Privacy-preserving federated machine learning in ATM: experimental results from two use cases

Quantitative proofs of the added value of private sensitive data in ATM operations

Sergio Ruiz
EUROCONTROL Innovation Hub
Bretigny-sur-Orge, France

Andre Rungger
SWISS International Airlines
Zurich, Switzerland

Javier Busto
SITA eWAS
Barcelona, Spain

Salman Toor
SCALEOUT
Uppsala, Sweden

Ignacio Martín
NOMMON
Madrid, Spain

Abstract—**This paper presents the experimental results conducted with a new technological solution that can enable the privacy-preserving exploitation of large private datasets through collaborative machine learning. The solution has been developed under the exploratory research project AICHAIN (SESAR2020 ER04). The final aim of this solution is to enable air traffic management (ATM) operations to be improved with the added value provided by datasets that may be subject to strict privacy requirements and cannot be shared. Experiments have been run with two relevant use cases around Demand Capacity Balancing (DCB) services. Results prove that airline´s private data can improve the machine learning models performance in operations.**

Keywords-component; privacy-preserving machine learning; federated learning; air traffic management; demand and capacity balancing.

I. INTRODUCTION

The European Air Traffic Management (ATM) is under such process of digital transformation and is moving towards what has been recently labelled as the Digital European Sky [1]. In this context, a known study about automation in ATM [2] has analysed the potential of Artificial Intelligence (AI) to enable the required levels of automation in ATM to support the transition towards Trajectory Based Operations (TBO). Such report concludes that machine learning (ML) –including Deep Learning (DL)– techniques are expected to have a key role in the future ATM due to their potential to enhance predictability and to enable advanced optimisation and decision-support tools.

In the early definition of the TBO concept, data sharing was identified as a key enabler for the increased coordination among parties. However, data privacy is a major challenge is such approach, because: i) some relevant pieces of operational information are strategic and business-sensitive for ATM stakeholders (e.g. cost structure of flights, fuelling policies, aircraft weight, etc.) and will unlikely be shared; and ii) data protection laws (i.e. GDPR) may impose additional privacy constraints to data sharing (e.g., passenger data). As a result, some of the pieces of information that are relevant to achieve the desired efficient and resilient ATM operations are just non-shareable, i.e. data subject to strong privacy requirements.

The cost of opportunity of not exploiting such private/non-shareable data might be large in the context of the digital ATM. AI techniques typically require as much data as possible to perform well, from which the afore-mentioned non-shareable data might be important sources. The limits of the data sharing approach to access the private operational data from stakeholders may undermine the ability of ML techniques to improve the predictability of ATM operations and ultimately optimise the quality of service of ATM system.

Federated Learning (FL) [3] is a new enabler technology that can facilitate the development of ML models while preserving the privacy of the private data sources. FL is an evolution of traditional ML/DL methods that allow training a model in a distributed and collaborative way: the model parameters are first locally updated by each data owner and afterwards combined into a single model through novel aggregation techniques. This way, the federated model leverages all available training data, while no single data record is processed by other than the data owner itself.

On this basis, the AICHAIN project has developed a technological solution that uses federated learning to train –and serve– machine learning models while respecting the privacy of stakeholders' sensitive data. The solution has considered the combination of federated learning with other technologies (e.g. blockchain) to satisfy all the privacy and operational requirements in the ATM domain [4]. Such solution has been partially validated experimentally with two ATM case studies: the prediction of flight Estimated Time of Take-off (ETOT) and the prediction of the 2D routes flown by Airspace Users (AUs) in the tactical phase of flight.

This paper presents the experimental results from the two case studies and discusses about the added value of the FedML approach and its potential to improve the ML models used in ATM operations. The experiments involved the use of private datasets that normally cannot be available through the conventional data sharing paradigm. The private data features used in this paper have been kindly provided by SWISS International Airlines and made available to the AICHAIN project consortium to facilitate the experiments.

The reminder of the paper is structured as follows: Section II provides the state of the art of federated learning and clarifies de concept. Section III presents the proposed case studies and the experimental plan followed to conduct them. Section IV discusses the experimental results and illustrates the value of the private data within the case studies. The paper ends with the conclusions and the identification of next steps in Section V.

II. STATE OF THE ART

A. Federated machine learning

Federated Machine Learning (FedML) is a new concept proposed by Google in 2016 [5][3][6][7], in which input data privacy can be preserved while ML models can be trained leveraging the best data available.

Figure 1 illustrates the basic FedML concept and its topologic logical architecture. FedML consists of sharing ML model parameters (e.g. coefficients) instead of the data, and training them in a distributed and collaborative manner. Only the ML parameters after the local training at nodes are shared among the federation members and then aggregated in the master node. The underlying idea can be described in four iterative steps: 1) the server sends the last version of the global model (e.g., a neural network) to the nodes; 2) the nodes locally update that model with training iterations using their local data; 3) each node sends back the model updated locally to the server; and 4) the server composes the updated version of the general model through mathematical aggregation functions. Note that this way, the private datasets never left their owner premises, while it is still possible to build, through collaboration, a complete ML model that can then be used to make predictions and/or optimise a certain process (e.g. an ATM process).

It is worth noting that, in addition to the privacy protection of the input data, FedML technologies are also useful to speed up the training of large and complex ML models. This is due to the de-centralised and parallelised nature of the federated training. Therefore, federated training presents a great potential for real time ML applications (training, prediction and optimisation of the ML models) [22]. The use of bandwidth is also more efficient (i.e. less costly and faster) compared to the data sharing approach, since exchanging ML model parameters typically require exchanging much less amount of information than sharing the raw data.

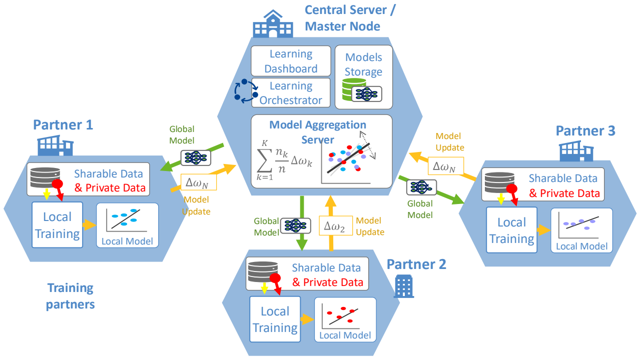

Figure 1. Federated machine learning concept and architecture

Many traditional machine learning methods have already been successfully adapted to the FedML case, including deep Convolutional Neural Networks (CNNs), Long-Short-Term-Memory (LSTMs), Support Vector Machines (SVMs), conformal predictors, and ensemble-type methods [7][8][9][10]. However, while all these efforts helped to increase the levels of security and input data privacy compared to traditional learning (centralised and dependent on pooled data), there still are some open gaps that need to be addressed, like trust (transparency, tamper-proofness and auditability), data privacy, and model lifecycle governance [11].

B. Challenges and limitations of federated machine learning

The FedML approach presents some challenges and limitations that need to be addressed before it can be used in operational environments.

Some of these shortcomings are related to cyber-security, trustworthiness, runtime performance and/or other requirements. Hence, a combination of privacy enhancing tools is normally needed to ensure full covering of all privacy requirements (e.g., input and output privacy, as well as input and output verification/trust, and information flows governance).

An example of a privacy open challenge related to cyber-security and output privacy can be illustrated by a bad-intentioned "attacker" that could use the intermediate training gradients, i.e., parameters that are shared by the local nodes after the training of the model with the local private data, to infer important information about the private data used during the training of the shared ML model [12][13]. This would be a major issue in the ATM domain since the air transport industry is a very competitive environment and some stakeholders could potentially monitor and learn from others to improve their market strengths. This risk of potentially losing market competitiveness could rapidly discourage the participation of a wide number of stakeholders to the collaborative training of FedML models, consequently undermining the expected operational benefits of the technology.

Another important aspect to consider when designing a federated learning system is that the effective collaboration of the data owners is required to correctly train the global model.

In practice, it should not be assumed by default that all the federated parties (data owners) will be willing to cooperate in the collaborative training of a model. Some contributors could limit their collaboration either in look for cost reduction or simply due to unwillingness to participate. Hence, a system of governance that steers the federated alliance and aligns the incentives/interests of each individual with the goals of the system is required.

It is worth mentioning that the AICHAIN solution has addressed the above privacy and governance challenges. Several technologies have been identified to enhance the federated learning core engine of the solution [4]. For instance, in the last years the novel technology of Blockchain [14] has emerged with great potential to provide trust to some processes of the FedML approach. In addition, Blockchain can also enable the use of trustable digital tokens, which could be used to enable advanced governance and incentives mechanism based on the exchange of digital assets with added value for the token holders. More details about how the AICHAIN project has addressed the topic of governance and incentives in deliverable D4.2 "Governance and Incentives model" (available in the project webpage [4]).

For the experiments presented in this paper (aimed to prove the feasibility and added value of exploiting private datasets in ATM applications), the challenges mentioned above can be simplified and disregarded, since it has been assumed that all the federated participants act as *unconditional honest participants*.

III. Experimental design and use-cases

Two ATM case studies have been developed with a double purpose: i) to validate the AICHAIN solution prototype; and ii) to demonstrate the potential value of the FedML approach in the context of the Digital European Sky.

The case studies selected address some well-known ATM challenges and are both considered high priority cases by the Network Manager (i.e. EUROCONTROL). Some improvements have been recently achieved in the two use cases thanks to the introduction of machine learning models to address them. However, these models were trained using data features that were collected and available by EUROCONTROL only. The hypothesis validated experimentally in this paper is that the performance of such ML models could potentially be enhanced if new relevant private/non-sharable data features are available to the model training processes through FedML. Following subsections present each case study, the data available and the experimental methods followed.

A. Use case 1: Estimated Take-Off Time (ETOT)

At present, the Estimated Take-Off Time (ETOT) of each individual flight is obtained from the Enhanced Tactical Flow Management System (ETFMS) Flight Data (EFD), which is regularly updated since the submission of the Initial Flight Plan (IFP) up to the Actual Take-Off Time (ATOT). The ETOT calculated by the ETFMS system is subject to many sources of uncertainty that hinder the actual predictability of the traffic at sectors. Some sources of discrepancy between the ETOT and the ATOT include: unforeseen network congestion; severe weather constraints; reactionary delays; and the reaction (flight plan changes) of the airspace users to these changing conditions.

A study published by EUROCONTROL about A-CDM (airport collaborative decision making) [15][16] presented quantitative evidence showing that improving the take-off time predictability can significantly prevent sector over-deliveries, which could be used to reduce the en-route capacity buffers (latent capacities) without compromising the required levels of safety. Figure 2 illustrates the relationship between higher predictability and the possibility of reducing the buffers in the declared/published sector capacities while maintaining the same levels of safety.

The same study concluded that, following a wider implementation of Airport CDM, the benefits could be:

- Potential increase of sector capacity within the core area by up to 4% which equates to between 1-2 aircrafts per sector
- Reduction of en-route delays of between 33%-50%.
- Some sectors which are expected to be saturated are not really saturated. As a result, some imposed regulations may not be required.

A more recent study [17] updated and refined the previous analysis pointing that the benefits could be less significant than the ones suggested in the previous study, but still reaffirmed that increasing ETOT predictability may lead to significant increase in the actual sector capacities.

Similarly, the aim of the case study addressed in this paper is to improve the Estimated Take-Off Time (ETOT) accuracy before the departure of flights, for the following reasons:

- To reduce the capacity buffers at sectors (less buffers will lead to higher declared capacities).
- To reduce the number of regulations (better accuracy in the assessment of potential sector overloads).
- To enable traffic complexity management (i.e. more effective ATCFM/DCB measures at the level of 4D trajectories instead of at the level of flows).

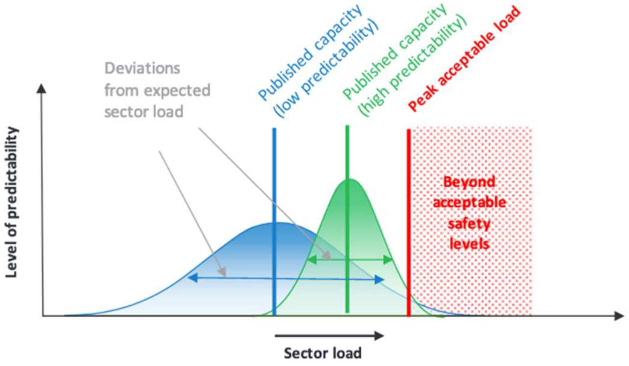
Deviations from expected sector load
Published capacity (low predictability)
Published capacity (high predictability)
Peak acceptable load
Beyond acceptable safety levels
Level of predictability
Sector load

Figure 2. Higher predictability can lead to more sector capacity (source: the Airspace Architecture Study)

Regarding the complexity management possibilities, it is worth noting that the Maastricht Upper Area Control Centre (MUAC) has developed the Traffic Predictions Improvements (TPI) project with the purpose of optimising the use of Air Traffic Controllers (ATCOs) time to reduce the uncertainty of capacity predictions. The project aimed at improving predictability of traffic, allowing MUAC to make a more accurate view of the incoming traffic and consequently apply a better management of capacity and a better anticipation of the ATFCM and Air Traffic Control (ATC) measures on the demand. To achieve the goal MUAC developed ML models to address the following sub-problems: route prediction, 4D trajectory prediction and sector sequence prediction. The model used in this paper for this ETOT case study is a simplified version of one that was deployed as part of the TPI project [18].

Trained against (ATOT - ETOT) target, this model provides an improved ETOT value based on the initial ETOT (and other features as input). This prediction will then contribute to the improvement of the Demand Profile. A more accurate ETOT can help the ANSPs and the NM to better assess the traffic demand at sectors. For some short-term tactical planning timeframes (e.g. less than 1 hour) more accurate ETOTs may facilitate the management of traffic complexity and even trajectory interactions (strategic conflict management). This should result in less unnecessary constraints on the traffic demand and in a better planning of the capacity resources, thus potentially leading to a reduction of the latent capacities (fewer capacity buffers).

The model structure of each training sample consists of a set of input features (like the departure and destination airports, the turn-around time, the ATFM delay, the ETOT, the LAT with respect to ETOT, etc.) and a target to be predicted, which is the ATOT of the corresponding flight. A set of input features extracted from the (private) data provided by SWISS for the same flights has been used to generate a new enriched/augmented model. Table I and Table II show, respectively, the most important non-private and private features that influence the most in the model performance. The ones identified as the most important will be mentioned in the section of results. The comprehensive list of the NM data and SWISS features used to develop the ETOT enhancer model can be found in Deliverable D3.2 (Operational value) of the AICHAIN project, which is based on the former one published in [18].

TABLE I. NON-PRIVATE FEATURES DESCRIPTION FOR USE CASE 1

ADEP	Aerodrome of departure of the flight
ADES	Aerodrome of destination of the flight
ADEP_LEG	Aerodrome of departure of the previous flight leg
ADEPETO_IFP_TO_ADEPETO	Difference between the current Estimated Take-Off Time (ETOT) and the ETOT according to the initial flight plan
ADEPETO_IFP_TO_ADEPETO_LEG	Delay of the previous leg
ARCTYP	Aircraft Type (ADEXP)
ATFMDELAY	The ATFM delay allocated by the ETFMS system (via CTOT) to that flight (at the time the sample was generated)
CDMSTATUS	cdmstatus values as available in ETFMS (not present for non-CDM airports)
DAY	Day of the week
EOBT_IFP_TO_EOBT	Difference between the current Estimated Off-Block Time (EOBT) and the EOBT according to the Initial Flight Plan
EOBT_IFP_TO_EOBT_LEG	Same as EOBT_IFP_TO_EOBT but for the previous flight leg
EVENT	Type of sample: message, system event or manual
EVENT_LEG	Last source of the sample sent by the previous flight leg (see EVENT feature)
FLTSTATE	Status: filed, active, regulated, slot_issued
FLSTATE_LEG	Last flight status of the previous flight leg (see FLTSTATE feature)
FLIGHT_DURATION_LEG	Scheduled duration of the flight
HOUR	Hour of the day
MONTH	Month of the year
RWY	Last letter of the Standard Instrumental Departure (SID) procedure. Used as a 'proxy' for the take-off runway
TIME_FROM_REG_CHANGE	Time from allocation or change of aircraft registration number (i.e., airframe) to the flight.
TIMESTAMP_IFP_TO_TIMESTAMP	Time form the reception of the initial flight plan at the ETFMS
TIMESTAMP_LEG_TO_TIMESTAMP	Time from last sample of the previous flight leg
TIMESTAMP_TO_TSAT	Time to Target Start-up Approval Time (TSAT)
TIMESTAMP_TO_EOBT	Time to Expected Off-Block Time (EOBT)
TIMESTAMP_TO_TOBT	Time to Target Off -Block Time (TOBT)
TURNAROUND_LEG	Expected turn-around time (can be negative if previous flight has a large delay)
TAXITIME	The taxi-time-field contains the most recently known taxi-time value by ETFMS at the time of the sample

TABLE II. PRIVATE FEATURES DESCRIPTION FOR USE CASE 1

NUMPAXBOOKED	Number of passengers that paid a ticket
SWISS_RWYNUM	Runway ID allocated to the flight
NUMPAXFLOWN	Number of passengers that took the flight
SWISS_ETOT_TO_ETOT	ETOT predicted by SWISS (delta from ETFMS ETOT)
DEPARTURE_GATE	Geodesic longitudinal separation between origin and destination
OCCUPATION	Geodesic latitudinal separation between origin and destination
SWISS_EOBT_TO_EOBT	EOBT predicted by SWISS (delta from ETFMS EOBT)
SWISS_EOBT_TO_SOBT	EOBT predicted by SWISS (delta from ETFMS SOBT)
CREW_CONNECTION_TIME_PREVIOUS_FLIGHT_SCHEDULED	Time scheduled for the crew connection with the current flight
SCDINPAX_G6	Number of passengers with scheduled connection time of more than 60 min from other flights to the selected flight

B. Use case 2: ATM 2D route prediction

As part of the strategic, pre-tactical and tactical demand and capacity processes, the airspace users are required to provide their flight plans as soon as possible to the Network Manager. In their flight plan, AUs must declare –among other aspects of the flight– the route (set of airspace 2D waypoints) and the flight level they intend to fly. These parameters can change at the time of operations due to different reasons (route availability, network congestion, weather constraints, and others). As in the ETOT case, the lack of predictability of the route and flight level of the traffic might generate capacity overloads and consequently lead to an unwanted increase of the capacity safety buffers required at many sectors and to more traffic regulations.

This second AICHAIN case study aims at developing a ML prediction model to improve the predictability of the 2D route and flight levels well in advance before operations. This can help the NM to allocate the capacity resources and the traffic demand in an optimal way while minimising the number of restrictive network constraints (e.g. ATFM regulations).

TABLE III. NON-PRIVATE FEATURES DESCRIPTION FOR USE CASE 2

Feature name	*Description*
Airport GDP	Gross Domestic product of the Origin/Destination surroundings areas
Airport population	Population density of the Origin/Destination surroundings areas
Airline TOW	Measured Take-Off weight by the airline of each flight
CAPE	Used as a storm proxy
Charges	The charges paid for the current route for a given aircraft
Connecting passengers	Number of passengers that have a flight connection in the destination airport
Daily flights	Number of flights for each od pair and day
Direct costs	Sum of the fuel and charges costs
DoY	The day of year in which the flight takes place
DoW	The day of week of the flight codified accordingly
Flight Time	The ETOT hour of the flight
Fuel cost (estimation)	Estimation of the cost of fuel for each given route
Humidity	The relative humidity observed along the route, that is a requisite for thunderstorms to occur
K-index	Weather metric that approximates the probability of a thunderstorm to happen
Latitude diff	Geodesic latitudinal separation between origin and destination
Longitude diff	Geodesic longitudinal separation between origin and destination
Market share	Airline's flight share for each OD pair and day
Military zones	The route crosses a closed military zone, not use as a feature but to discard routes
Regulations	The duration of the regulation that affects the route
Route length	The length in kilometres of a given route
Wind at origin/ destination	Variable that measures how aligned and in what value local wind at the airport is
Wind length	Length of the route in kilometres adjusting the effect of the along wind

TABLE IV. PRIVATE FEATURES DESCRIPTION FOR USE CASE 2

Feature name	*Description*
Airline TOW	Measured Take-Off weight by the airline of each flight
Connecting passengers	Number of passengers that have a flight connection in the destination airport

The proposed model is a two-step approach that models the prediction of the route to be flown as a binary classification ML problem.

- The first step consists of a *DBSCAN clustering that aggregates the routes of each OD pair into "relevant routes"*, that is, a set of representative routes of all the ones observed for a given period. This way, the problem is to choose among representative routes which one will be flown (multi-class classification).
- The second step develops a *ML classifier that determines whether a given representative route is to be flown at each OD pair*. To perform predictions, the features of each potential route are obtained by subtracting the observed features of the most flown route in the previous month to the observed features of each possible route.

For this paper only the second step which develops the supervised ML model is relevant. The original model was developed in [19] and relies on NM's data and public weather data sources to generate its features. The details of the baseline (non-augmented) version of the model used in this case study can be found in [20].

Table III and Table IV show the non-private and private features of the model developed for Use Case 2. A subset of features provided by an AU (SWISS) have been used as part of the augmented model. These features are:

- *Flight Take-Off Weight:* the measured weight at take-off as measured by the AU. The actual weight of the aircraft can determine taking a more direct route (e.g. the flight is full and cannot delay other flights at destination).
- *Number of passengers with flight connections at destination:* this is a good proxy indicator for the model to "learn" the importance of the cost of delay and connectivity constraints for that flight (e.g. a lot of people with connections might make the AU take a faster route).

For further details on the case studies, the reader can refer to the deliverable D3.2 "AICHAIN Operational Value – Final Report" of the AICHAIN project (publicly available at the project webpage [4]).

C. Machine learning algorithms used

Both use cases have been developed and experimentally assessed with both neural networks (NN) and decision trees (DT) models. Due to the current prototype development state, the federated learning experiments were only executed with neural networks. For the sake of simplicity, the results presented in this paper will combine results from the different experiments where the two types of models were used. All the details regarding the experiments conducted and their results can be found in the public deliverables of the project [4].

D. Dataset description

To perform the experiments of each case study, the AICHAIN consortium had access to Network Manager data required to develop both case studies using regular ML. In addition to that, SWISS Air has prepared a dataset including some private features of their own flights that have been identified as relevant for the use cases.

All the data available (both NM data and AU data) corresponds to the following areas and periods:

- Case study 1 (ETOT prediction): all flights crossing Maastricht Upper Area Control Centre (MUAC) sectors from 27th June 2019 to 28th February 2020.
- Case study 2 (2D route prediction): all flights connecting Switzerland and great London area airports from 27th June 2019 to 28th February 2020.

It is worth noting that the dataset provided by SWISS has been shared with the project partners to bring the highest degree of control and transparency as possible to the experiments (see next section). However, under normal (non-research) operational conditions the AICHAIN solution would not need that the AUs share their data with anyone to train a model.

E. Assessment framework and experimental methods

The experiments presented in this paper aim at demonstrating the feasibility of the proposed federated approach and to generate quantitative evidence of the added value that the AU's data may have for the proposed case studies.

To this aim, the experimental method consists of the generation and benchmarking of three different scenarios. The scenarios differ from each other in the way the ML models were trained (dataset augmented or not, and federated training or not):

- *Baseline model* (V0), a.k.a. *bottom boundary performance scenario*. In this case the machine learning model has been trained in a non-federated manner with the NM data only. For each use case under consideration, this model can be considered as *equivalent* to the machine learning models currently used by NM in operations.
- *Augmented model* (V1), a.k.a. *upper boundary performance scenario*. The training of the machine learning model has been performed with the features available to NM *plus* the private sensitive variables from the AU. The two datasets were merged before the training, thus expanding the number of features for the same samples available at the NM data. The model has not been trained following the federated learning approach, but with a single computer having full access to the augmented dataset. This experiment aimed at finding the value of the private datasets under ideal conditions, i.e., assuming that all the data (sharable and non-sharable) is available to train the model without the need for federating. The rest of the training conditions are the same as in V0. Thus, this scenario determines the upper boundary of model performance that can be achieved in each use case by exploiting the private data available.
- *Federated augmented model* (V2), a.k.a. *solution prototype performance scenario*. As in the previous scenario, the federated model has been trained with the NM's variables as well as the private AU's variables, but through federated learning with private data distributed in users' data silos. The training of the federated model has been done using the AICHAIN prototype. The performance of the federated model is expected to be equal or very close to the V1 ones.

Figure 3 shows the general methodology for benchmarking the different scenarios. Each experiment consisted in comparing a reference scenario against the benchmark scenario.

Since there has been private data only from one AU (SWISS), all data points in the federated experiments have been randomly split into two "synthetic AUs", each one having 50% of the available flights. In the federated experiments, two independent nodes, yielded to two model updates at each step.

IV. RESULTS OF THE EXPERIMENTS

A. Use case 1: ETOT prediction

Results of this use case are presented based on two different executed experiments: The first experiment was focused on assessing the *private data value*: both the baseline (V0) and augmented (V1) models were modelled with Decision Trees (*LightGBM*), with the purpose of proofing the value of private data. The second experiment was focused on assessing the *technological feasibility of the federated learning approach*. due to the current prototype development state, the federated learning model (V2) experiments were only executed with neural networks (further details can be found in the public deliverables of the project [4]).

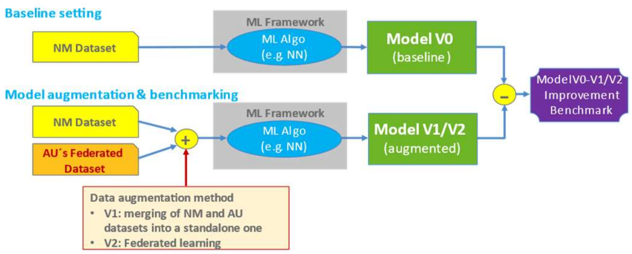

Figure 3. Methodology and setup for benchmarking scenarios/models

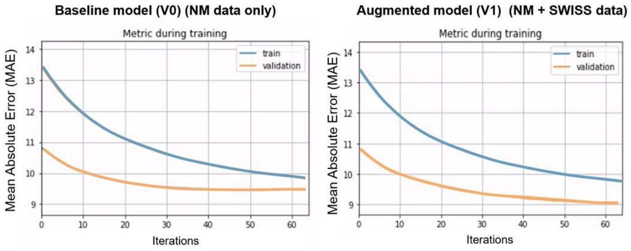

Figure 4. Learning curves of scenarios V0 and V1 for use case 1

The available dataset contained around 5 M samples from 1.1 M flights, and it was split in three different sets for the experiment: training set (70% of the samples), validation set (15%) and test set (15%). Each model have been trained using the training set and validated with the validation and test sets. Figure 4 shows the learning curves as a function of the training steps for the scenarios V0 (baseline) and V1 (augmented).

Table V shows the results for the bottom and upper boundary scenarios in terms of *average prediction absolute error (MAE) as a function of the prediction look-ahead time*. The augmented scenario V1 is better in terms of predictability than the baseline V0, proving that the private data of SWISS has added value to the model and could increase the predictability in this particular case study. Note that the private data could improve the error mean more significantly for the prediction look-ahead times in the range between 30 and 120 minutes than for short or very long look-ahead times (i.e. less than 30 and more than 120 minutes).

Table VI shows the aggregated statistical distribution description of the absolute error on average for the scenarios V0 and V1. To complete the comparison, the figures obtained from the legacy system ETFMS (basic ETOT predictions without using machine learning models) are also shown.

The following results can be inferred from the first experiment:

- Average improvement with NM data only (scenario V0) with respect the ETFMS predictions: +11.4%
- Improvement with NM+SWISS data (scenario V1) with respect the ETFMS predictions: +14.3%
- Relative improvement (calculated with ((ETFMS-V1)/(ETFMS-V0)-1): +25%
- The improvement is observed in all the distribution (note: the slight increase in the second quartile of the V0 model with respect the ETFMS can be explained due to the presence of more "small errors" compared to the presence of "large errors", relative to each of the distributions).
- The mean value is notably higher than the median value (i.e. second quartile), which suggests the right-skewed distributions typical of ETOT errors.
- The performance improvement when private data is introduced (V1 vs V0) is more noticeable in the prediction look-ahead times between 30 to 120 minutes before flights departure.

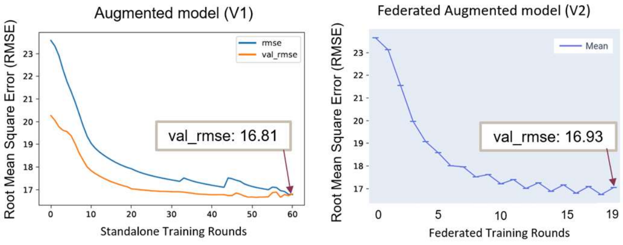

Figure 5. Learning curves of scenarios V1 and V2 for the second experiment of use case 1(with neural networks). Metric: root mean square error (RMSE).

TABLE V. ABSOLUTE AVERAGE ERROR OF THE PREDICTIONS ON THE TEST SET AT DIFFERENT PREDICTION LOOK-AHEAD-TIMES (IN MINUTES)

Id	*(0, 30]*	*(30, 60]*	*(60, 90]*	*(90, 120]*	*(120, +)*
Baseline (V0)	5.12	7.87	9.16	9.47	11.19
Augmented (V1)	5.06	7.70	8.94	9.19	11.04

TABLE VI. ABSOLUTE AVERAGE ERROR OF THE PREDICTIONS ON THE TEST SET (IN MINUTES)

Statistic	ETFMS (legacy system)	Baseline (V0)	Augmented (V1)
Mean	10.5	9.3	9.0
Std deviation	15.7	13.0	13.0
1rst quartile	2.9	2.7	2.6
2on quartile	6.0	6.1	5.8
3rd quartile	12.0	11.4	10.8

Regarding the second experiment, Figure 5 shows the root mean square error (RMSE) evolution for both the augmented and the federated augmented scenarios, once the prototype was ready and the use-case modelled with Neural Networks. As expected, the federated experiments led to very similar results compared to the non-federated augmented scenario V1, proving not only the feasibility of the federated approach, but also the capability of the solution to extract all the value from the private datasets. Note that in the figure the horizontal axis has a different scale at each chart shown because the federated training cycles works differently than with classical ML approach. Also note that RMSE metric was used in this experiment and cannot be directly compared with the MAE metric of the first experiment.

It is also interesting to analyse the importance of the features on the predictability performance of the model. Figure 6 shows the importance feature analysis for the upper boundary case (NM+SWISS data). In the figure the private features provided by SWISS have been highlighted in red. It can be observed that some of these private features contributed significantly to the increase of predictability of the take-off time, e.g.: the number of passengers booked and already onboarded, the runway assigned, the own ETOT predictions made by SWISS, the assigned gate, the number of passengers with connection, and the time of connection of the previous flight. Other private features used can be found in the project deliverable D3.2 [4].

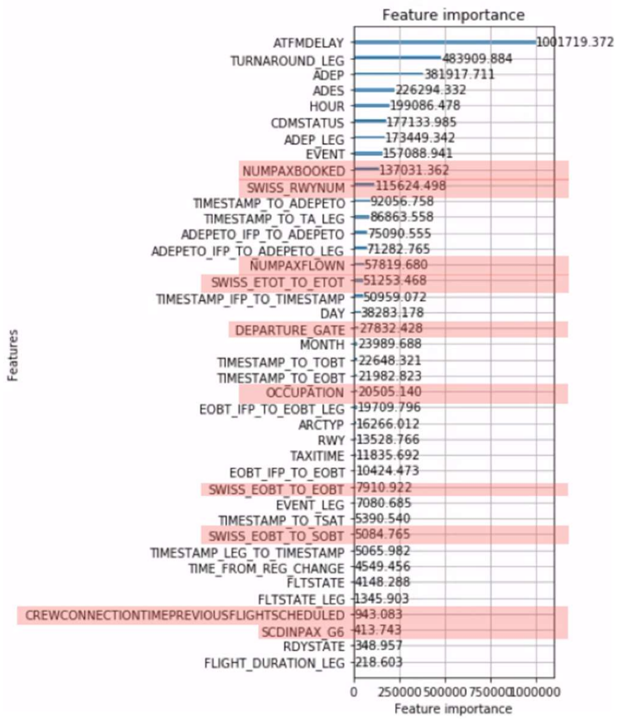

Figure 6. Feature importance analysis (NM + SWISS data)

B. ATM 2D route prediction case study

The available dataset was split into two subsets to facilitate the training and validation of all models according to time: training set (85% of the samples), and validation set (15%). The performance metric used in this case study is the accuracy of route selection, that is, the amount of route predictions that were correctly assigned to their afterwards flown route.

Table VII summarises the accuracy achieved by each model scenarios V0, V1 and V2 assessed using the validation set in the 2D route prediction case study. In addition, to set a reference performance similar to the legacy system of the previous case study, a simple model based on heuristics has been defined: the most flown route (MF), which makes a prediction based on the route of each OD pair that was the most flown route in the previous month from the prediction time.

These results indicate that the simplified heuristic of the most flown route can already provide a relatively high accuracy of 87.6%. This is due to the fact that AUs tend to fly the same route pretty often, which simplifies the problem of prediction. Regardless, the application of machine learning models with the baseline scenario V0 in which only NM data was used could bring a significant improvement with respect to the heuristic of the most flown route of around 8%. It is worth noting that the accuracy reached was actually very high in V0, i.e. 95%. Nonetheless, the results of the augmented V1 scenario show that after adding just two extra private AU's features to the model, the performance could still increase over the baseline case V0, with a relative improvement of about 4%. The federation of the experiment, V2, again shows similar results as in V1, perhaps showing slight less improvement, but the accuracy achieved is still above the obtained without the AU's dataset.

Figure 7 shows the feature importance analysis of the variables using the SHAP values approach [21]. As observed, there are several variables that contribute significantly to the model among which both TOW and number of connections are included. Moreover, it is worth noting the fuel cost variable. The fuel cost variable used in the model is an approximation to the actual cost of fuel for the airline based on average fuel prices. The reason for not using actual airline fuel prices is because it was found too sensitive information to share it with the project consortium in the context of the research project, even if the dataset was protected under strict non-disclosure agreements. Due to the importance that this private feature has in this use case, it is expected that the model performance could still be improved significanly more if the actual value of such feature could be exploited directly in the SWISS premises through federated learning.

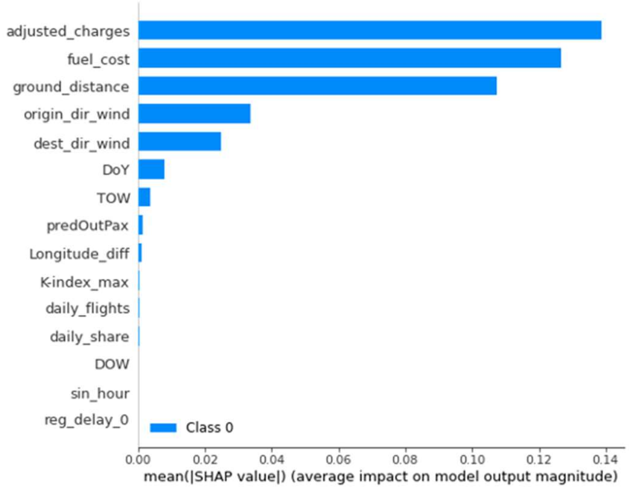

Figure 7. Feature importance analysis of the UC2 (SWISS + NM data)

TABLE VII. RESULTS OF 2D ROUTE PREDICTION (ACCURACY)

Id	*Test description*	*Federated*	*AU features*	*Accuracy*
Most flown (MF)	Non ML model baseline	-	-	0.876
Baseline (V0)	No AU properties	No	No	0.95
Augmented (V1)	AU properties	No	Yes	0.954
Federated (V2)	Federated model with AU properties	Yes	Yes	0.953

V. Conclusions

The AICHAIN solution enables the privacy-preserving exploitation of large private datasets from different stakeholders to enrich operational machine learning applications. This is achieved through privacy-preserving federated machine learning, where the training and serving of the federated models can be done at the data owners' facilities in a cyber-secured and trustworthy manner without sharing any data. Thus, private data owners can remain in full control of their dataset's privacy

From the experiments conducted with the two ATM use cases, the following conclusions can be extracted:

1. The exploitation of private data can improve ML models performance.
2. The performance improvement of the models augmented with private data may differ from use case to use case. It is expected that private data will always yield model performance improvements, which can be of different orders of magnitude depending on several factors (e.g. the problem complexity, the number of private features needed by the model, the importance of the private features in the model, to name a few).
3. The features contributing most significantly to a model can be private/confidential features. This is the case of UC2, where the fuel cost has been the most important feature of the model but it had to be approximated because, due to its high sensitivity for the air transport industry competition model, this feature was not available in the experiments. In those cases where these significant private features cannot be easily approximated, the federated learning approach can enable significant model performance improvements.

A limitation in the shown experiments is the private data used was provided by only one airline (i.e. SWISS) and for a reduced sample of flights (scenarios limited in space and time). It is expected that the addition of more data from more airspace users will lead to additional significant model improvements. Future work should include more realistic experiments with a larger number of airspace users contributing with their private datasets to the federated alliance. Additional use cases can be also explored in different domains of ATM and air transport, potentiallyt including U-Space and multimodality.

Acknowledgment

The authors would like to acknowledge the support of the SESAR 2020 programme through the AICHAIN project (grant no. 894162). Similarly, the authors would like to acknowledge the contributions from Ramon Dalmau, Federica Lionetto, Joao Banha, and Manuel Mateos in the development and experimentation of the machine learning models.

References

[1] SESAR Joint Undertaking, 2020, "The Digital European Sky, Blueprint", doi:10.2829/61772

[2] SESAR Joint Undertaking, 2020, "SESAR Automation in Air Traffic Management".

[3] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

[4] Sesar 2020 exploratory research: Aviation. AICHAIN. (n.d.). Retrieved September 9, 2022, from https://www.aichain-h2020.eu

[5] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Inger- man, A., Ivanov, V., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H. B., et al. Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046, 2019.

[6] TensorFlow Federated : https://www.tensorflow.org/federated, https://github.com/tensorflow/federated [Last accessed: Feb 5, 2021]

[7] H. Brendan, et al., 2017, "Communication-Efficient Learning of Deep Networks from Decentralized Data", arXiv:1602.05629.

[8] Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S., Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, Lecture Notes on Computer Science, 2019.

[9] O. Spjuth, L. Carlsson., N. Gauraha., Aggregating Predictions on Multiple Non-disclosed Datasets using Conformal Prediction, ArXiv 1806.04000, 2018.

[10] Gauraha, N. and Spjuth, O., Synergy Conformal Prediction, DiVA preprint. 360504 (2018). URL: urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-360504.

[11] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How To Backdoor Federated Learning. arXiv:1807.00459 [cs], August 2019.

[12] C. Song, T. Ristenpart, and V. Shmatikov, Machine learning models that remember too much, ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017.

[13] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Inference attacks against collaborative learning", arXiv preprint arXiv: 1805.04049, 2018.

[14] J. Weng, J. Zhang, M. Li, Y. Zhang and W. Luo, "DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-based Incentive," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2019.2952332.

[15] EUROCONTROL, 2010. Airport CDM network impact assessment. Technical Report.

[16] EUROCONTROL 2005, AIRPORT CDM Cost Benefit Analysis (CBA). Technical Report.

[17] Pickup, S. and Huet, D., 2017. Airport – Collaborative Decision Making (A-CDM) Local and Network Impact Assessment. Proceedings of 12th ATM R&D Seminar

[18] R. Dalmau, F. Ballerini, H. Naessens, S. Belkoura, S. Wangnick, 2021, An explainable machine learning approach to improve take-off time predictions, Journal of Air Transport Management.

[19] Martín Martínez, I., Mateos Villar, M., García, P., Herranz, R., García Cantú-Ros, O., & Prats Menéndez, X. (2020). Full-scale pre-tactical route prediction: machine learning to increase pre-tactical demand forecast accuracy. In ICRAT 2020: papers & presentations.

[20] Mateos Villar, M., Martín, I., Alcolea, R., Herranz, R., García Cantú-Ros, O., & Prats Menéndez, X. (2021). Unveiling airline preferences for pre-tactical route forecast through machine learning. An innovative system for ATFCM pre-tactical planning support. In Proceedings of the 11th SESAR Innovation Days.

[21] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[22] Morgan Ekmefjord, Addi Ait-Mlouk, Sadi Alawadi, Mattias √Ökesson, Prashant Singh, Ola Spjuth, Salman Toor, Andreas Hellander, 2022. Scalable federated machine learning with FEDn. Version V2. Published in arXiv. Doi: 10.48550/ARXIV.2103.00148