Interdisciplinary Description of Complex Systems 20(2), 162-178, 2022

FRAMEWORK FOR FEDERATED LEARNING OPEN MODELS IN E-GOVERNMENT APPLICATIONS*

Emanuel Guberović[1,] **, Charalampos Alexopoulos[2], Ivana Bosnić[1] and Igor Čavrak[1]

[1]University of Zagreb, Faculty of Electrical Engineering and Computing
Zagreb, Croatia

[2]University of the Aegean
Athens, Greece

DOI: 10.7906/indecs.20.2.8
Regular article

Received: 16 January 2022.
Accepted: 25 April 2022.

ABSTRACT

Using open data and artificial intelligence in providing innovative public services is the focus of the third generation of e-Government and supporting Internet and Communication Technologies systems. However, developing applications and offering open services based on (open) machine learning models requires large volumes of private, open, or a combination of both open and private data for model training to achieve sufficient model quality. Therefore, it would be beneficial to use both open and private data simultaneously to fully use the potential that machine learning could grant to the public and private sectors.

Federated learning, as a machine learning technique, enables collaborative learning among different parties and their data, being private or open, creating shared knowledge by training models on such partitioned data without sharing it between parties in any step of the training or inference process. This paper provides a practical layout for developing and sharing machine learning models in a federative and open manner called Federated Learning Open Model. The definition of the Federated Learning Open Model concept is followed by a description of two potential use cases and services achieved with its usage, one being from the agricultural sector with the horizontal dataset partitioning and the latter being from the financial sector with a dataset partitioned vertically.

KEY WORDS

e-Government, open data, machine learning, federated learning open model

CLASSIFICATION

JEL: D80

*This is the extended version of the abstract published in: Vujić, M. and Šalamon, D., eds.: Book of abstracts of the National Open Data Conference. University of Zagreb, Faculty of Traffic and Transport Sciences, Zagreb, 2021.

**Corresponding author, η: emanuel.guberovic@gmail.com; -; -

INTRODUCTION

Utilization of Internet and Communication Technologies (ICT) by various governments worldwide to supply its citizens and other interested parties a whole new plethora of capabilities centered around the data and services that fall within its domain is known as electronic government (or shorthand e-Government). Six distinct governance properties experience improvements by using e-Government activities, including quality of public services, administrative efficiency, open Government (OG) capabilities, ethical behavior and professionalism, trust and confidence in government, and social value and well-being [1]. Three different generations of e-Government [2], differ in their final goals and essential ICT tools used to achieve them. The first generation focuses on informational and transnational services through ICTs and web technologies. The second generation focuses on improving openness and interoperability through web 2.0 concepts. Finally, the third generation aims to achieve innovative governance by exploiting disruptive technologies such as artificial intelligence.

Different governments utilize a plethora of varying open data policies [4], with a good potential found in all of them embracing further openness in increasing the participation and interaction of open data consumers and producers, resulting in positive results such as stemming corrupt behavior [5]. Simultaneously, there is a pronounced sense of privacy in personal data sharing resulting in numerous data protection regulations and acts appearing in recent years. The United Nations Conference on Trade and Development's (UNCTAD) publication on Data Protection regulations and international data flows [6] analyzed data protection laws that were current in the year 2016 (in e.g. GDPR [7]). It concluded a recognized set of core data protection principles in binding international and regional agreements and guidelines, including a limited and secure collection of personal data. Their enforcement poses a challenge to artificial intelligence usage because many of its applications owe their successful implementation to personal data used in training and inference of the models. Adherence to their requirements is a logical next step in the evolution of the implementation of machine learning in the cohabitation of ethical computing and intelligent services, as privacy is found to be one of the ethical guidelines for artificial intelligence [8].

In recent years, a new machine learning technique called federated learning (FL) has helped the field of artificial intelligence to abide by data privacy regulations. Standard machine learning aggregates data from different sources on a central server, the model training process takes part. The central learning principle partakes with different dataset instances firstly being aggregated on a single central point; in this way, the central dataset can be perceived as a per data source partitioned data shards database. On the other hand, FL is based on the distributed learning principle. Each data owner partakes in the training process with their local data shard. This process emphasizes transferring model parameters between respective data owners instead of sharing their data. Because data never leaves the data source, this method is private by design. FL is a machine learning method that elevates knowledge derived from one instance by aggregating individual latent values extracted through the training process of the crowd or multiple instances.

As an ICT e-Governance tool of the third generation, it allows using new technologies for accomplishing crowd intelligence that supports data-wise and evidence-based public services.

RELATED WORK

THIRD GENERATION E-GOVERNMENT

The primary objective of the research and practice in the domain of Digital Government (DG) is the exploitation of ICT in government and the provision of ICT-based services to their potential users: citizens, private and public companies, as well as public servants. However,

the change in needs (and expectations) of citizens and societies also mandates the evolution in capabilities offered by ICT - not merely restricted to increase in performance and the number of services offered, but by shifting the focus of DG - thus driving the evolution of the digital government domain itself. Two major factors influence the evolution of the DG domain; the first one is defined by the wider external environment (economic, social, and political), and the second one by its technological environment. Nevertheless, a common pattern can be identified when observing evolutions in the DG domain; the first step preserves the existing practices, processes, and services and merely automates/supports them through existing or innovative ICT. Only in the second step the existing practices and processes are incrementally transformed and/or completely new practices adopted, usually through incremental ICT-based improvements introduced by the government [9].

Big Data generated by the Internet of Things (IoT) and Open Government Data (OGD) movement, Blockchain Technologies (BCT), Artificial Intelligence (AI), and particularly Machine Learning (ML) algorithms are some of the technologies used for modernizing the previous services provided by all of the governments around the world [10]. As Scholl [11] argues, future trends in DG that include "*smart approaches, many of which are Data Science-based, rely on the use of Artificial Intelligence (AI) and Machine Learning (ML) in combination with big structured and unstructured data to identify patterns and predictive models, which inform and evaluate decisions of human actors or non-human actors in real-time*". The latest generation in the digital government domain, namely, e-Government 3.0, is described exactly like that: "*e-Government 3.0 refers to the use of new disruptive ICTs (such as big data, IoT, analytics, machine learning, AI), in combination with established ICTs (such as distributed technologies for data storage and service delivery), and taking advantage of the wisdom of the crowd (crowd/citizen-sourcing and value co-creation), for supporting data-driven and evidence-based decision and policy making*" [3]. Vast amounts of data collected and aggregated in government agencies represent a massive potential for employing machine learning and other artificial intelligence techniques, thus unlocking the potential of that data by constructing descriptive and predictive models invaluable in supporting and enhancing government decisions and policymaking.

Considering AI, it is a broader concept that could be described by smaller and specific concepts: big data, machine learning, and decision-making. Castro and New [12] argue that "*AI is a field of computer science devoted to creating computing machines and systems that perform operations analogous to human learning and decision-making*". So, it needs the final concept of "automated decision making" in order for an application to be described as an AI one (i.e. face detection, voice recognition, and autonomous vehicles). The rest of the applications could be characterized as ML ones. As Abbod et al. [13] mentioned, "*Learning can be used to train a machine, so that it optimizes its rule base in a model and then new parameters may be tested in that model*", so the machines can learn with no use of explicit programming. Machine learning is a set of techniques that provides knowledge to any user or machine based on probabilistic algorithms applied to specific data. The most common techniques are classification and regression trees; Neural Network (Multilayer Perceptor); Bayesian Neural Network; Support Vector Regression (SVR); K-nearest neighbor model (KNN) and Gaussian Processes.

In recent years, governments have increasingly outlined ML as a research priority for a better understanding of government's data and implementing more efficient government solutions [14]. When it comes to a government, ML algorithms can help in the identification of significant factors and not yet defined interrelations. As such, they can be used to decrease the complexity of social phenomena that are related to policy problems.

In the literature, ML is applied to a plethora of sectors and fields regarding also the nature of data. In the legal and policy sector, the research focuses more on the analysis of the text. It deals with

Natural Language Processing and text mining, which includes techniques like arguments, topics and rules extraction, clustering, similarity check, and sentiment analysis. This could be further applied to comments or whole texts in several domains like legal texts [15], consultation platforms [16], and social media [17] enhancing the democratic process through participation and better interpretation of the results or finding contradictions in a specific legal system. Furthermore, they are used to classify news [18] or detect fake news [19, 20].

Other fields include cybersecurity and in terms of finding the related research of a domain as well as in multiple business domains [21]. For example, the topic modeling and the collaborative filtering algorithms (ML algorithms) are often used for the improvement of users' experience and for revenue increasing [22, 23]. ML is also used for information extraction from raw data and it can be used for a variety of purposes (e.g. prediction, understanding) [24]. Predictive modeling is defined as the analysis of large data sets to make inferences or identify meaningful relationships that can be used to predict future events [25, 26]. ML techniques in predictive modeling are used for the analysis of both current and historical facts for predictions making either for future or unknown events. Furthermore, ML is applied in the concept of smart cities dealing with traffic prediction and transportation. Accurate traffic prediction based on machine and deep learning modeling can help to minimize the issue [27, 28] of the tremendous rise in traffic volume causing a series of serious problems in modern society's quality of life, such as traffic congestion, delays, increased CO pollution, higher fuel prices, accidents [29], etc.

The list is continuously growing as more applications are included in the healthcare, environment, food, education, and agricultural domains. However, a series of challenges exist in the utilization of ML in the DG domain. As it is highlighted in [30] there is a list of barriers towards the full exploitation of the ML power with two of them being the most important ones. The first one is the combination of various ML techniques towards the production of proper results. Different ML techniques need to be tested to check their performance [31]. The second one is the availability of data. In many cases, the collection of personal data, the ownership of personal data, are subject to General Data Protection Regulation preventing the realization of the benefits from their processing. Policies like GDPR protect the corresponding entities regarding personal or even sensitive data. The publication of such data entails the risk of leading to privacy and ethical issues [21]. Furthermore, ML also depends upon collecting and processing data from society. This data may be explicitly sensitive (e.g., racial origin, religion, health data, ethnic origin) [32]. There are ways of preventing this phenomenon by applying anonymisation techniques before data publishing. But data anonymization in itself is not a fool proof system, being prone to de-anonymization attacks [33]. Even more, with the exponential growth of open data, de-anonymization techniques could work better maximizing the privacy and ethical risks. Based on the lack of the availability of proper data, quality issues occur that in turn, decrease the quality and quantity of the whole ML system [34]. Thus, in many cases, equilibrium should be achieved between these two major barriers. In addition, there can be difficulties of gaining regulatory approval of accessing data (for instance in healthcare), or even lack of data (geographical data) in order for an ML system to be properly trained for quality results. One of the challenges in producing e-Government services built on FL is in ensuring fairness and reproducibility, which is well emphasized in a paper on an analysis framework suitable for governmental scenarios in FL applications [35].

OGD could partially tackle the data availability issue since in most cases the usage of private data knowledge could increase the ML performance. A new solution is needed in order to safeguard legal and ethical issues regarding access to specific data while in parallel increasing the performance of ML algorithms. Federated ML and the proposed framework is

moving towards this direction and by proposing a proper solution handling these barriers. This study describes and applies the framework at hand in two separate cases. The first use case revolves around a horizontally partitioned environment, with a goal of agricultural commodity price prediction by combining data from the EUROSTAT price index [36] and FAO product import/export dataset [37]. This data is partitioned on a country level, with each one being a distinct data unit. Using FLOM in this example allows individual producers to gain better information about the cost-effectiveness of producing each commodity. This new knowledge can be discovered without the need for producers to exchange their production cost data, often confidential. The second use case relies on the constructed dataset from the anonymized private data created for a loan approval task containing credit record data and some client-specific private data. By vertically separating the dataset into credit balance data and private data, we compare the gains achieved using FL with the knowledge extracted from the complete dataset versus using only the credit balance data.

OPEN MACHINE LEARNING MODEL INITIATIVES

Machine learning (ML) training data sets are stored in well-known data formats that include unstructured text formats, tabular text-based file formats, columnar data file formats, nested text file formats, binary text file formats, array-based formats, hierarchical data formats, language-specific formats, and various image, video, and document file formats [38].

When it comes to defining data models themselves, different ML frameworks use different formatting: TensorFlow uses protocol buffers [39], Keras models are stored as .h5 files [40] and both PyTorch and Scikit-Learn store models as pickled file formats [41].

By using language, framework and environment agnostic formats for defining ML models, they can be made more easily interoperable, facilitating adherence to open data attributes [42], thus making models open themselves. Formats for open models include common formats successfully implemented and used in previous years. Data Mining Group (DMG) pioneered the search for a common format for defining an open standard for defining ML model exchange types with their design of Predictive Model Markup Language (PMLL) [43] and newer Portable Format for Analytics (PFA) [44].

More recently an extensive work by different industry partners has been done in defining formats for language-agnostic neural network models exchange that include two distinct projects: Neural Network Exchange Format (NNEF) by Khronos Group [45] and Open Neural Network Exchange Format (ONNX) [46] originally authored by Facebook and Microsoft, now a Linux Foundation project.

PMML

PMML is an XML-based open standard for model interchange first developed by DMG in 1997, with the newest release, as of writing this paper, being version 4.4 released in November 2019.

PMML files are described within well-defined parts that include [47]:

- header: general information about the PMML document, including its description, copyright, and timestamp,
- data dictionary: definitions for all the possible fields used by the model, including a description of valid, invalid, and missing data,
- transformation dictionaries: definitions of user data mapping that include: normalization, discretization, value mapping, aggregation, and functions mapping,
- model(s): contains the definition of the models themselves that includes mining schema (per data dictionary), local transformations, targets, outputs as well as model-specific contents.

PMML currently supports 16 different model types combined into more complex ensembles. Furthermore, PMML models are fully interchangeable between different PMML-compliant systems, of which some of the most notable are the pmml package for R language [48] and jpmml [49] for SParkML.

PFA

PFA is a JSON-based open standard for model interchange also built by DMG, with the most current release dating to November 2015.

It is based on AVRO schemas for defining data types and encoding custom functions (actions) applied to inputs. The actions are built using a set of inbuilt functions and language constructs (such as control flow), essentially making PFA a mini functional math language with schema specification. On the other hand, PMML allows building model functionality using only a set of predefined models.

Open Data Group projects spearhead PFA implementation for full implementation for Java Virtual Machine (Hadrian), Python (Titus), and R (Aurelius). Unfortunately, both PFA and PMML currently lack support for standardized operators for describing deep learning models.

More recently, PFA models have been used in the Medical Informatics platform of the Human Brain Project [50] to achieve models built using medical data that are shared in an FL manner.

ONNX

ONNX was initially released as Toffee by Facebook (an interchange format between PyTorch and Caffe), with development later joined by Microsoft and now completely maintained as an open-source project. It uses protobuf as a data structure format and is built using the principle of computational acyclic graphs with built-in operators and standard data types. Each computational node has one or more inputs and outputs and a call to an operator. Definitions of the different operators are implemented externally to ensure that every framework supporting ONNX provides implementations of built-in operators.

Although relatively new, with its first release as ONNX in September of 2017, the project is actively developed, with the latest release being 1.9.0 dated to April 2021. The active development of ONNX is incremental to its success and adoption as it stays current with changes in the deep learning ecosystem of frameworks that support its format, of which some of the most notable include: TensorFlow, Keras, PyTorch, Caffe, and ScikitLearn.

NNEF

Kronos Group developed NNEF**,** initially released in December 2017, with the latest release date to July 2019. Although a similar project to ONNX, its main focus is on inference interchange, especially with a focus on edge devices. NNEF standard is by definition less frequently evolving with its governance done by a multi-company group.

Technically the main differences between the two standards include using structure definition in a text-based procedural format, the capability of defining compound operators, and avoiding references to machine representations by describing quantization on a more conceptual level, thus allowing for machine-specific inference optimizations favorable for usage in edge devices.

FEDERATED LEARNING

Federated learning is privacy by design and collaborative machine learning technique. In its essence, it allows machine learning to comply with the recently emerging data privacy

regulations, incidentally creating a new possibility of using machine learning collaboratively without the need for a central data silo during the training process.

To achieve a collaborative learning process, a data-parallel distributed learning model uses one of the iterative model aggregation mechanisms as a center of the iterative learning process federation down to the data producers themselves.

In FL, every data owner N is the training process participant, updating global model weights by training purely on his local data D_l. Central aggregation server utilizes one of the aggregating algorithms to these new unique data owner model weights $w(M^l)$ on a central server, resulting in the new weights $w(M^F)$. This process is displayed in Figure 1. On the other hand, in standard central ML, the central server first aggregates all of the data owner's data shards before starting the training process to generate new model weights. Furthermore, since the central server only needs model weights for global model calculation, the need for data owners to exchange the original training data, often private, is eliminated.

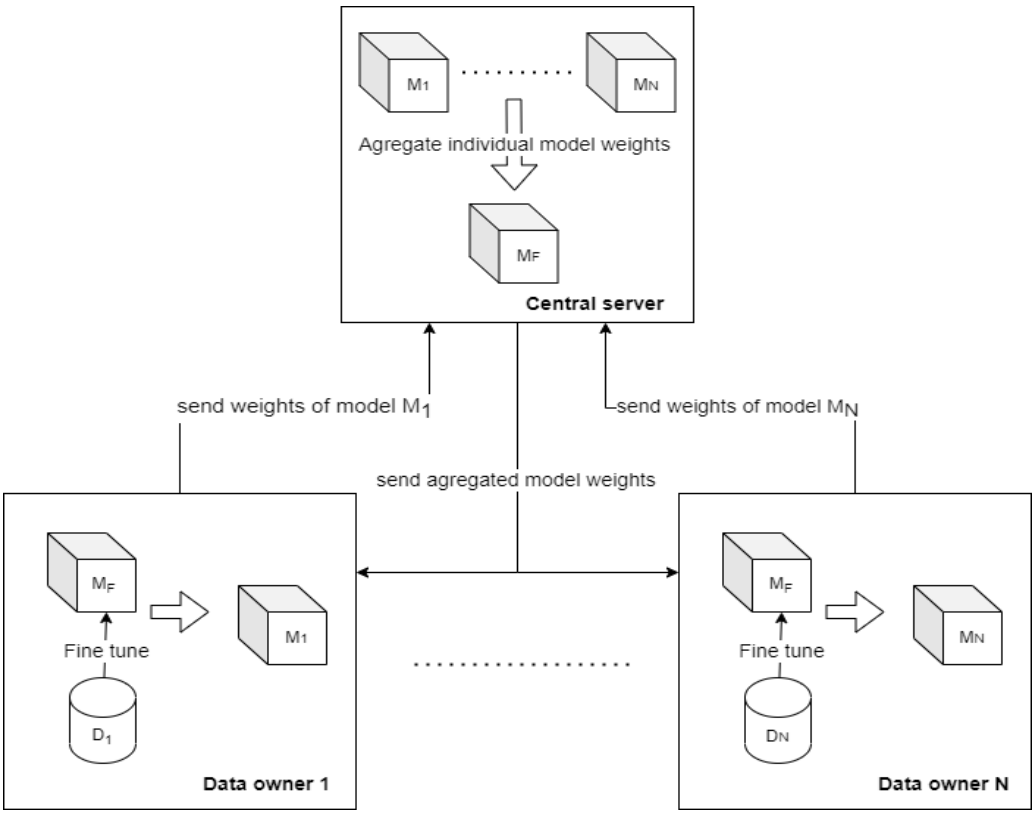

Figure 1. Federated learning process.

In general data-parallel machine learning, there are two ways the data shards (subsets of records in a dataset; physically stored in different locations but logically forming a complete dataset) can partition: horizontally and vertically. The main difference is sharing the same feature sample set in horizontally partitioned datasets (shown in Fig. 2) and contrastingly sharing the same sample set in vertically partitioned ones (shown in Fig. 3). E.g., if different hospitals had the same kind of data of different individual patients - the data is partitioned horizontally. However, if these hospitals had different data on the same patients, their datasets would be partitioned vertically.

Although the original FL model presented by McMahan et al. [51] is designed for horizontally partitioned datasets, several vertical FL models were designed in research that followed [52-54].

However, it is essential to note that the exchange of model weights and their storage on different data owner devices does pose a new possible vector of attacks, commonly known as model inversion attacks. If there is no control over the FL training process, there are vast possibilities of individual data owners tainting the global model by providing model weights trained on local datasets of low quality.

There are also a lot of technical challenges in achieving needed communication requirements for the training process and in eliminating potential problems that could emerge from a significant heterogeneity in data owner's device availability and data quality.

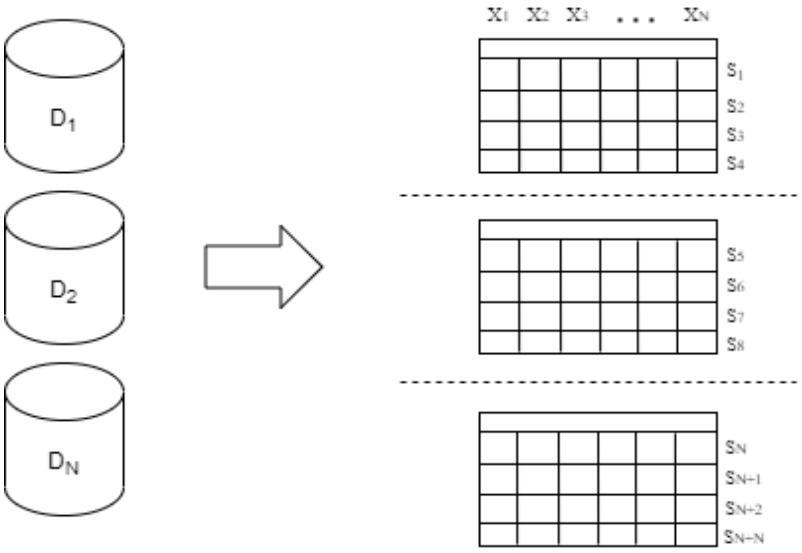

Figure 2. Horizontal data partitioning with each data shard sharing the same feature space.

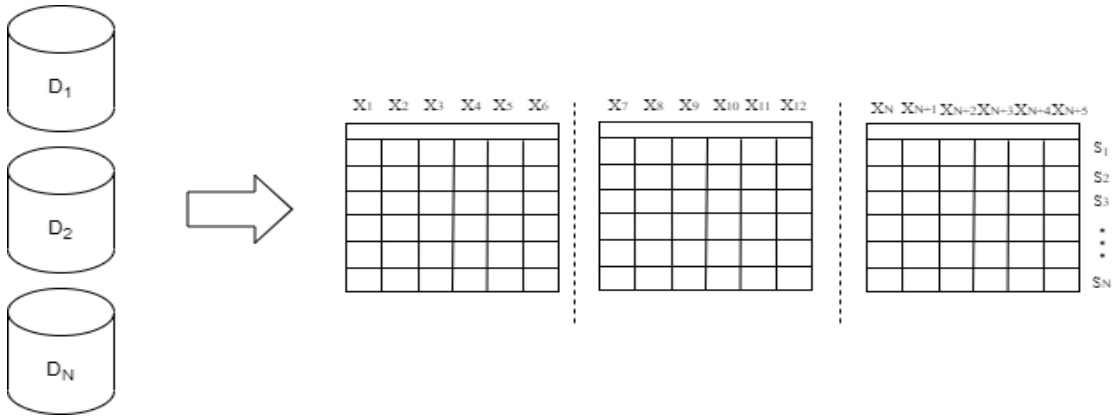

Figure 3. Vertical data partitioning with each data shard sharing the same sample space

Although there are many challenges in creating a real-world working use case, FL can be used as a tool for building cross-enterprise and cross-domain ecosystems for big data and artificial intelligence, where centralized machine learning and cloud-centric paradigms failed to overcome barriers for its inception. Authors in [55] emphasized the importance of coupling the practical usages with the evolution of business models that would accompany it by proposing the usage of FL in data alliances of enterprises.

FLOM FRAMEWORK

This artice introduces a new concept based on the symbiosis of the general federated learning process with that of open model specifications called Federated Learning Open Model

(FLOM). FLOM provides a layout for using FL in the creation of open models and federated training processes with the primary goal of overcoming the technical barriers to using FL. In essence, it allows an easier generation of business models built on the federation of the learning process and using global knowledge without sharing any private or confidential data.

FLOM is a framework for developing an open ML training process done in a federated manner, with model sharing being done by exchanging model definitions in an open standard.

FLOM is accompanied by a technical specification that consists of descriptions of:

- client (individual data owner) data and device requirements,
- a central (aggregation) server specifications and requirements,
- an inferable and trainable model shared with an open standard specification (e.g. PFA, PMML, ONNX),
- an Application Programming Interface (API) with the implemented endpoints for all of the necessary steps for achieving FL process.

The General FL process takes four specific steps that get iteratively repeated during the lifecycle:

1. clients send their model updates,
2. aggregation of model updates into new global model weights (learnable and non-learnable parameters of ML models),
3. disseminating the new global model weights to the client,
4. clients update their local models and start the new iteration.

From the client's side, the FLOM process has a few additional steps to acquire client and server specifications and register the client to the central server (steps 1-4 in Fig. 4).

In essence, our contribution by defining FLOM is in adding these extra steps available through an API endpoint enveloping a traditional FL process with model definitions in one of the open specification formats. By doing this, we hope to help facilitate the usage of ML models in an open and approachable manner that makes it easier to set up and use. On the more practical level, it allows for an easy integration of different prosumers to a collaborative ML process that FL made possible, and FLOM made more accessible.

CLIENT DATA AND DEVICE REQUIREMENTS

Client data and device requirements include a definition of the necessary minimum data quality metrics and optionally capabilities of the client needed to partake in the training process. These can include required not-null data attributes, data generation frequency, data quantity, and any possible additional metrics [56].

Critical endpoints accompany these requirements on the API side for receiving the human and machine-readable specifications and intrinsic procedures to check the quality of the newly generated model weights and the time that took the client to send the newly generated weights [57]. In addition, rules should be applied to drop out and late clients to ensure the model quality.

CENTRAL (AGGREGATION) SERVER SPECIFICATION AND REQUIREMENTS

Central server specifications and requirements include the description of the maximal number of clients and the estimated time it takes to conclude a single training iteration. Estimated training time is analogous to hardware and network capabilities of the server that is used for achieving aggregation and dissemination of model weights as well as serving client requests on the API endpoints.

ML MODEL

The description of the model that is used as the base for the service achieved by the FLOM process is distributed in one of the open formats that include PFA, PMML, and ONNX. The open format allows for training and inference across different software and hardware environments used to achieve the training and inference on the client side.

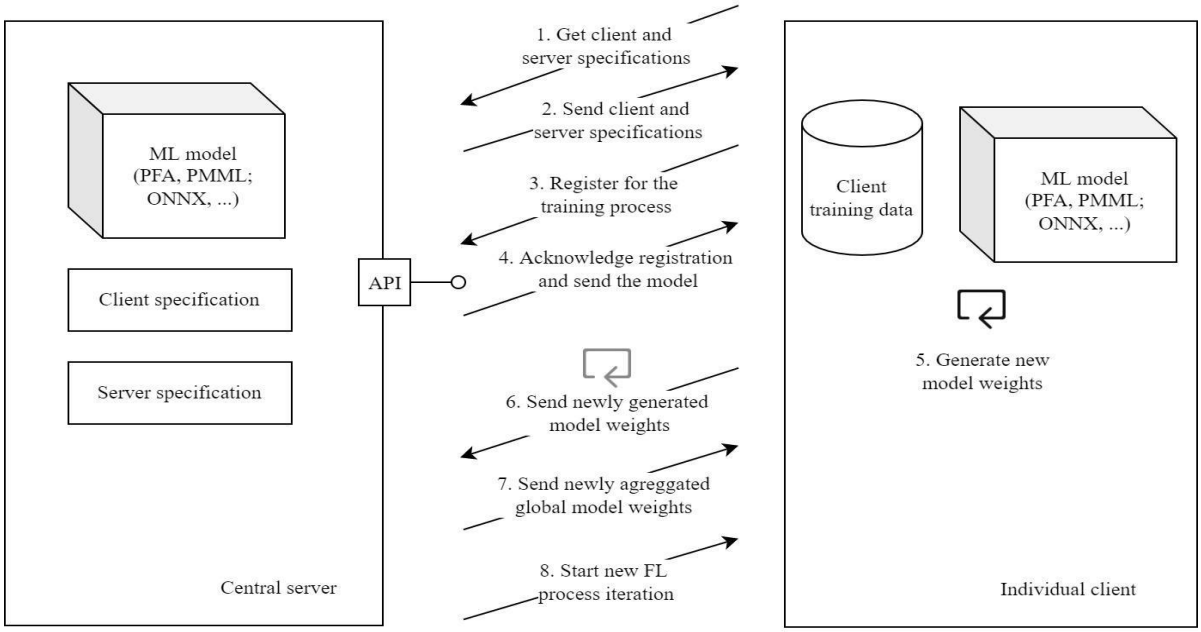

Figure 4. Steps of the FLOM process (steps 5-8 are iteratively repeated during the whole training lifecycle).

API INTERFACE

API interface is located on the central server and needs to support endpoints for registering clients, receiving client model updates, and sending the new global weights to all participating parties. From the client-side, these include actions to:

- receive client requirements description,
- receive central server specification,
- register for participation (generates a unique ID for internal client references),
- receive ML model in an open format,
- send client-specific model weights,
- receive a message with new model weights and a synchronization message to start of the new training iteration.

API interface is achieved using one of the many open-source web frameworks based on well-known standards for client-server communications.

APPLICATION

HORIZONTALLY PARTITIONED ENVIRONMENT

The first use case is created with a dataset containing horizontal data partitions. It aims to attain price prediction of agricultural commodities by incorporating data from the EUROSTAT price index [36] and FAO product import/export dataset [37], which are, in essence, horizontally sharded datasets with partitioning done on the per-country level. An excerpt from these datasets is displayed in Table 1.

The environment is partitioned horizontally on a market region level, where participants could be certain countries, regions, private companies and other organizations.

Although these datasets are open, the use of FLOM in this application allows new business models wherein individual organisations are incentivized to join the training process. The extra incentive is gained from better price forecasting by joining their privately built and historical knowledge on their market area with the latent knowledge located in the more globally distributed knowledge extraction.

Table 1. Data for countries Croatia and Greece found in the FAO and EUROSTAT datasets.

Country code	FAO Commodity ID	Description	Import (t)	Export (t)	Year
HR	882	Milk, whole fresh cow	189435	25849	2020
GR	882	Milk, whole fresh cow	91162	25849	2020

Country code	EUROSTAT Agricultural price ID	Description	Price index (% of 2015 price)
HR	121100	Cow's milk	102.9
EL (GR)	121100	Cow's milk	102.21

Country code	Description	EUROSTAT product price, € per100 kg
HR	Raw milk	34
EL	Raw milk	39

In this use case, the frequency of data generation is once per year, so aggregation server and client hardware specifications are not that stringent.

FLOM consists of:

- linear regression model distributed in the PFA format,
- server specifications that need computational capabilities to run the model aggregation on a yearly basis, with an estimation done by benchmarking using historical country data,
- API endpoints that are defined in the previous section with their locations referenced in OpenAPI format,
- client specifications that define the needed data frequency with yearly samples including organizational area extent in geoJSON format, historic price data in USD, and production and trade data in millions of tons.

By joining the FLOM trading process, the individual organizations would help build the global model by including more finely grained samples than the ones found in open datasets, that are generally per country level. This would further enhance the benefits that organizations would get from forecasting potential further prices, allowing them to compare the potential profits for the upcoming years, regarding changes to their own area of interest, production and trade data.

Using FLOM in this example allows individual producers to understand better the cost-effectiveness of producing each commodity. This new knowledge can be discovered without the need for producers to exchange their production cost data, often confidential.

VERTICALLY PARTITIONED ENVIRONMENT

The second example is built on a vertically partitioned data set that is artificially constructed from a loan ratification ML models analysis [58]. The loan approval prediction system has

the goal of automatically calculating the weight of each attribute of the clients taking part in loan processing and ultimately making the decision whether a new applicant should be approved for the loan or not. Originally, these could be achieved using different ML models, including logistic regression, random forests classifiers, support vector machines, etc. Originally this dataset was a horizontally partitioned dataset with an individual sample being each client (person). The vertical partitioning is done on the client's attributes, and they are separated into two groups: private data and financial data. Private data being: gender (male or female), marital status, number of dependents, education qualification, whether the person is self-employed, and the financial data being: the person's income, co-applicant income, loan amount, loan amount term, credit history and property area (urban/suburban). An excerpt from this dataset can be seen in Table 2.

Table 2. An excerpt from the loan prediction task dataset, with private and financial partitions.

Loan ID	Gender	Married	Dependents	Education	Self employed
LP001032	Male	No	0	Graduate	No
LP001034	Male	No	1	Not Graduate	No
LP001036	Female	No	0	Graduate	No

Loan ID	Income	Co-applicant income	Loan amount	Loan amount term	Credit history	Property area	Loan status
LP001032	4950	0	125	360	1	Urban	Y
LP001034	3596	0	100	240		Urban	Y
LP001036	3510	0	76	360	0	Urban	N

Since people have become accustomed to safeguarding their personal data and becoming more and more unwilling to share it, this could hinder potential loan providers in using services of loan approval prediction systems. However, one could build a service-oriented around ML loan approval where training on private data is done on the client's side using FLOM with more financial data training done on the loan provider's side. This process could be implemented in mobile banking applications wherein, user's private data would stay on their own mobile device.

FLOM consists of:

- tree regression model distributed in the PFA format,
- server specifications that needed computational capabilities to run the model aggregation on a monthly basis, with an estimation done by benchmarking using historical data,
- API endpoints that are defined in the previous section with their locations referenced in OpenAPI format,
- client specifications define the needed computational capabilities to run the monthly training.

By joining the FLOM training process potential applicants could have the benefits of using automated loan approval prediction systems themselves, and better plan their financial future without needing to share their private data. Their presence in the training process would benefit the global model to build the weights that are applied to private data, unavailable to the loan provider.

Both use cases are displayed in Figure 5, which focuses on the process defined in Figure 4, with respect to the client and data partition specific to each use case.

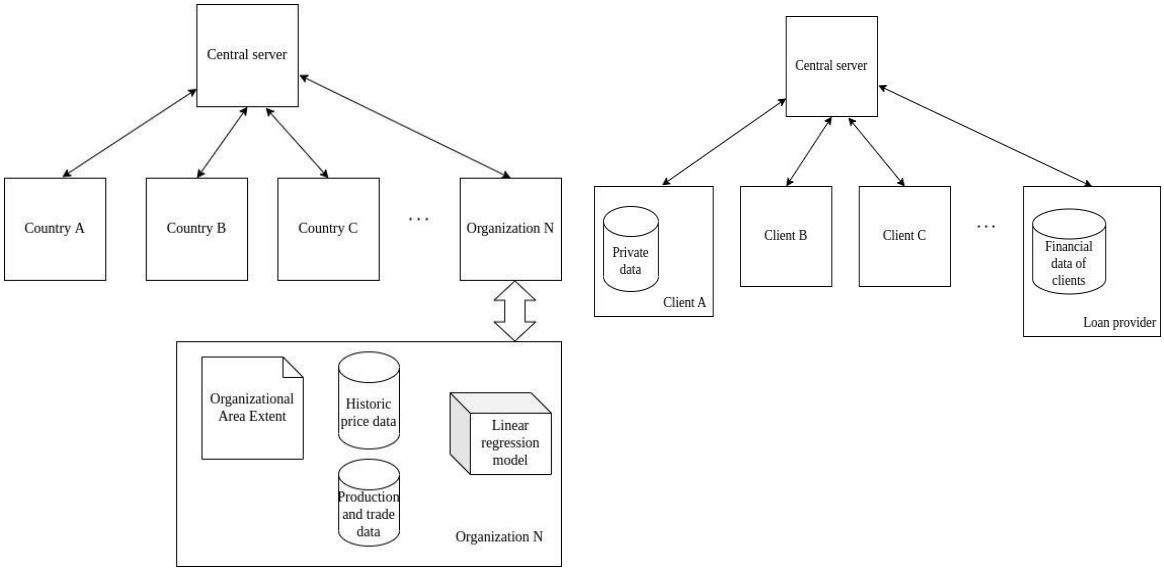

Figure 5. Diagrams of FLOM in the first (agricultural commodity price prediction) and second (loan approval prediction) use case.

CONCLUSION

In this article, we presented a framework for using open standard model formats in a federated machine learning manner. The FLOM framework represents a blueprint for defining open models and the requirements that support the federated learning processfor both the clients and the central server that are accompanied by a model defined in one of the currently available open standards.

The framework encourages the design of new tools, services, and applications for many previously not practically feasible domains. We see this framework as a tool for facilitating collaborative model training and sharing, allowing the combination of knowledge creation from both open datasets and datasets closed due to regulatory or confidentiality reasons. Its potential capabilities as an eGovernance tool are showcased using two potential use cases that leverage openly available and closed datasets attained through the collaborative FL. The use cases showcase the multitude of possible application domains and collaborations, with the first being private business-oriented and the second being private person-oriented.

Future work should be done in evaluating the use cases regarding central ML models that lack the knowledge extruded from private and confidential data. Further disseminating runnable FLOM examples that could easily be reused would encourage broader research on using FL in more general use cases. Using FLOM in the eGovernment context could enable many innovative services that could further citizen participation and incentivize private organizations to build and publicly provide intelligent services in collaboration with governmental and various public organizations.

ACKNOWLEDGMENTS

This research is part of TODO project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857592.

REFERENCES

[1] Twizeyimana, D. and Andersson, A.: *The public value of e-government – a literature review.* Government Information Quarterly **36**(2), 167-178, 2019, http://dx.doi.org/10.1016/j.giq.2019.01.001,

[2] Lachana, Z.; Alexopoulos, C.; Loukis, E. and Charalabidis, Y.: *Identifying the different generations of e-government: an analysis framework*.
The 12th Mediterranean Conference on Information Systems (MCIS), pp.1-13, 2018,

[3] Charalabidis, Y.; Loukis, E.; Alexopoulos, C. and Lachana, Z.: *The three generations of electronic government: From service provision to open data and to policy analytics*.
International Conference on Electronic Government. Springer, pp.3-17, 2019,

[4] Zuiderwijk, A. and Janssen, M.: *Open Data Policies, Their Implementation and Impact: A Framework for Comparison*.
Government Information Quarterly **31**(1), 17-29, 2013,
http://dx.doi.org/10.1016/j.giq.2013.04.003,

[5] Bertot, J.; Jaeger, P. and Grimes, J.: *Using ICTs to Create a Culture of Transparency: E-Government and Social Media as Openness and Anti-Corruption Tools for Societies*.
Government Information Quarterly **27**(3), 264-271, 2010,
http://dx.doi.org/10.1016/j.giq.2010.03.001,

[6] UNCTAD: *Data protection regulations and international data flows:Implications for trade and development data protection regulations and international data flows: Implications for trade and development*, 2016.

[7] EU General Data Protection Regulation (GDPR*): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, OJ 2016 L 119/1.

[8] Jobin, A.; Ienca, M. and Vayena, E.: *The global landscape of AI ethics guidelines*.
Nature Machine Intelligence **1**, 389-399, 2019,
http://dx.doi.org/10.1038/s42256-019-0088-2,

[9] Janowski, T.: *Digital government evolution: From transformation to contextualization*.
Government Information Quarterly **32**(3), 221-236, 2015,
http://dx.doi.org/10.1016/j.giq.2015.07.001,

[10] Lachana, Z.; Alexopoulos, C.; Loukis, E., and Charalabidis, Y.: *Identifying the different generations of Egovernment: an analysis framework*.
The 12th Mediterranean Conference on Information Systems (MCIS), pp.1-13, 2018,

[11] Scholl, H.J.: *Digital government: looking back and ahead on a fascinating domain of research and practice*.
Digital Government: Research and Practice **1**(1), pp.1-12, 2020,
http://dx.doi.org/10.1145/3352682,

[12] Castro, D. and New, J.: *The promise of artificial intelligence*.
Center for Data Innovation, 2016,
https://www2.datainnovation.org/2016-promise-of-ai.pdf,

[13] Abbod, M.F.; Catto, J.W.; Linkens, D.A. and Hamdy, F.C.: *Application of artificial intelligence to the management of urological cancer*.
The Journal of Urology **178**(4), 1150-1156, 2007,
http://dx.doi.org/10.1016/j.juro.2007.05.122,

[14] Leonard M.: *Government leans into machine learning*.
GCN |Technology, Tools and Tactics for Public Sector IT, 2018,
https://gcn.com/articles/2018/08/17/machine-learning.aspx,

[15] Avgerinos Loutsaris, M.; Lachana, Z.; Alexopoulos, C. and Charalabidis, Y.: *Legal Text Processing: Combining two legal ontological approaches through text mining*.
The 22nd Annual International Conference on Digital Government Research, pp.522-532, 2021

[16] Arana-Catania, M., et al.: *Citizen Participation and Machine Learning for a Better Democracy*.
Digital Government: Research and Practice **2**(3), 1-22, 2021,
http://dx.doi.org/10.1145/3452118,

[17] Androutsopoulou, A.; Charalabidis, Y. and Loukis, E.: *Policy Informatics in the Social Media Era: Analyzing Opinions for Policy Making*.
Proceedings EGOV-CeDEM-ePart 2018 Conference, 2018,

[18] Singh, R.; Chun, S.A. and Atluri, V.: *Developing Machine Learning Models to Automate News Classification*.
The 21st Annual International Conference on Digital Government Research, pp.354-355, 2020,
[19] Wani, A., et al.: *Evaluating deep learning approaches for covid19 fake news detection*.
International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation. pp.153-163, 2021,
[20] Sahoo, S.R. and Gupta, B.B.: *Multiple features based approach for automatic fake news detection on social networks using deep learning*.
Applied Soft Computing **100**, No. 106983, 2021,
http://dx.doi.org/10.1016/j.asoc.2020.106983,
[21] Jordan, M.I. and Mitchell, T.M.: *Machine learning: Trends, perspectives, and prospects*.
Science **349**(6245), 255-260, 2015,
http://dx.doi.org/10.1126/science.aaa8415,
[22] Zhou, Y.; Wilkinson, D.; Schreiber, R. and Pan, R.: *Large-scale parallel collaborative filtering for the netflix prize*.
Algorithmic Aspects in Information and Management. pp.337-348, 2008,
[23] Ghoting, A., et al.: SystemML: *Declarative machine learning on MapReduce*.
IEEE International Conference on Data Engineering (ICDE), pp.231-242, 2011,
[24] Sarker, I.H.: *Machine learning: Algorithms, real-world applications and research directions.*
SN Computer Science **2**(3), 1-21, 2021,
[25] Guszcza, J.: *Analysing Analytics.*
Contingencies, American Academy of Actuaries, 2008,
[26] Biswas, P. and Bishnu, P.: *Application of data mining and CRM in banking sector medical insurance*.
International Journal of Innovative Research in Computer and Communication Engineering **3**(1), 38-46, 2015,
http://dx.doi.org/10.15680/ijircce.2015.0301007,
[27] Essien, A.; Petrounias, I.; Sampaio, P. and Sampaio, S.: *Improving urban traffic speed prediction using data source fusion and deep learning*.
IEEE International Conference on Big Data and Smart Computing (BigComp), pp.1-8, 2019,
[28] Gregurić, M.; Vujić, M.; Alexopoulos, C. and Miletić, M.: *Application of deep reinforcement learning in traffic signal control: An overview and impact of open traffic data.*
Applied Sciences **10**(11), 2020,
[29] Guerrero-Ibáñez, J.; Zeadally, S. and Contreras-Castillo, J.: *Sensor technologies for intelligent transportation systems*.
Sensors **18**(4), No. 1212, 2018,
http://dx.doi.org/10.3390/s18041212,
[30] Alexopoulos, C., et al.: *How machine learning is changing e-government*.
Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance, pp.354-363, 2019,
[31] Singh, R.; Chun, S.A. and Atluri, V.: *Developing Machine Learning Models to Automate News Classification*.
The 21st Annual International Conference on Digital Government Research, pp.354-355, 2020.
[32] Goodman, B. and Flaxman, S.: *European Union regulations on algorithmic decision-making and a "right to explanation"*.
arXiv preprint:1606.08813, 2016,
[33] Ding, X.; Zhang, L.; Wan, Z. and Gu, M.: *A Brief Survey on De-anonymization Attacks in Online Social Networks*
International Conference on Computational Aspects of Social Networks, pp.611-615, 2010,
http://dx.doi.org/10.1109/CASoN.2010.139,

[34] Netten, N.; van den Braak, S.; Choenni, S. and van Someren, M.: *A Big Data Approach to Support Information Distribution inCrisis Response*.
Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, pp.266-275, 2016,
http://dx.doi.org/10.1145/2910019.2910033,

[35] Balta D., et al.: *Accountable Federated Machine Learning in Government: Engineering and Management Insights*.
In: *ePart 2021: Electronic Participation*. Lecture Notes in Computer Science **12849**, Springer, Cham, pp.125-138, 2021,
http://dx.doi.org/10.1007/978-3-030-82824-0_10,

[36] EUROSTAT: *Price index*.
https://ec.europa.eu/eurostat/cache/metadata/en/apri_pi_esms.htm,

[37] FAO: *Product import/export dataset*.
https://www.fao.org/faostat/en/#rankings/commodities_by_country_imports,

[38] Kleppmann, M.: *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*.
O'Reilly Media, Inc., 2017,

[39] –: *A Tool Developer's Guide to TensorFlow Model Files*.
https://chromium.googlesource.com/external/github.com/tensorflow/tensorflow/+/r0.10/tensorflow/g3doc/how_tos/tool_developers/index.md,

[40] Keras API Reference.
https://keras.io/api/models/model_saving_apis,

[41] –: *Python Object Serialization*.
https://docs.python.org/3/library/pickle.html,

[42] Open Knowledge Foundation: *What is Open Data?*
https://opendatahandbook.org/guide/en/what-is-open-data,

[43] Data Mining Group: *Predictive Model Markup Language Version 4.4.1*.
http://dmg.org/pmml/pmml-v4-4-1.html,

[44] Data Mining Group: *Portable Format for Analytics Version 0.8.1*.
http://dmg.org/pfa,

[45] –:*Neural Network Exchange Format: Version 1.0.4, Revision 1*.
The Khronos NNEF Working Group, 2021,

[46] –: *Open Neural Network Exchange*.
https://onnx.ai,

[47] Guazzelli, A.; Zeller, M.; Lin, W. and Williams, G.: *PMML: An Open Standard for Sharing Models*.
The R Journal **1**(1), pp. 60-65, 2009,
http://dx.doi.org/10.1002/j.1941-9635.2009.tb00398.x,

[48] –:*CRAN Predictive Model Markup Language*.
https://cran.r-project.org/web/packages/pmml/index.html,

[49] –: *Java PMML API*.
https://github.com/jpmml,

[50] Levitan, S. and Claude, L.: *Open standards for deployment, storage and sharing of predictive models. PMML / PFA / ONNX in action*, 2019.
http://dx.doi.org/10.13140/RG.2.2.31518.89920,

[51] McMahan, H.B., et al.: *Federated learning of deep networks using model averaging*.
ArXiv preprint:1602.05629, 2016,

[52] Wu, Y. et al.: *Privacy preserving vertical federated learning for tree-based models*.
Proceedings of the VLDB Endowment **13**(12), 2090-2103, 2020,
http://dx.doi.org/10.14778/3407790.3407811,

[53] Sun, J., et al. *Vertical Federated Learning without Revealing Intersection Membership*.
ArXiv preprint: abs/2106.05508, 2021,

[54] Romanini, D. et al.: *PyVertical: A Vertical Federated Learning Framework for Multi-headed SplitNN*.
ICLR 2021 Workshop on Distributed and Private Machine Learning, 2021,

[55] Yang, Q.; Liu. Y.; Chen, T. and Tong, Y.: *Federated machine learning: Concept and applications*.
ACM Transactions on Intelligent Systems and Technology **10**(2), 1-19, 2019,
http://dx.doi.org/10.1145/3298981,

[56] Sidi, F. et al.: *Data quality: A survey of data quality dimensions*.
International Conference on Information Retrieval & Knowledge Management, pp.300-304, 2012,
http://dx.doi.org/10.1109/InfRKM.2012.6204995,

[57] Shyn, S.K.; Kim, D. and Kim, K.: *FedCCEA : A Practical Approach of Client Contribution Evaluation for Federated Learning*.
ArXiv preprint:abs/2106.02310, 2021,

[58] Kumar, A.; Ishan, G. and Sanmeet, K.: *Loan approval prediction based on machine learning approach*.
IOSR Journal of Computer Engineering **18**(3), 79-81, 2016,
http://dx.doi.org/10.9790/0661-1803017981.