




A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D5.4: Block-chain based distributed learning and AI passport

Reference	D5.4_ EuCanImage_UB_v1
Lead Beneficiary	UB
Author(s)	Akis Linardos, Socayna Jouide, Alexandros Tragkas, Davide Zaccagnini
Dissemination level	Public
Type	Report
Official Delivery Date	30 September 2022
Date of validation of the WP leader	30 September 2022
Date of validation by the Project Coordinator	30 September 2022
Project Coordinator Signature	

EuCanImage is funded by the European Union's H2020 Framework
Under Grant Agreement No 952103

1. Version log

Issue Date	Version	Involved	Comments
10 Aug 2022	V0.1	Akis Linardos, Socayna Jouide	First draft of the FL framework
15 Sep 2022	V0.2	Alexandros Tragkas & Davide Zaccagnini	First draft of the block-chain section
26 Sept 2022	V0.3	Kaisar Kushibar & Anais Emelie	Feedbacks
30 Sept 2022	V0.4	Davide Zaccagnini	Revised version
30 Sept 2022	V1	Anais Emelie, Karim Lekadir & Kaisar Kushibar	Revised and corrected final version.

2. Executive Summary

Distributed data and algorithmic systems are rapidly establishing themselves as the next digital framework thanks to their more efficient operations, security and privacy attributes. EuCanImage is adopting this paradigm in data storage and the development of AI through a federated learning infrastructure (T5.4). These frameworks envision human operators running computations over locally stored data to train and validate AI agents and, in future scenarios, the use of AI tools as network services exposing their functions in support of medical decision making or other clinical processes. In this view, the present task has also designed and developed a blockchain-based provenance and permissioning system to equip AI algorithms with unique metadata schemas that, on one side will allow to trace and validate authorship and other type of attributes such development status, intended use etc and, on the other, to orchestrate their use, or their training on target data sources under ethically and legally binding permissions enforced by the blockchain.

This work stems from Lynkeus' experience and previous developments in the area of blockchain-based permissioning which began with the MyHealth-MyData project and was further developed in the euCanSHare and Kraken initiatives and University of Barcelona (UB) new development on federated learning infrastructures. The Lynkeus blockchain has been configured and deployed into a public hosting environment, configured to implement a generic data and network governance protocol and exposed to the EuCanImage infrastructure via REST APIs. Future work on this component, as AI tools developed during the project become available for wider distribution, will include the assignment of AI passports and the testing of distributed transactions over the EuCanImage network.

Table of Contents

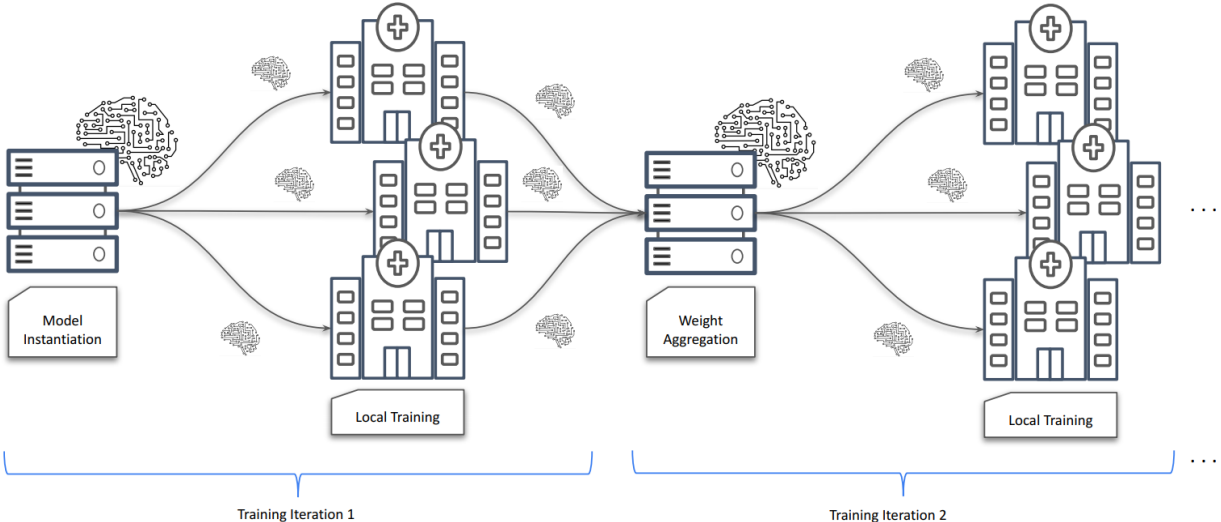
1.	Version log	2
2.	Executive Summary	2
1	Federated Learning Framework	4
1.1	Technical Design	4
1.1.1	The Flower Library	4
1.1.2	Containerization	4
1.1.3	Pipeline Description	5
1.2	Simulation Experiments	6
1.2.1	Datasets (5 simulated centres)	6
1.2.2	Comparative Analysis Experiments: Motivation and Results	6
2	Addressing Individual Centres	7
2.1	Addressing the Ethical Committees - Requirements and Constraints	7
2.2	Deployment Tests	8
3	Block-chain and AI passport	9
3.1	Blockchain Architecture	9
3.1.1	Distributed, permissioned data governance	9
3.1.2	Implementation strategy	9
3.2	The AI Passport	10
3.2.1	AI Passports as Smart Contracts	10
3.2.2	Ensuring AI Agents Authenticity	11
4	Conclusions	11

Acronyms

Name	Abbreviation
Federated Learning	FL
Collaborative Data Sharing	CDS
Graphical Processing Unit	GPU
Virtual Private Network	VPN
The Cancer Imaging Archive	TCIA
Membership Service Provider	MSP
Certificate Signing Requests	CSP
Distributed Ledger Technology	DLT
Certificate Signing Requests	CSR
University of Barcelona	UB

1 Federated Learning Framework

Federated Learning, first introduced by Google in 2017, alleviates security and privacy risks of data transfer by allowing training models in a decentralised manner. Essentially, federated learning reverses the information exchange between collaborating parties: instead of pooling data to one central server, the server defines an AI model which is then distributed across data centres. Training occurs locally, and the models are aggregated on the central server after a defined number of epochs (which we will be referring to as EPR: Epochs Per Round) and the process is repeated for a pre-defined number of federated rounds (FR).



1.1 Technical Design

In terms of design choice, the most important factor is the feasibility of deployment across hospitals. A major constraint when establishing a network is to bypass the firewalls of each Centre. Commonly, a VPN is used to verify secure communication between the collaborating parties. Another predictable pitfall is that of systems heterogeneity—i.e. each Centre works with different machines and as such will require a way for our framework to work irrespective of the system variability. Note that it is nonetheless advisable for the centres to coordinate their hardware as best they can as there are aspects (such as the requirement of a powerful GPU to run deep learning models) that cannot be circumvented.

1.1.1 The Flower Library

UB's research led to choosing the Flower library to base the core architecture. It's an open-source library, which allows free usage of all their modules. A major bonus to using this library is the pull mechanism—i.e. instead of the server making requests from each collaborating Centre, it's the centres that request the updated model at each federated round. This has the significant advantage of bypassing firewall constraints without requiring a VPN, as the centres are never pinged by the central server.

1.1.2 Containerization

A significant challenge to deploying federated networks is the heterogeneity in terms of both hardware and software. To overcome this hurdle, we use container technologies, which wraps our code in a package that installs its own dependencies irrespective of the system it's running on. This is the deployable part of the code.

1.1.3 Pipeline Description

Following the architecture of [Flower](#), our pipeline, available at [Gitlab](#), works as displayed in figure 1.

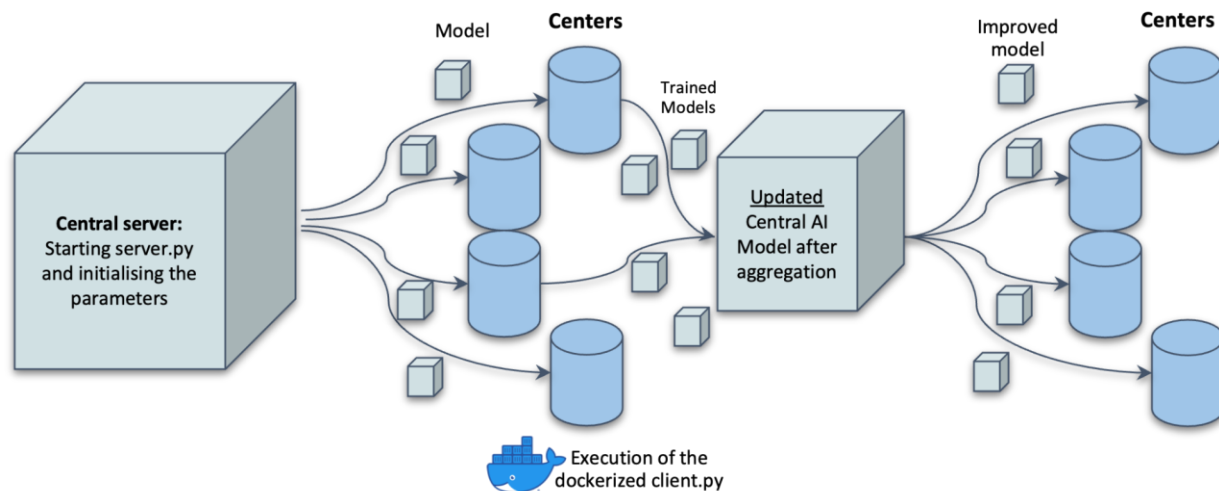


Figure 1. Federated-learning framework. A script called `server.py` is initiated on the central server. The model is defined and its parameters are initialised there along with all the parameters of the federated learning pipeline. Federated learning parameters include the number of rounds to run, and how many epochs (iterations over the whole data of each Centre) will take place each round. Once the `server.py` is initiated, it listens for incoming requests from collaborating centres.

On the Centre side, a script called `client.py` (which has been containerized either by Singularity or Docker depending on the Centre's individual needs) will be run and request the model from the central server to initiate the training locally. This script loads several modules, including a data pre-processing module that rescales the local data, performs histogram matching to pre-defined landmarks.

```
(fl_env) user@eucanimage-dev:~/BFP/src$ cat nohup.out
Using cache found in /home/user/.cache/torch/hub/pytorch_vision_v0.10.0
INFO flower 2022-08-10 16:06:53,292 | app.py:77 | Flower server running (insecure, 120 rounds)
INFO flower 2022-08-10 16:06:53,292 | server.py:118 | Initializing global parameters
INFO flower 2022-08-10 16:06:53,292 | server.py:304 | Requesting initial parameters from one random client
INFO flower 2022-08-10 16:07:30,391 | server.py:307 | Received initial parameters from one random client
INFO flower 2022-08-10 16:07:30,391 | server.py:120 | Evaluating initial parameters
INFO flower 2022-08-10 16:07:30,391 | server.py:133 | FL starting
DEBUG flower 2022-08-10 16:08:09,773 | server.py:251 | fit_round: strategy sampled 5 clients (out of 5)
DEBUG flower 2022-08-10 16:09:54,563 | server.py:260 | fit_round received 5 results and 0 failures
Using cache found in /home/user/.cache/torch/hub/pytorch_vision_v0.10.0
WARNING flower 2022-08-10 16:09:55,867 | aggregator.py:338 | No fit_metrics_aggregation_fn provided
DEBUG flower 2022-08-10 16:09:56,091 | server.py:201 | evaluate_round: strategy sampled 5 clients (out of 5)
DEBUG flower 2022-08-10 16:10:56,277 | server.py:210 | evaluate_round received 5 results and 0 failures
Generating experiment 0067_TRIAL_Averaging
Length results is 5
{'accuracies_aggregated': [], 'total_val_loss': [], 'time_spent': []}
Saving round 1 aggregated_parameters...
[0.87, 0.91, 0.89, 0.88, 0.86]
Round 1 accuracy aggregated from client results: 0.882
```

Figure 2. Federated-learning framework test. The example consists of one server and five clients. Clients are responsible for generating individual weight-updates for the model based on their local datasets. These updates are then sent to the server which will aggregate them

to produce a better model. Finally, the server sends this improved version of the model back to each client. A complete cycle of weight updates is called a round. Furthermore, we display the server logs, where we can see the progress of the execution for the different clients and finally the accuracy value after aggregation (0.882).

1.2 Simulation Experiments

Before deploying the FL network in the challenging set-up of the real world, we first perform the vast majority of our experiments in a simulation—where one machine behaves as both the central server and the collaborating centres. This is done using open-source data we have available on UB. The task we are solving is Breast Cancer Screening—i.e. classifying Benign versus Malignant tumours which addresses Use Case 8 of EuCanImage.

1.2.1 Datasets (5 simulated centres)

We use 4 datasets, 3 of which were obtained as part of the EuCanImage project (OPTIMAM, InBreast, BCDR) and an additional open-source dataset available from TCIA called CMMD. OPTIMAM is the only multi-centric dataset among those, containing 3 centres in total. However, as one of the centres contains only malignant cases, it’s impossible to train locally without overfitting our model to always predict malignant cases. As such we use two centres from OPTIMAM (stge and jarv). The total number of patients and the representation of each class is shown in the table below:

	Total Patients	Total Images	Benign Images	Malignant Images
OPTIMAM (stge)	2468	15197	5350	9847
OPTIMAM (jarv)	3171	19591	4190	15401
InBreast	134	342	242	100
BCDR	229	544	439	105
CMMD	1774	5198	1108	4090

1.2.2 Comparative Analysis Experiments: Motivation and Results

Our first step was to compare state-of-the-art classification networks on the task of Breast Screening. We used models of small parametric size that have shown great performance in the task of ImageNet classification (EfficientNetB0) moving up in parametric size to DenseNet121, ResNet18 and ResNet50. We used two set-ups: In one we pool all the data as if it were coming from one Centre (Collaborative Data Sharing, or CDS) and ran 120 iterations over the whole data (epochs) while in the other set-up we considered each dataset as a separate Centre and run 120 federated learning rounds over it. We used the traditional Federated Averaging, but also Federated Median to investigate how the elimination of outliers in the weight aggregation affects performance overall.

Our results show that our federated approach not only reaches the “golden standard” of CDS but also overcomes it:

	ResNet50			ResNet18			DenseNet121			EfficientNetB0		
	CDS	FL Med	FL Avg	CDS	FL Med	FL Avg	CDS	FL Med	FL Avg	CDS	FL Med	FL Avg
InBreast	0.9190	0.9003	0.8618	0.8880	0.8900	0.8363	0.8984	0.9309	0.6522	0.9530	0.8363	0.7481
BCDR	0.9197	0.9282	0.9166	0.9048	0.9514	0.9144	0.8647	0.9838	0.8079	0.9523	0.9144	0.7338
OPTIMA M (stge)	0.9323	0.8434	0.9499	0.9387	0.7356	0.9482	0.9259	0.7397	0.8970	0.9488	0.9483	0.8994
OPTIMA M (jarv)	0.9196	0.8324	0.9497	0.8946	0.7535	0.9362	0.8975	0.7531	0.9539	0.9439	0.9362	0.9398
CMMD	0.9120	0.9372	0.9287	0.9140	0.9381	0.9163	0.9069	0.9072	0.8751	0.9113	0.9163	0.8812
Average	0.9205	0.8883	0.9213	0.9080	0.8537	0.9103	0.8987	0.8629	0.8372	0.9419	0.9103	0.8405

2 Addressing Individual Centres

The teams conducted several meetings to update the collaborating clinical centres and help them understand the intricacies of coordinating Federated Learning experiments across a real world network. At the same time, we received feedback and gained a better understanding on what is needed from our side and what we have to provide.

2.1 Addressing the Ethical Committees - Requirements and Constraints

We drafted a 10-page long document to address the Ethical Committees of all centres. Specifically we addressed the following requirements from the ethical committees:

1. A clear definition of the task—i.e. Breast Cancer Screening.
2. Ascertained we under no circumstances have direct access to the data provided.
3. Provided a clear request of the data relevant to the task at hand—i.e. Mammograms of breast cancer patients with their meta data defining the tumour as malignant or benign.
4. Provided a work hypothesis and objectives of our investigation:
 - a. To develop more efficient federated algorithms for distributed learning across hospitals and institutions.
 - b. To study and develop fair AI models by integrating data from centres across the world and with varying data sizes.
 - c. To develop a state-of-the-art model for breast cancer diagnosis and open-source this model and allow its usage by future research
5. Provided a methodology and study design:
 - a. The mammogram imaging data of the participants is grouped into the clinical variables “benign”, and “malignant”. The AI model in this study is tasked with identifying this clinical variable from a mammogram image. To do so, the model is distributed across centres and trained locally on the mammograms and clinical variables. Once trained, the AI models return to a central server

- (Barcelona SuperComputing Centre), where they are aggregated. The process is then repeated.
- b. The deploy federated learning algorithm adheres to a “privacy-by-design” principle: A computing network connects the central server (hosted in Barcelona Supercomputing Centre) to hospitals and institutes around the world. In this setup, a data scientist connects to the central server and distributes the model across clinics, while the data remains at the hands of its owners.
6. Declared who are the Intended Users of Data:
 - a. The User of the data is only the local AI model. Local means that this AI model is trained inside the clinical centre. Only the trained models are sent to the central server, which is hosted at the University of Barcelona.
 7. Biases derivable from the variables:
 - a. The entirety of the collected mammogram imaging data can contain biases towards particular patient populations. In the federated setup of the present study, this effect is reduced by the aggregation of AI models from different centres from various regions and underlying patient populations

Furthermore, we outlined the requirements of our project both in terms of Hardware and Data.

The main constraint is our requirement of NVIDIA GPUs which is a prerequisite for our deep learning models to run successfully. These have been addressed across all centres by now, although some installation is pending.

2.2 Deployment Tests

To make sure our framework works across all collaborating parties, we first used a subset of our own datasets to run local experiments. A subset from CMMD in particular was used—which is already publicly available by TCIA and bears no constraint to share—and distributed across all collaborating parties once they declared their hardware was ready. We have thus far successfully tested 4 centres and verified our framework works correctly. Once the rest of the centres have addressed the remaining concerns we will proceed with a full-scale evaluation of our best performing models.

Affiliation	Hardware/OS/Network progress	Network Checked
Universitat de Barcelona	Ready	✓
Hospital Germans Trias i Pujol	Ready	✓
Hospital Parc Taulí	Ready	✓
Maastricht University	Pending	
Medical University of Gdańsk	Ready	✓
Hospital Italiano de Buenos Aires	Ready	✓
Erasmus MC	Pending	
La Fe University and Polytechnic Hospital	Ready	✓
Aristotle University Thessaloniki	Ready	✓

3 Block-chain and AI passport

Blockchain is a type of Distributed Ledger Technology (DLT) which tracks network events in a peer-to-peer transparent, and immutable manner though practically unbreakable cryptographic schemes realising a trust-less system that incentivizes each participant to make the right decision and not act with malicious intent. In this sense, it was deemed ideal for the needs of the EuCanImage project which is implementing a distributed data infrastructure spanning different jurisdictions, data types, actors and geographies.

3.1 Blockchain Architecture

We used Hyperledger, from the Linux Foundation, because of the highly versatile and modular infrastructure for enterprise use cases, in particular for the flexible permissioning framework that was needed within EuCanImage, especially toward hospitals and clinical data centres. Fabric allows for pluggable and extendible consensus mechanisms which we utilised to encode key terms of the GDPR, to capture and enforce informed consent terms and hospital data policies. Fabric also provides the ability to customise the network in terms of membership authorization and access control management thus providing, in the later part of the project, the ability to set the data and network governance framework. This will become a key feature as the platform will be open to third party users and data providers as it will offer an explicit, transparent set of protocols governing data transactions on the platform.

3.1.1 Distributed, permissioned data governance

In a HLF network every entity is an organisation and every member making transactions must belong to one. In EuCanImage, an organisation can be a hospital, a research lab or any entity belonging to the consortium. In future expansion of the platform, third parties will be enrolled as organisations as well. Each organisation can set up specific roles for their members. Technically, an organisation is represented by a folder with crypto material called Membership Service Provider (MSP). The MSP is the means by which validation is performed when organisations or members transact. Lastly, organisations form consortiums, set consensus policies, and decide on the network configuration.

More specifically, the governance controlling how data will be accessed in the federated learning transactions, have been set in the EuCanImage "channel" a sub-network inside the HLF network which allows for private communication between EuCanImage organisations. HLF offers the possibility of creating additional channels as needed and this option will be investigated further, for instance, in the case of US-EU transactions involving UAMS because of the specific legal constraints in that scenario. Key to the functioning of a channel are smart contracts which encode data access control policies.

3.1.2 Implementation strategy

Blockchains are computationally intensive system which do not excel in time performance, for this reason we developed a cache database as an off-chain utility to provide faster performance as well as greater flexibility for query handling. This will be crucial when audits on given AI tools will be required using blockchain data.

Our Cache Database was implemented as a MongoDB instance which replicates the data on the ledger via block event listening. As most of the blockchain platforms with smart contracts, HLF allows setting events inside the contracts to notify applications of updates on the ledger. Thus, a block listener updates the database with the data associated with the event and is also used to provide updates on the client side of the application.

3.2 The AI Passport

In this context, Lynkeus and UB started gathering from the consortium AI development teams, high level requirements for the creation of the AI passport, a digitally signed smart contract designed for AI Algorithm serving as a secure, transparent and reliable identification and tracking systems to train, validate and the deploy AI tools over distributed environments. In this view, and after multiple feedback gathering sessions with AI team, the AI Passport was designed as a permanent, public but extensible data schema carrying not only the signature of the AI developer and his/her institution but also a provenance record, as indicated below, to be stored inside the HLF blockchain. Any user of the network will, in this way, have the ability to view the entire set of published AI passports referring to either developing or completed, follow the progress of these algorithms, including performance metrics.

The AI Passport, in this view, is stored as an object inside a Smart Contract and only the creator has permission to modify their Passport.

3.2.1 AI Passports as Smart Contracts

The EuCanImage blockchain now hosts prototypical versions of AI Passport in the form of Smart Contracts embedded objects. While the initial version of the AI passport it's still limited in scope, additional parameters will be added as AI tools start being deployed on the federated learning infrastructure and training of these agents begin. The initial version of the passport's parameters list is shown below and it supports querying, storing, modifying, getting the history of the AI tool directly from the blockchain ledger.

AI Passport
ID
Owner
Description
TrainingType
DatasetIDs
Train/Validation Details (metrics, flagged biases, etc. These are currently being established in WP6)

3.2.2 Ensuring AI Agents Authenticity

HLF provides authentication mechanisms built upon a Certificate Authority (CA) architecture¹. We leveraged this functionality to ensure that any user will be able to transact as a member of the network after receiving a certificate (X509 type) by one of the consortium's CA. As the federated learning infrastructure becomes operational, CA will be identified among consortium members, besides Lynkeus and UB which have already been elected to such a role. Subsequently, all transactions will be validated by the network using the Public Key Infrastructure (PKI) and in this way, through native control logic integrated into smart contracts, the network will ensure that only the AI algorithm provider can modify the algorithm, and indeed, the passport itself.

To this extent, the task teams investigated at length not only the Digital Signature Service (DSS)² from the Connecting Europe Facility, but also a variety of commercial and other open-source e-signature systems which did not offer additional benefits compared to the HLF functions in regard to handling transaction signing. The teams, however, believe that as these frameworks evolve, they will become relevant to the project and will therefore be kept under close observation.

4 Conclusions

The blockchain network deployed and configured during this task is now ready to support both federated learning and the distributed use of AI agents in the EuCanImage project offering transaction transparency, provenance tracking, and authenticity controls while inspiring AI developers to engage with the distributed ecosystems this infrastructure supports. This implementation will be of great importance as EuCanImage scales into an open platform for both imaging data and algorithms attracting more data providers and developers in the second part of the project.

¹ [Identity — hyperledger-fabricdocs main documentation](#)

² [Digital Signature Service - DSS \(europa.eu\)](#)