increasing the number of classes leads to greater class imbalance.
- *Dataset distribution across available clients*—The distribution of the dataset across clients can be independent and identically distributed (IID) or non-IID.
- *Dataset quality disparity*—Disparity in the quality of the dataset can come from sources such as accuracy of labelling and dataset privacy protecting measures applied before training.

The first three of these parameters are defined. Optimisation of these parameters is outwith the scope of this work and therefore are determined from literature.

Local client model architecture

The WorldCereal product³⁵ is obtained by training a random forest model. As random forest operates on each pixel rather than patch, spacial autocorrelation is not accounted for. Therefore, as the dataset already includes the spatial correlation element of a patch of pixels, an architecture which does allow for spatial autocorrelation is considered optimal for this study. Therefore, a U-Net is determined as the most suitable model architecture for this parameter. A U-Net model³⁶ is a type of convolutional neural network architecture specifically designed for image segmentation tasks. Originally developed for biomedical image segmentation, U-Nets have gained popularity across various domains due to their effectiveness in precisely delineating the boundaries of objects within an image. The architecture of a U-Net model is characterised by a symmetric U-shaped design as seen in Fig. 5, with an encoder path that captures context and a decoder path that enables precise localisation. This structure allows the network to effectively learn both global and detailed features, making it highly suitable for segmentation tasks where accuracy and detail are critical. Another reason, the U-Net is used in this work is it excels in scenarios with limited training data, as they can achieve high-quality results even with relatively small

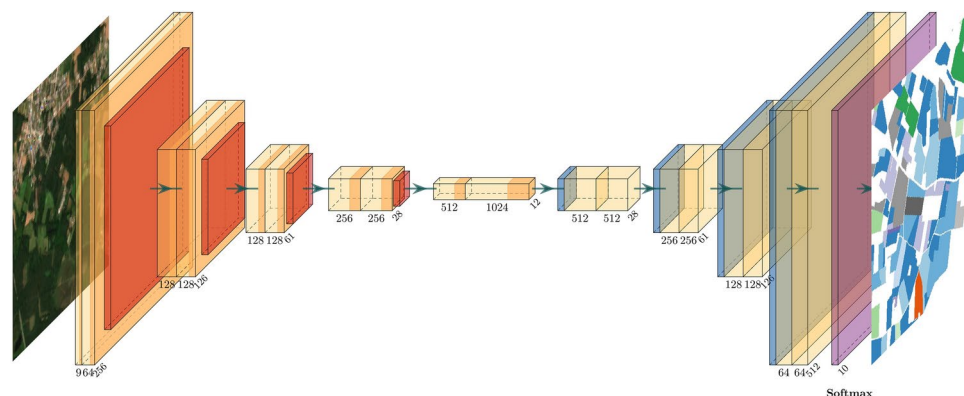


Figure 5. Diagram representing the U-Net used for this work.

datasets. In this work, at the micro level, farmers may provide very small datasets if their farms are limited in size, therefore, the U-Net is chosen to mitigate this problem.

Federated learning aggregation

After the client has trained the model on their local data over a predefined number of epochs, the U-Net model is aggregated together with all the other client side models in the federated learning network. In this study, the local model is trained for 10 epochs, which represents another hyperparameter that can be further tuned to explore local learning capabilities. The aggregation of U-Net models can be performed using various methods, as outlined in³⁷. The comparison of these methods is beyond the scope of this paper as they have already been studied in³⁸. In this paper we use the Federated Averaging (FedAvg)¹² aggregation scheme as this is the most commonly used in federated learning as well as methods similar to FedAvg having been already designed for decentralised networks such as³⁹.

Learning optimisation

U-Net, is computationally demanding and benefits significantly from Adaptive Moment Estimation's (ADAM)⁴⁰ ability to dynamically adjust learning rates per parameter. This adaptability is especially useful in multi-class classification, where gradients vary widely due to imbalanced class distributions and sparse labels. ADAM combines elements of momentum and adaptive learning, resulting in faster convergence and more stable learning compared to traditional optimisers like Stochastic Gradient Descent. This is particularly advantageous in pixel-wise classification with U-Net, as it accelerates training and improves stability even with challenging input distributions. For deep architectures such as U-Net, ADAM's robustness helps ensure consistent performance without extensive fine-tuning, allowing for efficient training with reliable convergence. This quality is advantageous in multi-class tasks where U-Net's complex structure might otherwise require significant hyperparameter adjustments. This allows for a general value to be set for learning rate reducing the hyperparameter search required. Therefore, the learning rate for the ADAM is set to 0.001, a value deemed high enough to provide meaningful learning, while maintaining model generalisation by avoiding overfitting. As ADAM automatically adjusts learning rates per parameter and avoiding overfitting is important, a learning rate optimiser or hyperparameter study is not included in this study.

Loss ratio and output confidence study

A separate hyperparameter study to determine the best value for the loss function parameter λ , that represents the ratio between two loss functions (cross entropy loss and dice loss), and the threshold of the soft-maxed output is undertaken. Testing these parameters on a centralised model provides us with information on the relation between these parameters and the global learning behaviour. The training is done with the same 80%/20% train/test split as that of the federated model and is trained over 50 epochs with 10 crop classes. These conditions where as the models learning begin to plateau. The loss function used for training is given in Eq. (1). Cross entropy loss is chosen to provide more stable learning and dice loss is used to improve learning on class imbalanced datasets such as the one used in this study. Thresholding is studied to understand what level of confidence in the outputs should be required to provide the highest accuracy in the model. The threshold is applied to the output of the soft-maxed probabilities of each class for each pixel. The results of this hyperparameter study, shown in Table 2, provide us the optimal values for λ and the thresholding level to use this study. From these results, a λ of 0.5 with no thresholding gives the highest weighted F1-score and thus is chosen for the federated model. Weighted F1-score including the background class is prioritised over the weighted F1-score without the background class as false positives across the background could artificially inflate the quantity of each crop.

$$loss = \lambda \cdot CrossEntropyLoss + (1 - \lambda) \cdot DiceLoss \quad (1)$$

Number of clients study

For the set values of the hyperparameters described above, a study to determine how the number of clients effects model learning is undertaken. For this study, the number of crop classes is fixed to 10 plus the background, and the 4792 training and validation patches are evenly distributed between a set of varying number of clients. Training occurs over 10 local client epochs for each of 20 federated learning aggregation rounds. These values are selected because, as observed in the results section, they represent the point at which learning begins to plateau. As the independent variable in this simulation, the number of local clients are set to 2, 4, 8, 16, 32, 64, 128 and 256. These are chosen to illustrate the training performance across scenarios with both small and large variations in the number of clients. For each simulation of the federated learning, for each number of clients, the

Loss ratio (λ)	No threshold		0.5 Threshold		0.75 Threshold	
	F1-W	F1-W-NB	F1-W	F1-W-NB	F1-W	F1-W-NB
0.1	0.875	0.725	0.875	0.725	0.870	0.713
0.3	0.883	0.741	0.882	0.739	0.870	0.710
0.5	0.900	0.766	0.898	0.761	0.887	0.732
0.7	0.898	0.779	0.895	0.773	0.860	0.690
0.9	0.897	0.778	0.895	0.775	0.873	0.724

Table 2. Hyperparameter study for centralised model. Significant values are in bold.

Global federated learning parameters		Local client training parameters	
Parameter	Value	Parameter	Value
Total number of clients	2, 4, 8, 16, 32, 64, 128, 256	Client epochs	10
Overall train/test split	80%/20%	Client train/validation split	90%/10%
Aggregation rounds	20	Optimiser type	Adaptive moment estimation (ADAM)
Aggregation method	FedAvg	Optimiser learning rate	0.001

Table 3. Global and local federated learning hyperparameters (number of clients study).

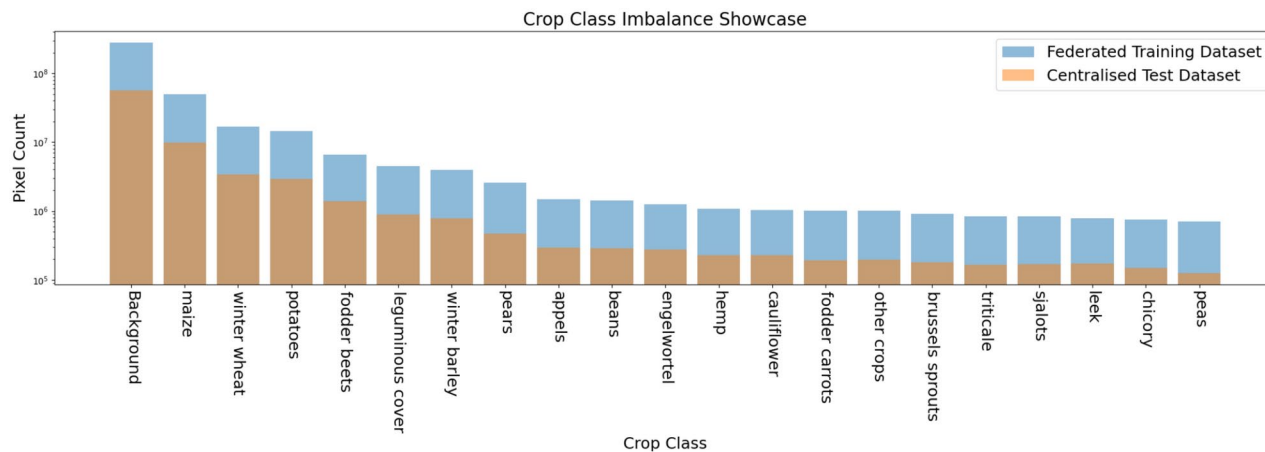


Figure 6. Class imbalance across dataset for federated training set and centralised testset. The crops are sorted from highest number of pixels in entire dataset to smallest. The first 20 classes plus the background class are shown. When 10 classes are used in models, the smallest 10 out of 20 classes are merged into the background class.

experiment is conducted five times. This repetition allows for the calculation of an average performance metric, thereby mitigating the impact of variability introduced by the random partitioning of the dataset prior to the commencement of the training process.

The federated learning aggregation is simulated using the Flower Federated Learning Framework⁴¹, to simulate many clients simultaneously. Table 3 shows an overview of the parameters used across the client side and the global federated learning for the number of clients study, however, further details can be found on the linked source.

Dataset imbalance study

To study the effect of dataset imbalance on learning performances, we can consider changing the number of classes, the distribution of patches across each client, the distribution of classes across clients, and the quality of the dataset across different clients or regions. In this study, only the first two parameters are studied, while disparity in datasets quality and class distribution across clients is only discussed with potential for future work. Based on the results from the study on the number of clients, we identify two values: one for the number of clients that yields the best-performing global model and another for a distinctly lower-performing global model that's learning curve is significantly different in shape from the highest performing model while minimising the difference in number of clients. Therefore, the number of clients in this study is set to 2 and 32, the best performing and lesser performing global models respectively. The parameters on model architecture, aggregation and learning are fixed to the values discussed above, so that only the interaction between number of clients and class imbalance is explored.

The number of crop classes is set at either 10 or 20, plus the background class, this influences not only the total number of classes but also the degree of class imbalance. The crop types are ranked from the most to the least prevalent by pixel count across the entire dataset. Consequently, as more crop classes are included, crops that occur less frequently are incorporated into the background class. This can be seen in Fig. 6 which showcases the 20 crop classes used. When only 10 classes are utilised, the 10 least prevalent crops are absorbed into the background class.

To measure dataset imbalance over the clients, two setups are defined, an IID distribution and a non-IID distribution. The IID dataset distribution is a random but even distribution of all patches in the dataset. The non-IID dataset distribution is a random and non-even distribution of all patches in the dataset. This non-even distribution is defined by each client being assigned an ID number and the number of patches each client receives is linearly correlated with the client ID. The client IDs are distinct and in the set $\{1, N\}$, where N is the total number of clients. Class imbalance between the clients is not studied due to the crop classes density being

Global federated learning parameters		Dataset parameters	
Parameter	Value	Parameter	Value
Total number of clients	2, 32	Dataset distribution	IID, non-IID
		Number of crop classes	10, 20 (plus background)

Table 4. Federated learning and dataset hyperparameters (dataset imbalance study).

Per Class F1-Score for Different Numbers of Clients

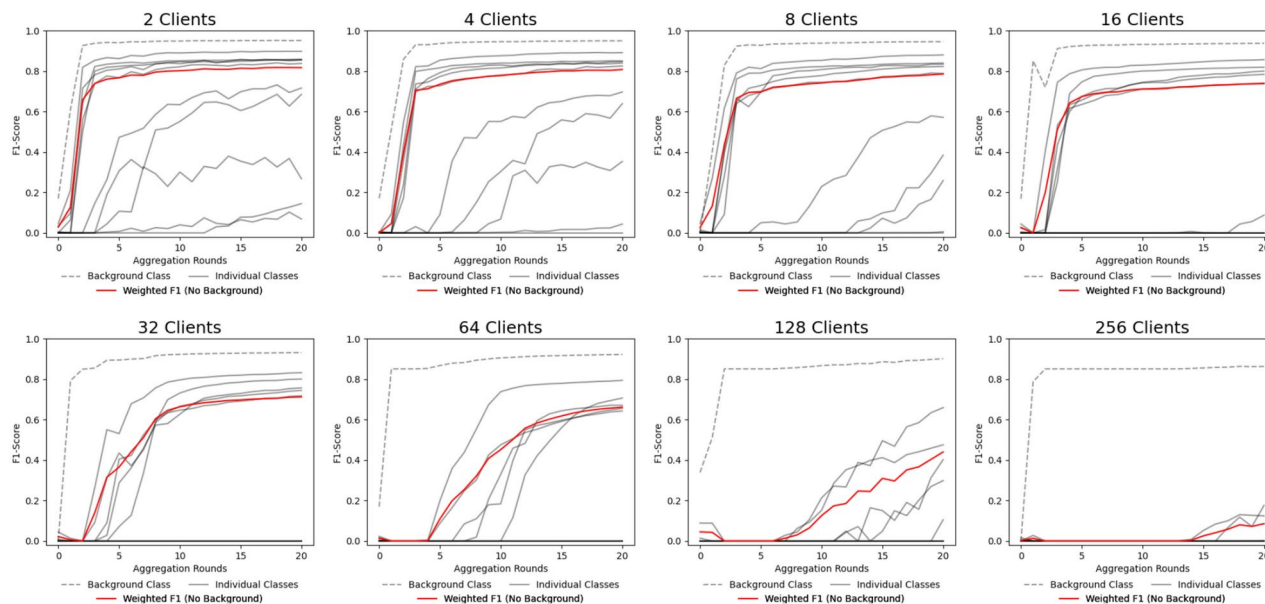


Figure 7. Comparison of per class F1-scores for different number of clients in federated learning (averaged over 5 repetitions). The dataset is distributed in a IID way with 10 crop classes plus the background class. The grey lines represent the F1 score of each class over the aggregation rounds, with the dashed line being the background class. The red line represents the weighted F1 score without the background.

varied throughout each patch, but could be an area for future work. The variable parameters in this study can be seen in Table 4. Other non-IID dataset distribution setups exist, such as exponential or natural distributions⁴², however this is chosen as a simple example to study the effect on training.

The dataset imbalance generated in this simulation is compared to the real world distribution of crop yield data at different aggregation scales. By proxy, this helps determine how the different aggregation scales, and there accompanying dataset imbalances, will impact the training of a global federated learning crop yield model.

Results

Number of clients study

Figure 7 demonstrates that models trained with fewer clients (2, 4, 8, 16) exhibit a relatively steep initial increase in F1-score, which then quickly tapers off as the aggregation rounds progress. This rapid initial increase indicates that with fewer clients, the model can quickly adapt and improve, likely due to each client holding more data or contributing more significant updates during each round of aggregation. In contrast, models trained with more clients (32, 64, 128, 256) show a slower increase in F1-score, with low class F1-score values persisting over a larger number of aggregation rounds. This pattern suggests that as the number of clients increases, the impact of individual client updates on the overall model diminishes, possibly due to smaller data subsets per client or noisier gradients. Consequently, models with fewer clients tend to achieve a higher final F1-score compared to those with more clients, implying that increasing the number of clients may introduce greater variability or noise into the training process, which can hinder the model's ability to converge to a lower loss. Interestingly, Fig. 7 also displays which classes are more difficult to classify, as the number of clients increases, per class F1-scores separate at different number of clients simulations, meaning later separation from the combined weighted F1 score implies an easier crop to classify. Distributed validation accuracy has high variability in F1-score at larger number of clients due to the small validation dataset sizes, as mentioned in Table 5 per client, bringing uncertainty in the metrics and is therefore omitted here. However, the centralised testset does not decrease in size, and therefore the accuracy achieved on the testset remains consistent over different client numbers.

Scenario			Results				
Clients	Train patches	Validation patches	Precision	Recall	F1-W	F1-W-NB	F1 Macro
2	2156	240	0.917	0.923	0.916	0.819	0.659
4	1078	120	0.913	0.919	0.913	0.808	0.63
8	539	60	0.905	0.91	0.904	0.785	0.576
16	270	30	0.877	0.897	0.886	0.739	0.456
32	135	15	0.863	0.886	0.874	0.713	0.435
64	67	7	0.844	0.866	0.854	0.662	0.399
128	34	4	0.757	0.817	0.783	0.442	0.258
256	17	2	0.616	0.747	0.667	0.1	0.108

Table 5. Scores of different number of clients in federated learning (averaged over 5 repetitions). Significant values are in bold.

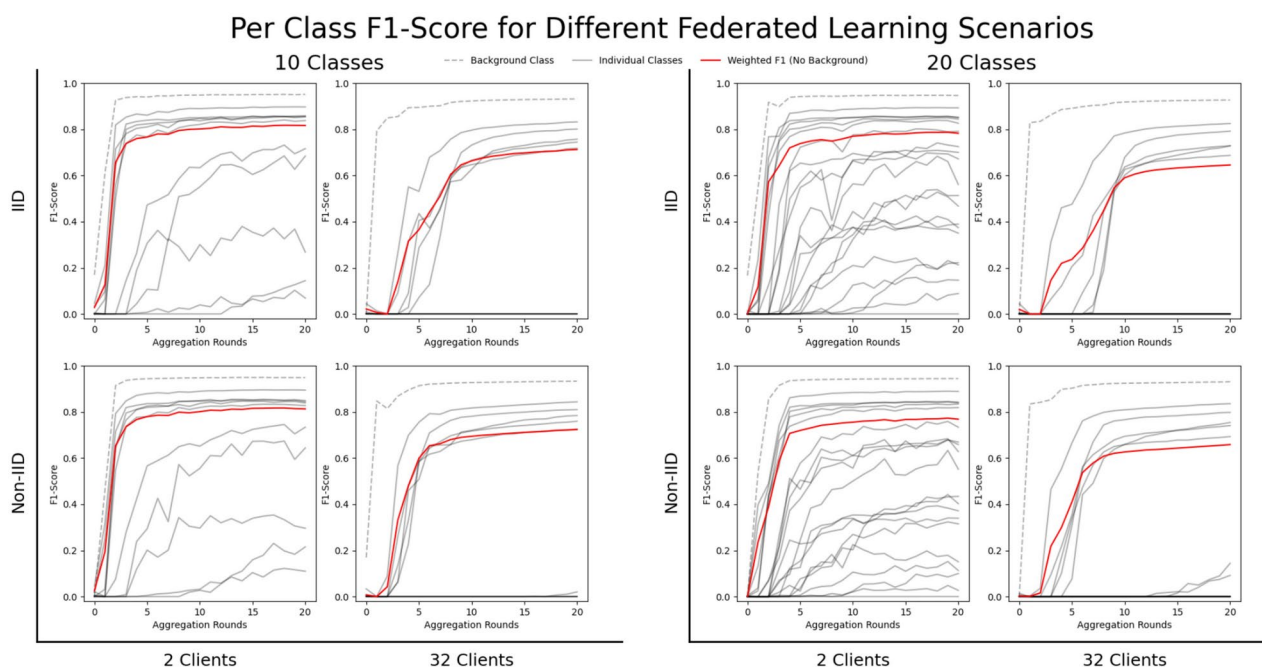


Figure 8. Comparison of per class F1-scores for different federated learning scenarios (averaged over 5 repetitions). This is a grid search across 2 and 32 clients, IID and non-IID distributed datasets, and 10 class and 20 class datasets (plus the background class). The grey lines represent the F1 score of each class over the aggregation rounds, with the dashed line being the background class. The red line represents the weighted F1 score without the background.

Looking at the metrics given in Table 5, it is obvious to see that increasing number of clients decreases all metrics, with 256 clients achieving a poor F1-W-NB score of 0.1. It can be seen that the background class, due to its large proportion of all pixels in the dataset, skews the F1-W scores and therefore both F1-W and F1-W-NB become important to determine the true accuracy of the model. This can also be seen in Fig. 7 where the dashed line, representing the background class F1-score, is consistently high.

Dataset imbalance study

In Fig. 8 the F1-score over aggregation rounds can be seen for parameter study on 2 and 32 clients, IID and non-IID distributed datasets, and 10 vs 20 class datasets. The most obvious differences in F1-score again come from those found in the number of clients study between 2 and 32 clients, however looking closer the difference between IID and non-IID in the 32 client simulations shows a difference in learning rate. IID learning dips in F1-score around 5 aggregation rounds in, before continuing to increase, however this dip does not exist when the dataset is distributed in a non-IID fashion. This can be seen in the 10 class simulations however is more prevalent in the 20 class simulations. Another point to highlight is that in 2 client 10 class simulations, all classes have a final F1-score of more than zero meaning that there must be a correct classification in the centralised testset of all crops occurring. However, this is not the case for 2 client 20 class simulations as some crop classes are achieving

Scenario			Results				
Clients	Classes	IID	Precision	Recall	F1-W	F1-W-NB	F1 Macro
2	10	True	0.917	0.923	0.916	0.819	0.659
32	10	True	0.863	0.886	0.874	0.713	0.435
2	20	True	0.902	0.91	0.902	0.79	0.564
32	20	True	0.826	0.869	0.846	0.645	0.223
2	10	False	0.915	0.92	0.915	0.819	0.668
32	10	False	0.868	0.89	0.878	0.724	0.443
2	20	False	0.897	0.905	0.896	0.774	0.528
32	20	False	0.836	0.874	0.853	0.658	0.237

Table 6. Scores of different federated learning scenarios (averaged over 5 repetitions). Significant values are in bold.

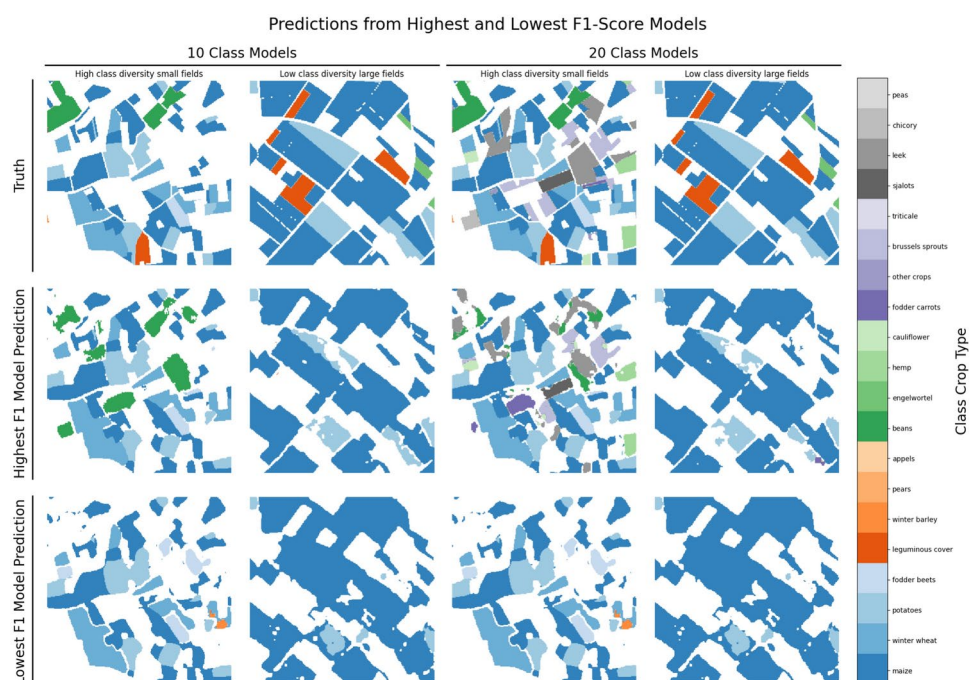


Figure 9. Predictions of highest (2 client IID) and lowest (32 client IID) F1-scoring models for 10 class models and 20 class models compared to the data labels/ground truth. The two patches are selected from the centralised testset representing a patch with small fields and high class diversity and a patch with large fields with low class diversity.

final F1-scores of zero, meaning they are never being classified. This may be because these crops are intrinsically difficult to classify or they occur such a small number of times in the dataset as to where they are never learned.

From the trend seen in Table 5 it can be seen that increasing the number of clients decreases the F1 score, and from this dataset imbalance study, in Table 6, increasing the number of classes also decreases all versions of the F1-score, precision and recall. However, counter-intuitively, when comparing non-IID client datasets to IID client datasets, the F1-score is consistent or improves. This starts to make sense when considering who has the majority of the dataset. For example, when there are 2 clients in the non-IID scenario, one client has 1/3 of the dataset and the other has 2/3 of the dataset. As the FedAvg aggregation algorithm is weighted by size of dataset, the client with the larger dataset contributes more to the global model, and has better training due to a larger training dataset available. This could be a reason for the non-IID scenario performing better than the IID scenario. More significantly, when considering the 32 client non-IID scenario, the client with the largest dataset has $32/496 \approx 6\%$ of the dataset. Again, the FedAvg algorithm is weighted towards the larger datasets, and considering the largest 10 clients hold over 50% of the whole dataset, meaning that the training is more centralised than the IID 32 clients scenario. In this scenario, clients with smaller datasets will contribute minimally, thus providing limited added value to the global model as they are carried through the learning process.

To analyse the structure of the model predictions, Fig. 9 presents results from four models, involving 2 clients and 32 clients, configured for 10 and 20 classes, respectively. These models were used to predict on two selected

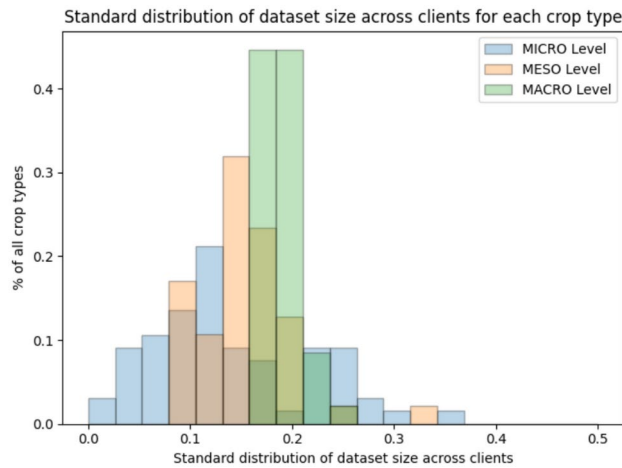


Figure 10. The three datasets are normalised per crop, per year (where years are given), and then the standard distribution is taken per year per crop. The x-values are the standard distribution and the y-values are how many crop types have that standard distribution.

patches from the centralised test set. These patches were chosen to exemplify scenarios with high class diversity and small fields, as well as low class diversity and large fields. This comparison is designed to represent the best and worst-performing models from the dataset imbalance study, each classifying one patch that is relatively easy to classify and another that is difficult. It can be seen that the highest F1 scoring models predict more classes which is to be expected from the F1-scores, but also classifies boundaries better than the low F1 scoring models. In the patches with small fields, the classes transition directly without any intervening background. Conversely, in the patches with large fields, the fields are separated by areas of background. The models with higher F1 scores accurately predict these distinctions, whereas models with lower F1 scores often predict the opposite. Additionally, the high F1 scoring models generally produce better-defined field edges, characterised by sharper edges and corners, compared to the low F1 scoring models, which tend to produce fuzzier edges and more rounded corners. Pre-training of a model with shapes such as on ImageNet⁴³, could help increase the learning speed to define sharper edges and corners as well as recognise other features required in crop type and yield prediction.

If these results from the IID and non-IID comparisons are to be applied to crop yield, we need to determine the dataset distribution across possible clients for the different aggregation levels. This is shown in Fig. 10 where we compare three different aggregation levels:

- Micro Level—The area of every feature in the Belgium crop type dataset²⁴.
- Meso Level—The area harvested per year, per crop, per NUTS level 2 region in Europe, from crop yield dataset³⁴.
- Macro Level—The area harvested per year, per crop, per NUTS level 2 region combined into each country in Europe, from crop yield dataset³⁴.

Analysis of Fig. 10 reveals that at the micro level, the balance of dataset size per client varies significantly by crop type. However, when averaged across all crop types, the micro aggregation level is the most balanced. In contrast, at the macro level, the variation in dataset size across clients is less dependent on crop type, but averaging across all types indicates that macro aggregation is the least balanced.

Discussion

Aggregation level effects on federated learning

The decrease in F1-score as the number of clients increases, as shown in Fig. 7, suggests that aggregation would be more effective with fewer clients. Therefore the Mega aggregation level might be optimal. Nonetheless, the size of each client's dataset is also crucial for effective learning. At the Meso and Macro aggregation levels, the datasets may remain sufficiently large to support adequate training.

Crop yield data

Crop yield models will likely be specific to individual crop types and therefore, each crop's model will have an associated imbalance in dataset size across all clients depending on the aggregation level chosen. As seen in the dataset imbalance study, dataset distribution imbalance does not necessarily reduce learning effectiveness, however in a real client scenario where the clients have data specific to their region, this may start to overfit the global model to specific regions with high crop areas such as large countries or farms. For example, the model will be specialised to certain countries where single farm holdings are very large, such as Canada and Australia rather than those of smaller single field farms, that actually make up the most of all holdings²¹. Therefore, addressing the imbalance in dataset size becomes crucial, and selecting the appropriate aggregation level can effectively achieve this. Crops with the lowest standard distribution in Fig. 10, should choose to operate on micro

aggregation levels, whereas when approaching a standard distribution of 0.1-0.2, meso and macro aggregation levels will become preferable. Again, this effect can be reduced if individual global models are trained for each continent, subcontinent or biome, however this will require smaller aggregation levels. Therefore, it will be preferable for some crops to operate on micro aggregation levels whereas other crop types may require meso or macro aggregation levels or individual global models for different global regions. Class imbalances will also occur across a global dataset due to different crops being grown in different quantities in different regions due to climate and culture. For example hotter climates grow maize in larger quantities and root vegetables are grown more frequently in colder climates. This means that many smaller, biome dependent, datasets will be preferred to produce better training.

The biggest problem is lack of labelled per field data for crop yield. Currently very little exists publicly over large enough areas to train a model. Until such data is made public, one of the only ways to produce a model is by the federated learning approach discussed. Without this data, testing and validating the model is also difficult without a large enough sample size of many regions across the globe. When acquiring ground truth data labels in a centralised or federated scenario labels will inevitably have different qualities. This issue may arise from accidental mislabelling but could also have more epistemic origins. It could stem from challenges at the micro-aggregation level, such as farmers facing difficulties with technology adoption, or from specific regions producing inaccurate or unreliable data. Standardisation of labelling and fact checking would be more suited to the larger aggregation levels where organisations or nations could align data labels. In a decentralised federated learning process a further consideration of adversarial clients should be considered. To reduce the effect of attackers on such a network, early stopping mechanisms for specific clients could be used. If a certain client is providing weights or tested metrics that are outside a distribution of the mean of all other clients, this client could be removed from aggregation. This approach would enable learning to progress with clients that align with the average behaviour of all clients, potentially stabilising the learning process. This issue is prevalent in open decentralised networks and could be mitigated by transitioning to permissioned networks. However, this solution would further centralise the process. Alternatively, implementing consensus mechanisms across each aggregation level may offer a balance.

Crop yield problem complexity

In Table 5, and viewing Fig. 7, the the F1-scores show that some classes are more difficult to classify. This may be due to the class imbalance across the dataset, but also could be due to the difficulty of identifying certain crops. If identifying certain crops is difficult, it will be even more difficult to estimate their yield. In this study only nine optical bands from Sentinel-2 L2A are used to attain sufficiently accurate models to study the effect of the number of clients in the federated learning scheme. Crop yield predictions will be a substantially more difficult problem. More modalities of data, such as synthetic aperture radar, will be required to detect parameters including height, weather and ground water content of the crops, landscape and atmosphere. The other major complexity is crop yield prediction takes place over substantial amounts of time. Monitoring crops from cultivation and planting through to harvest requires large amounts of temporal data.

Applying additional models, such as those for land cover and crop type to identify specific crop locations, along with cloud masking models, will also introduce their own inaccuracies into crop yield prediction models. Due to all these difficulties, large amounts of parameter optimisation would be required for optimisation of a global federated learning model, therefore, making smaller biome dependent models and using smaller aggregation levels more effective in attaining better results.

Privacy concerns

When considering smaller dataset sizes, differential privacy starts to fluctuate quickly between high-utility-low-privacy and low-utility-high-privacy. Therefore if small client datasets with high privacy exist, little to no information can be learned from these clients as the differential privacy noise applied can reduce any meaningful data. An increased dataset size and therefore larger aggregation level could help reduce these fluctuations between utility and privacy.

Further considerations

Decentralised data propagation due to larger data

In federated learning, large data volumes pose a significant challenge. Globally, there is 16 billion hectares of agriculture land, dividing that into 10 m^2 (same as the Sentinel data used in this work) gives, 1.6×10^{12} cells. This works out at approximately 0.37TB per band of data globally. This comes from the size of the crop type dataset in PNG file type, divided by the number of patches in the crop type dataset, multiplied by the number of global cells. As discussed previously, many extra bands, modalities and temporal data are required, potentially creating very large datasets for crop yield. Transmitting this entire dataset between clients and a central server is impractical due to the high communication overhead. To address this, federated learning optimises data sharing by transmitting model weights instead of raw datasets. Using decentralised protocols to manage these exchanges allows the system to handle large data efficiently while minimising computational and communication burdens. A trade off is made here where reducing the aggregation level will increase the number of weights distributed on the entire decentralised network however this could be larger in size than that of all the patches being used for training.

Computational requirements due to difficulty of problem

In the case of micro level aggregation, in federated learning, the computational requirements for running local training on farmers' devices and conducting decentralised aggregation can be substantial. Farmers, particularly those in rural areas, often rely on relatively modest computing resources, which may not be sufficient for the

intensive computational tasks involved in local model training. For instance, the local training process, requires significant processing power and memory. This concern is addressed in part by¹³ focusing on optimising federated learning algorithms to reduce their computational and communication overhead. Additionally, decentralised aggregation, where multiple clients contribute to the model updates without a central server, introduces further complexity. Different aggregation schemes, such as federated averaging on all client models simultaneously versus more advanced approaches like daisy chaining models in¹⁹, can significantly affect the computational power and data storage requirements of each client. These issues would diminish if aggregation levels such as Meso, Macro, and Mega are employed, as these levels involve fewer clients and each client possesses greater computational resources.

Conclusion

To use EO data in the context of crop yield prediction modelling is difficult due to the many modalities of data and the limited extent of data. This data is generally private or proprietary and limited examples exist publicly. To overcome such data privacy issues decentralised federated learning is introduced as a solution to the problem. However, decentralised federated learning is affected by the number of clients in the learning process as well as the distribution of data between each client, number of classes and imbalance of classes.

This paper has shown that increasing the aggregation level to fewer clients with more data reduces the impact of these problems. Balance of dataset size across different clients is also an issue where even distributions do not always produce the best learning but could produce skewed learning to specific parts of the world. The labelled datasets that do exist publicly for crop yield prediction at the farmer level are too unbalanced to provide meaningful weights to the global federated learning model. Increasing aggregation level to that of the described Meso, Macro or Mega levels is again a solution to this problem. When looking at dataset privacy mechanisms, systems such as differential privacy also break down with small datasets such as those on a per farm basis. Solutions to these three problems of number of clients, dataset balance, and privacy utility all lead to a more centralised aggregator being more effective for learning. Therefore, the more centralised the more effective the learning becomes at the trade off of privacy. Multiple, biome or region specific, models could be a compromise between these two problems.

Future research in this area could explore how different model architectures and sizes of model, are influenced by the number of clients involved. Another valuable direction would be to train an EO crop yield prediction model using private datasets, should they become available. Looking at generalisation difficulties across different biomes and time frames will also be an important step towards a global crop yield model. Understanding the difficulty of training a model for each crop type, to determine if dataset imbalance is the cause of some crops not being identified or if these crops are intrinsically difficult to identify from EO data will be a critical step for some crop yield models. Finally, studying the impact of differential privacy across datasets of different sizes and understanding how this affects the learning process could provide important insights.

Data availability

The datasets provided as part of this work are public at <https://huggingface.co/0x365>.

Code availability

The code associated with this work exists in a public repository at <https://github.com/strath-ace/smart-dao/>.

Received: 13 August 2024; Accepted: 12 March 2025

Published online: 26 March 2025

References

- Jeroen Degerickx, S. G. & Kristof Van Tricht. *Worldcereal: Seasonally updated global crop mapping*. Tech. Rep., ESA (2023).
- Dwork, C. et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
- Martins, P., Sousa, L. & Mariano, A. A survey on fully homomorphic encryption: An engineering perspective. *ACM Comput. Surv.* **50**, 66. <https://doi.org/10.1145/3124441> (2017).
- Antico, T., Moreira, L. & Moreira, R. Evaluating the potential of federated learning for maize leaf disease prediction. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional* 282–293 (SBC, 2022). <https://doi.org/10.5753/eniac.2022.227293>.
- Ahmed, R. et al. *Federated Learning-Based UAVs for the Diagnosis of Plant Diseases*. <https://doi.org/10.1109/ICEET56468.2022.1007133> (2022).
- Idoje, G., Dagiuklas, T. & Iqbal, M. Federated learning: Crop classification in a smart farm decentralised network. *Smart Agric. Technol.* **5**, 100277 (2023).
- Žalik, K. R. & Žalik, M. A review of federated learning in agriculture. *Sensors*. <https://doi.org/10.3390/s23239566> (2023).
- Demers, A. et al. Epidemic algorithms for replicated database maintenance. In *Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing* 1–12 (1987).
- Jeliasy, M. *Gossip* 139–162 (Springer, 2011).
- Zhao, Y. et al. Federated learning with non-iid data. arXiv preprint [arXiv:1806.00582](https://arxiv.org/abs/1806.00582) (2018).
- Li, X., Huang, K., Yang, W., Wang, S. & Zhang, Z. On the convergence of fedavg on non-iid data. arXiv preprint [arXiv:1907.02189](https://arxiv.org/abs/1907.02189) (2019).
- McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In Singh, A. & Zhu, J. (eds.) *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54 of *Proceedings of Machine Learning Research* 1273–1282 (PMLR, 2017).
- Kairouz, P. et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14**, 1–210 (2021).
- Fraboni, Y., Vidal, R., Kameni, L. & Lorenzi, M. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research* 3407–3416 (PMLR, 2021).
- Rai, S., Kumari, A. & Prasad, D. K. Client selection in federated learning under imperfections in environment. *AI* **3**, 124–145. <https://doi.org/10.3390/ai3010008> (2022).

16. Fraboni, Y., Vidal, R., Kameni, L. & Lorenzi, M. A general theory for client sampling in federated learning. In *International Workshop on Trustworthy Federated Learning* 46–58 (Springer, 2022).
17. Ribero, M. & Vikalo, H. *Communication-efficient federated learning via optimal client sampling*. arXiv preprint [arXiv:2007.15197](https://arxiv.org/abs/2007.15197) (2020).
18. Chen, W., Horvath, S. & Richtarik, P. *Optimal client sampling for federated learning*. arXiv preprint [arXiv:2010.13723](https://arxiv.org/abs/2010.13723) (2020).
19. Kamp, M., Fischer, J. & Vreeken, J. *Federated learning from small datasets*. arXiv preprint [arXiv:2110.03469](https://arxiv.org/abs/2110.03469) (2021).
20. Von Braun, J. *Small-scale farmers in liberalised trade environment* 21 (2004).
21. Lowder, S. K., Sánchez, M. V. & Bertini, R. Which farms feed the world and has farmland become more concentrated?. *World Dev.* **142**, 105455. <https://doi.org/10.1016/j.worlddev.2021.105455> (2021).
22. Union, E. *Statistical Regions in the European Union and Partner Countries* (2020).
23. Nations, U. *United Nations* (2024).
24. Boogaard, H. et al. WorldCereal open global harmonized reference data repository (CC-BY licensed data sets). <https://doi.org/10.5281/zenodo.7593734> (2023).
25. Cowlshaw, R. *Crop Type Segmentation Earth Observation Dataset of Belgium* (2024).
26. Sinergise, C., ESA. *Sentinel 2—Level 2A scenes and metadata* (2018).
27. Victoria, A. A brief guide to estimating crop yields (2024).
28. for Applied Systems Analysis (IIASA), I. I. & (WENR), W. E. S. G. *Worldcereal Reference Data Module* (2024).
29. Orynbaikyzy, A., Gessner, U., Mack, B. & Conrad, C. Crop type classification using fusion of sentinel-1 and sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies. *Remote Sens.* **12**, 66. <https://doi.org/10.3390/rs12172779> (2020).
30. Aiyim Orynbaikyzy, U. G. & Conrad, C. Crop type classification using a combination of optical and radar remote sensing data: A review. *Int. J. Remote Sens.* **40**, 6553–6595. <https://doi.org/10.1080/01431161.2019.1569791> (2019).
31. Tseng, G., Zvonkov, I., Nakalembe, C. L. & Kerner, H. CropHarvest: A global dataset for crop-type classification. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).
32. CLAAS. *Yield Mapping* (2024).
33. Murphy, D. P., Schnug, E. & Haneklaus, S. Yield mapping—a guide to improved techniques and strategies. In *Site-Specific Management for Agricultural Systems* 33–47 (Wiley, 1995).
34. eurostat. *Crop production in EU standard humidity by NUTS 2 regions*. https://doi.org/10.2908/APRO_CPSHR (2024).
35. Franch, B. et al. Global crop calendars of maize and wheat in the framework of the worldcereal project. *GIScience Remote Sens.* **59**, 885–913 (2022).
36. Weng, W. & INet, X. Z. *Convolutional networks for biomedical image segmentation*, 2021, 9. 10.1109/ACCESS.16591-16603 (2021).
37. Qi, P. et al. Model aggregation techniques in federated learning: A comprehensive survey. *Future Gen. Comput. Syst.* **150**, 272–293. <https://doi.org/10.1016/j.future.2023.09.008> (2024).
38. Nilsson, A., Smith, S., Ulm, G., Gustavsson, E. & Jirstrand, M. A performance evaluation of federated learning algorithms. In *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning* 1–8 (2018).
39. Sun, T., Li, D. & Wang, B. Decentralized federated averaging. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4289–4301. <https://doi.org/10.1109/TPAMI.2022.3196503> (2023).
40. Kingma, D. P. & Ba, J. *Adam: A method for stochastic optimization*. arXiv e-prints [arXiv-1412](https://arxiv.org/abs/1412) (2014).
41. GmbH, F. L. *Flower a friendly federated learning framework* (2024).
42. GmbH, F. L. *Flower dataset partitioners* (2024).
43. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255. <https://doi.org/10.1109/CVPR.2009.5206848> (2009).

Author contributions

R.C. conceived, conducted experiment and analysed results. N.L. and A.R. edited and provided feedback throughout the process. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025