Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 72 (2023) 2768–2775

Transportation Research Procedia

www.elsevier.com/locate/procedia

Transport Research Arena (TRA) Conference

Promoting sustainable and personalised travel behaviours while preserving data privacy

Noela Pina[a*†], Cláudia Brito[b†], Ricardo Vitorino[a], Inês Cunha[a]

[a]*Ubiwhere, lda., Travessa Senhor das Barrocas, 38, Aveiro, 3800–075, Portugal*
[b]*INESC TEC Universidade do Minho, R. da Universidade, Braga, 4710-057, Portugal*

Abstract

Cities worldwide have agreed on ambitious goals regarding carbon neutrality; thus, smart cities face challenges regarding active and shared mobility due to public transportation's low attractiveness and lack of real-time multimodal information. These issues have led to a lack of data on the community's mobility choices, traffic commuters' carbon footprint and corresponding low motivation to change habits. Besides, many consumers are reluctant to use some software tools due to the lack of data privacy guarantee. This paper presents a methodology developed in the FranchetAI project that addresses these issues by providing distributed privacy-preserving machine learning models that identify travel behaviour patterns and respective GHG emissions to recommend alternative options. Also, the paper presents the developed FranchetAI mobile prototype.

Keywords: Emissions monitoring; Citizens engagement; Data Privacy; Artificial Intelligence; Sustainable cities and communities; Multimodal transport;

1. Introduction

According to the World Economic Forum (WEF, 2022), "mobility is a fundamental human need, and an essential enabler of prosperity, but the current mobility paradigm is not sustainable". Quoting WEF, car travel causes millions of deaths every year, with a significant amount of Greenhouse Gas (GHG) emissions being transport-related and congestions causing heavy financial losses. The global mobility system is in the early stages of massive transformation worldwide, as novel technologies enable innovative related businesses. Moreover, policymakers seek ways to foster mobility that becomes smarter, cleaner, and more inclusive. The European Commission also acknowledges that

* Corresponding author. Tel.: +351-239-151-993
E-mail address: npina@ubiwhere.com
† Both authors did the same amount of work.

10.1016/j.trpro.2023.11.819

transport is the leading cause of air pollution in cities (EC, 2016). Cities worldwide have agreed on ambitious goals towards 2030 regarding GHG emissions and carbon neutrality. Based on that ambitious roadmap, smart cities face challenges regarding active and shared mobility due to public transportation's low attractiveness and lack of real-time multimodal information for citizens (that integrates public transport). These struggles have increased the community's lack of awareness of their mobility choices' carbon footprint and low motivation to change habits.

Additionally, according to Cisco's 'Consumer Privacy Survey' (CISCO, 2020), almost half of the consumers (48%) feel they do not have control over their data. Misuse or abuse of personal data is the top reason consumers lose trust in a company.

This study proposes a methodology created under the FranchetAI project (FranchetAI, 2022) to develop a solution that promotes personalised, sustainable travel behaviours while preserving data privacy and user trustworthiness. FranchetAI built the methodology on top of the following pillars: (i) state-of-the-art mechanisms to safeguard data collected from smartphones by not sharing private/sensitive data with any cloud service; (ii) compliance with European best practices in usability, accessibility and explainable AI to clarify in an understandable way how the data is being processed and how the results are achieved, and, finally, (iii) building up the experience on gamification and habit changing to promote incentives (rewards, vouchers, among others) to encourage the community to opt for sustainable trip choices as well as to create more awareness amongst them.

Based on this methodology, the final solution creates awareness of the saved emissions with periodic "Carbon digest" reports (daily and weekly) via a mobile app that showcases the individual environmental impact of their transportation decisions. The application also recommends sustainable alternatives that produce fewer or no emissions based on the available transit options in a city.

Leveraging explainable AI tools, plus usable and accessible interfaces, the methodology follows a user-centric approach to make data understandable by the various stakeholders (commuters, municipalities, transportation operators). Artificial Intelligence (AI) models process the data presented, which understand different mobility options and their environmental impact.

A different model responsible for GHG emissions estimates the users' carbon footprint. For instance, the Machine Learning (ML) model that predicts the user modal choice considers standard sensor data from smartphone devices (GPS/GNSS, accelerometer, etc.). It comes (off-the-shelf) trained to identify different types of transportation (car, bicycle, walking, etc.). To determine if a user is taking a specific bus or train route (or to recommend one afterwards), the solution requires data on the public transit network in GTFS (GTFS, 2022) format to train its AI models on the routes and schedules. The application then engages commuters to take greener options via gamification mechanisms (by comparing individual reports with the averages of local and European communities, as well as with the necessary targets to avoid climate changes), but especially with rewards/incentives via local challenges promoted and funded by NGOs, decision-makers and businesses willing to invest in carbon reduction. By nurturing these sustainable travel behaviours and changing commuters' habits, this methodology aims to contribute to CO_2 emissions reduction directly.

2. Methodology and main contributions

The methodology was split into two objectives to offer an encompassing solution: the AI approach and the GHG approach. First, to automatically understand the user's mobility patterns, we resort to Artificial Intelligence mechanisms, namely Federated Learning (AI). Secondly, after defining the type of transportation chosen by the user, we need to infer its carbon footprint (GHG).

With this, in the first step, we explain the need to use machine learning algorithms on top of mobility data and how we can focus on the privacy of the users' data while highlighting important information from it. Moreover, we also acknowledge the need to have prior knowledge about several aspects of transportation methods to measure GHG emissions correctly. Fig. 1 exhibits the pipeline of the proposed methodology.

2.1. AI Approach

Following the enormous deluge of data, Machine Learning became the de facto solution to leverage large quantities of data and extract insights from it. Nonetheless, new regulations (e.g., GDPR) impose new approaches to analysing data that may contain sensitive information. Also, Federated Learning (FL) has been emerging when focusing on not

independent and identically distributed (Non-IID) data (Bonawitz et al., 2019; Li et al., 2020). Such data is commonly generated on edge, local and mobile devices.

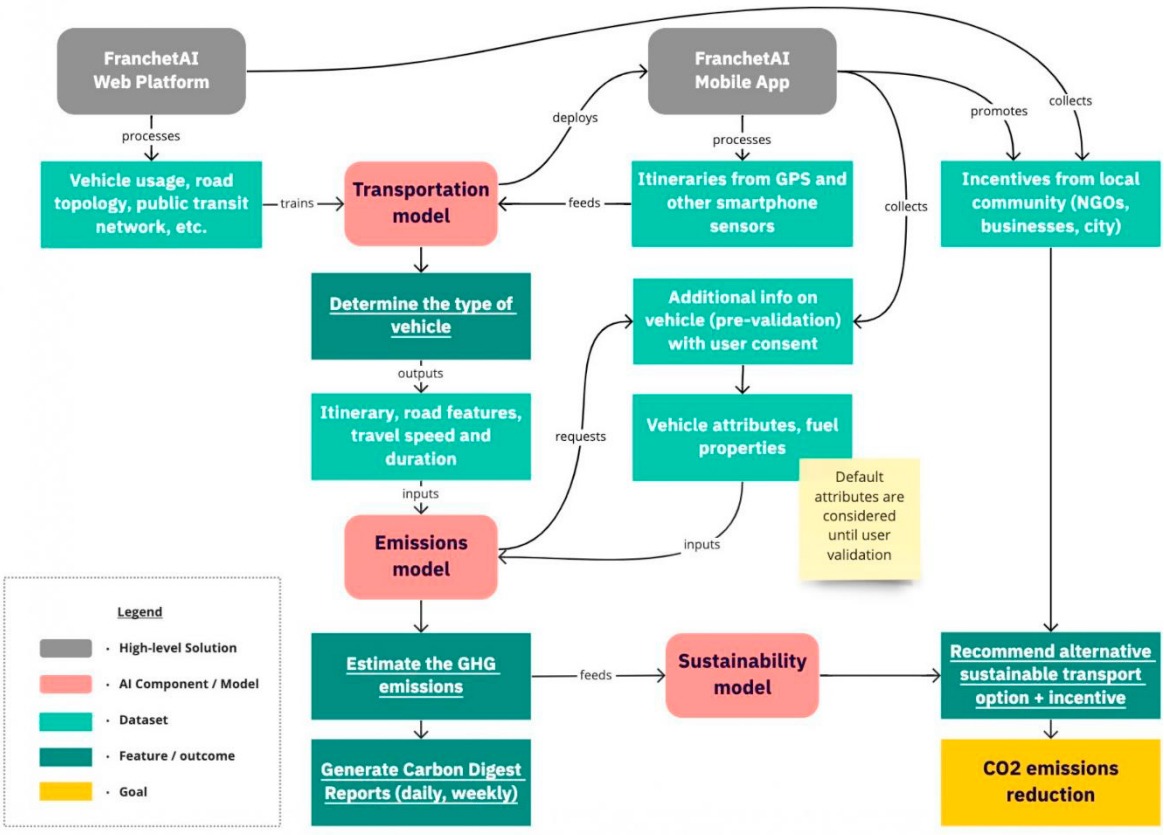


Fig. 1. The pipeline of the proposed methodology.

In brief, FL is a type of distributed machine learning where the data never leaves the data source. Sensitive data from different users can be leveraged to train a privacy-preserving model. In this setting, a centralised server has the initial trained model M_i and broadcasts M_i to every user. This trained machine learning model is a collection of parameters and hyperparameters calculated based on the training data. Typically, M_i is trained on previously collected data, open-sourced or private data. On the user side, the user device trains the model on the user's data. At each iteration, the centralised server asks N users for their new model parameters and calculates the average of all the obtained parameters. Each user can define if they will participate in the round or not. Similarly, the centralised server can decline the parameters broadcasted from the decentralised users. At the end of this cycle, the server informs the new parameters of every user, updating their local model (Bonawitz et al., 2019; Li et al., 2020). Also, one of the goals of FL is to preserve data privacy. To do so, FL resorts to differential privacy, adding an amount of noise to the user's data, guaranteeing that individual data cannot be disclosed (Wei et al., 2021). So, our AI approach is decoupled into two stages. First, we define the datasets to train our chosen models and deep learning architectures locally, and then we define the preferred framework for developing our federated system.

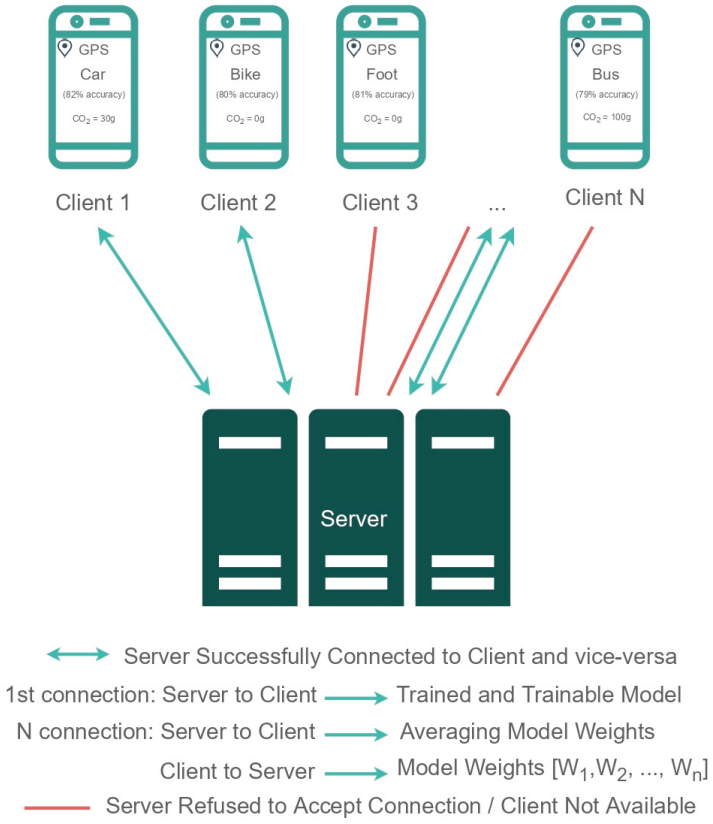


Fig. 2. Representation of the proposed federated system and representative visualisation of the output.

For such a development, we chose the GeoLife GPS Trajectories dataset (Zheng et al., 2011). It comprises GPS trajectory data from 178 users with latitude, longitude and altitude, containing 17,621 trajectories. This data defines which information we need the users to collect. Then, by working with such data, we describe the trajectory's velocity, acceleration, and distance as the main features to be calculated and used as the input of our models. Moreover, such a dataset encompasses several modes of transportation, and we focused on fewer transportation modes, e.g., car, motorcycle, bike, bus, or foot, to improve the accuracy of the models. This approach leads to the model more accurately labelling the transportation mode used by each user.

For the initial models, we used the Random Forest ($n = 15$, number of estimators) and Decision Trees (with a minimum of 2 tree splits), resorting to the library SciKit-Learn (Scikit-Learn, 2022). After using standard machine learning models, other algorithms, with emphasis on deep learning, were used. This was due to the federated learning system (further explained below). These models were built based on the assumption that the initial data will be loaded as CSVs and a time-series database. To this end, the models were developed with dense and long-short term layers. The literature supports the use of these architectures for similar use cases.

We focus our efforts on using state-of-the-art frameworks to offer the proposed encompassing solution. In the first stage, we tested our solution on top of PySyft (Ryffel et. al, 2018; Ziller et. al, 2021). However, due to the launch of its new version, it was essential to understand the changes and how they would influence the following development. After a first evaluation, it was possible to understand the lack of support from the new version to mobile settings. As such, there was a need to evaluate similar solutions, namely Flower (Beutel et al., 2020; Flower, 2022) and Tensorflow Lite (Tensorflow, 2022). While developing the first deep learning models, Tensorflow was the chosen framework. As such, both federated systems were valid alternatives. Nonetheless, even though the integration with the mobile APP still occurs, resorting to the compilation abilities of Tensorflow Lite, Flower provides integration with several machine

learning frameworks. This solution allows the training of federated algorithms not only on mobile settings. After this first stage, it was possible to define a federated system seamlessly for other models and use cases focusing on sensitive data, e.g., data obtained from each municipality or city. Flower also provides several averaging algorithms which comprise differential privacy alternatives. This main feature improves the solution's privacy-preserving goal and its trustworthiness. With this in mind, Fig. 2 represents the overall layout of the federated system.

Moreover, we acknowledge the need to employ Explainable AI tools to let users understand how their data is used. To do so, we apply the SHAP framework. Such a framework visually explains which information is used to train the models. Since we limit the number of labels and data used for this training, this Explainability allows the users to understand which features impact the outcome/output of the model.

2.2. GH Approach

The methodology adopted for estimating Greenhouse Gases (GHG) and air pollution emissions and measuring their reduction is based on the "tank-to-wheel" Life-Cycle Assessment (LCA), thus, only considering the operation of the vehicle. The emissions are estimated based on the CORe INventory AIR emissions (CORINAIR) system, i.e. the method approved by the European Environment Agency (EEA) to assess emissions. CORINAIR adheres to the IPCC guidelines (IPCC, 2006), used globally by environmental protection agencies for national and regional evaluations. According to the IPCC Guidelines for greenhouse gasses, a compiler builds a decision tree to select the appropriate methodology with different complexities and data requirements. Therefore, we apply the Tier 3 methodology from EMEP/EEA (EMEP/EEA, 2019) (formerly called the EMEP CORINAIR emission inventory guidebook) by considering equation 1. Moreover, the GHG estimations are based on the ultimate CO_2 emissions, which result from different processes (combustion of fuel; combustion of lubricant oil; and addition of carbon-containing additives in the exhaust).

$$E_{ik} = e_{ik}(v) \cdot a_k \tag{1}$$

where E_{ik} is the exhaust emissions of pollutant *i* induced by a vehicle technology *k* (in grammes); e_{ik} is the emission factor as a function of the vehicle driving speed (in grammes per kilometre); a_k is the transport activity in vehicle kilometres travelled (VKT) for vehicle technology k. The emissions are calculated individually for each client of the FranchetAI Mobile App by considering the average driving speed of the road links that constitute an individual trip. The previous AI approach provides information on traffic data (modal choice, trip route and distance and driving speed) necessary to calculate clients' emissions from traffic activity. Also, information on vehicle technology is required, i.e. the Euro Standard information, accessed by taking into account the age of the vehicle. Therefore, a user is asked to give this detailed information; otherwise, a default technology is used (Euro 4).

3. Results and Discussion

3.1. AI Approach

The focus of the first tests with the GeoLife dataset helped limit and create general labels for the transportation modes. We partitioned the initial dataset into test and train datasets, 30% and 70%, respectively. The first 70% of the dataset was used to train the model, while the remaining 30% was used to test it. Also, while limiting the used labels, we could diminish the tree length and improve the results of Random Forests and Decision Trees by around 5%, reaching an accuracy of 81% in less than 3 minutes of training. Moreover, the developed Deep Learning models, based on dense and long-short term memory layers, have yet to achieve similar results. Nonetheless, the results are identical to the previous ones, reaching an accuracy of 75% with a training run time of 5 minutes. Focusing on the current tests based on the proposed federated system, we followed a similar dataset definition. We further sharded the test dataset into ten shards to test the system with ten clients. Such initial results were promising, showing that the trained model can label the remaining trajectories with similar accuracy. With 1/10 of the dataset, each user can prepare a new local model and share its parameters with the centralised server.

The Explainability feature allows us to understand the weight given to each feature to label each class (Fig. 3). With classes (e.g., car, foot, bus) being the output (of the model), the mean velocity and the distance of each trajectory are the features of the model.

Although such a solution presents a viable option for mobile settings, the proposed models need yet to be comparable to others from the state-of-the-art. Further information could be leveraged to improve these models. Also, other data features can depend on what the mobile device collects on the user side. To this end, new tests should be made available, and new deep learning model architectures may be defined.

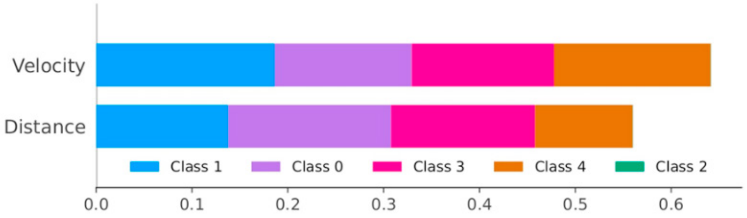
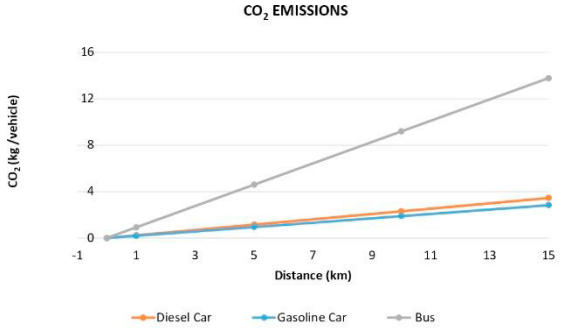


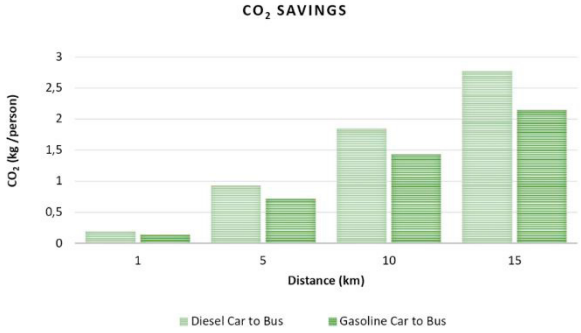
Fig. 3. Example of the current Explainability outcome.

3.2. Example GHG Approach

The emissions model currently proposed needs the following information: *i*) the mean velocity of the trajectory; *ii*) the type of fuel of the car; *iii*) the category of the vehicle (i.e., Passenger, Bus, Heavy Duty and, Motorcycle); *iv*) the total distance of the trip; *v*) the year of the vehicle. To calculate such emissions, we need the trajectory's velocity and total distance, together with the type of vehicle and its category. The year and category are further used to define the Euro Standard. Still, when the user does not disclose such information, the GHG emissions are estimated based on Euro 4, and the previously trained AI models define the vehicle category.



(a) CO_2 emissions of a user commuting with a diesel car, gasoline car and a bus.



(b) CO_2 savings of a user commuting with a diesel car, gasoline car and a bus.

Fig. 4. Example of the proposed solution for GHG emissions.

Fig. 4 presents an example of the usage of our emissions model. For instance, for a user commuting for 30 min at 30km/h, one may analyse the CO_2 induced by different vehicle modes (diesel/gasoline car or bus) in kilograms per vehicle. Also, by assuming that an urban bus will have an occupancy of 20 people, the impact of using such a more environmentally friendly vehicle is presented.

3.3. FranchetAI Mobile Prototype

The Cotoneaster franchetii inspires FranchetAI (see Fig. 5), a super plant acknowledged for filtering 20% more emissions (namely automobile air pollutants) than other shrubs. FranchetAI is a digital rewarding mechanism for people opting for sustainable mobility options (public transit, electric vehicles, and other light modes), ensuring transparency and trustworthiness between the user and the different stakeholders creating the incentives. Overall, such a solution aims to let users understand their carbon footprint while offering travelling alternatives and rewards for travelling more sustainably by changing their current habits.

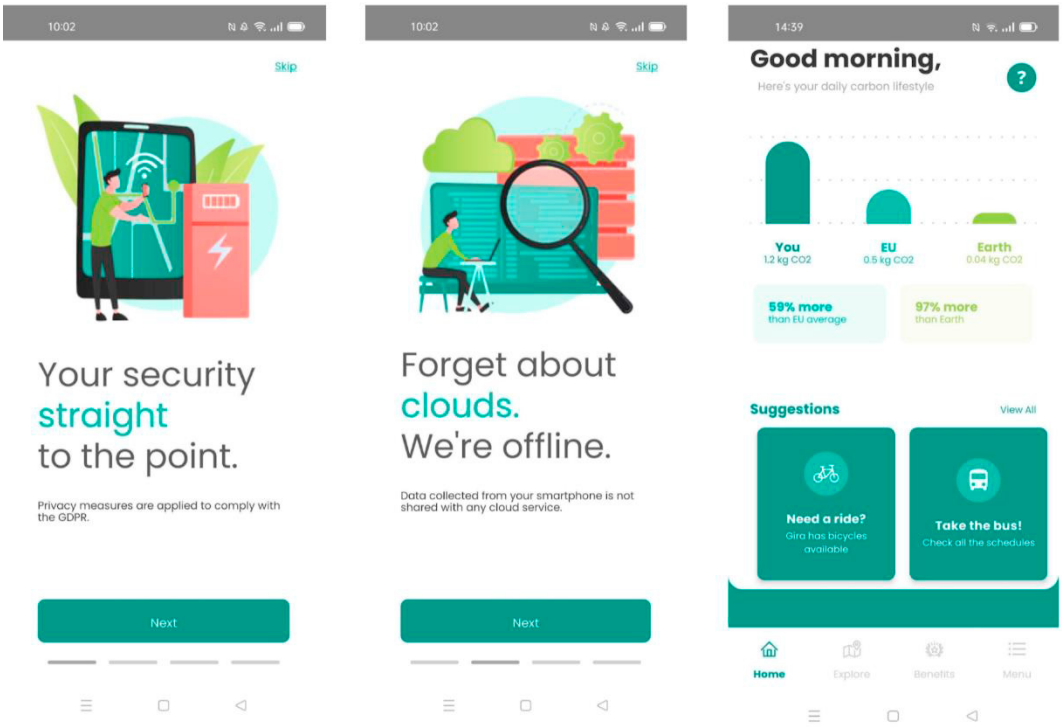


Fig. 5. FranchetAI Mobile App.

To improve the models that classify the transportation modes and the emissions, a few attributes and parameters are sent to our server to retrain them. These do not reflect the user's location or any other personal data that might have been shared with our application, specific to the model parameters. Hence, data is collected from a smartphone only with the user's consent and is not shared with any cloud service. Besides, FranchetAI helps increase the users' awareness of their transport environment impact choices while rewarding them when a "good behaviour" is made, provided by local stores and services. Therefore, FranchetAI plays a crucial role in achieving the SDGs regarding climate change and helps build cities' economies while promoting local businesses. This prototype must be fully implemented and deployed into the pilot stage for a complete proof-of-concept solution. With this, the focus will be on the young adult generation, who are typically more prompt to test new environment-aware solutions. This stage will also allow us to try and improve the feasibility of such a novel solution.

4. Conclusion

This methodology leverages best practices for differentiating on privacy and engaging citizens. Specifically, it uses AI/ML models to classify personas based on user feedback collected in the mobile app and provide recommendations on routing options that produce fewer/no emissions. Also, the solution uses AI/ML models to detect the transportation

modes based on user content and data-sharing consent and to understand the impact and quality of incentives. The users have complete control of their data, knowing which data is used for locally inferring the system's models and which is used for training new models in a secure and privacy-preserving manner. Following a federated learning setting and security protocols, sensitive data must not leave the users' premises at any given moment.

Regarding future work, the methodology will be validated within real-world scenarios. Data from past and ongoing initiatives (namely other R&D projects) is being used as well as open data and third-party platforms to ensure the solution is as off-the-shelf as possible (although local context and data will help personalise it to the target communities). Also, we want to evaluate different strategies for the federated system and understand the impact of adding differential privacy to real-world use cases.

Acknowledgements

The FranchetAI project is part of the AI4 Cities project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 871914. This work was also supported by the Portuguese Foundation for Science and Technology through a PhD Fellowship (SFRH-BD-146528-2019 – Cláudia Brito).

References

CISCO, 2020. "2020 Consumer Privacy Survey". Retrieved at: https://bit.ly/3lmoap5

EC (European Commission), 2016. "A European Strategy for low-emission mobility". Retrieved at: https://ec.europa.eu/clima/policies/transport_en

EMEP/EEA, 2019. "Exhaust Emissions from Road Transport. Passenger Cars, Light-Duty Trucks, Heavy-Duty Vehicles Including Buses and Motor Cycles". European Monitoring and Evaluation Programme (EMEP), Air Pollutant Emission Inventory Guidebook 2019, EEA Report No 13/2019

FranchetAI, 2022. "Absorbing Traffic Pollution with AI". Retrieved at: http://franchet.ai

GTFS, 2022. General Transit Feed Specification. Retrieved at: https://developers.google.com/transit/gtfs

IPCC (The Intergovernmental Panel on Climate Change), 2006. "2006 IPCC Guidelines for National Greenhouse Gas Inventories". ISBN 4-88788-032-4

WEF (World Economic Forum), 2022. "Shaping the Future of Mobility". Retrieved at: https://www.weforum.org/platforms/shaping-the-future-of-mobility

Zheng Yu, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li, 2011. "Geolife GPS trajectory dataset-user guide." Geolife GPS trajectories 1 (2011).

Bonawitz, Keith, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon et al., 2019. "Towards federated learning at scale: System design." Proceedings of Machine Learning and Systems 1 (2019): 374-388.

Li, Tian, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, 2020. "Federated learning: Challenges, methods, and future directions." IEEE Signal Processing Magazine 37, no. 3 (2020): 50-60.

Wei, Kang, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H. Vincent Poor, 2020. "Federated learning with differential privacy: Algorithms and performance analysis." IEEE Transactions on Information Forensics and Security 15 (2020): 3454-3469.

Scikit-Learn.org, 2021. "Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.0.2 Documentation." https://scikit-learn.org/stable/.

Beutel, Daniel J., Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D. Lane, 2020. "Flower: A friendly federated learning research framework." arXiv preprint arXiv:2007.14390

TensorFlow, 2022. "TensorFlow Lite | ML for Mobile and Edge Devices.". 2022. https://www.tensorflow.org/lite.

Ryffel, Theo, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach, 2018. "A generic framework for privacy-preserving deep learning." arXiv preprint arXiv:1811.04017

Ziller, Alexander, Andrew Trask, Antonio Lopardo, Benjamin Szymkow, Bobby Wagner, Emma Bluemke, Jean-Mickael Nounahon et al, 2021. "Pysyft: A library for easy federated learning." In Federated Learning Systems, pp. 111-139. Springer, Cham.

Flower. 2022. "Flower: A Friendly Federated Learning Framework." Flower.dev. https://flower.dev/.