**MSc in Data Analytics (SB+) - Sept 2023 - 2024 - YR1**

Author: Jose Maria Rico Leal

e-mail: sba23021@student.cct.ie

Student ID: sba23021

Federated Learning for Data Privacy Enhancement in Key Industries: A Machine Learning Approach.

## **Abstract**

*This research proposal comprises the key components for the summer capstone project in an early version. All necessary sources have been reviewed, the sampling strategy and primary research methodology have been proposed, and ethical aspects have been addressed to align with European regulation.*

[**Abstract** 1](#_Toc166439239)

[**1.** **Introduction** 3](#_Toc166439240)

[**2.** **Proposed title and topic area of Data Analysis project** 3](#_Toc166439241)

[2.1. Title 3](#_Toc166439242)

[2.2. Topic area 3](#_Toc166439243)

[**3.** **Research Objectives and topic alignment** 4](#_Toc166439244)

[3.1. Research Objectives 4](#_Toc166439245)

[3.1.1. Evaluate the performance of existing Federated Learning frameworks 4](#_Toc166439246)

[3.1.2. To develop a Federated Learning system 4](#_Toc166439247)

[3.1.3. Comparative analysis 4](#_Toc166439248)

[3.2. Topic alignment 4](#_Toc166439249)

[**4.** **Literature review** 5](#_Toc166439250)

[4.1. Introduction 5](#_Toc166439251)

[4.2. Federated Learning frameworks 5](#_Toc166439252)

[4.3. Distributed Machine Learning 9](#_Toc166439253)

[4.4. Federated Learning implementation 11](#_Toc166439254)

[4.5. Commonly used datasets 12](#_Toc166439255)

[4.6. Real world Federated Learning scenarios 12](#_Toc166439256)

[4.7. Conclusion 13](#_Toc166439257)

[**5.** **Proposed sampling strategy, population, sampling method and type** 14](#_Toc166439258)

[**6.** **Proposed primary research methodology** 15](#_Toc166439259)

[**7.** **Ethical and legal/regulatory aspects sourcing and analysing the data. Strategies. Addressing these aspects according to the European Commission Ethics and Data Protection handout.** 16](#_Toc166439260)

[**8.** **Bibliography** 18](#_Toc166439261)

# **Introduction**

The concept of Federated Learning emerges from training decentralized machine learning models while maintaining data privacy (H. Brendan McMahan et al., 2017).

Imagine various hospitals that need to train a model with patient data; here, data privacy becomes a concern. Federated Learning addresses this by distributing the training process across each hospital and then sending the results to a central server that aggregates and updates the global model (Guan et al., 2024). This concept does not only apply to the healthcare sector; it is also embedded in today society. Each mobile phone acts as a node, training global models for technological companies that operate on them (Kang et al., 2020).

This field is growing rapidly with many frameworks available to implement Federated Learning; however, these frameworks are often difficult to implement. By the end of this project, a Federated Learning system will be created to demonstrate the benefits of this recently developed concept.

# **Proposed title and topic area of Data Analysis project**

# Title

Federated Learning for Data Privacy Enhancement in Key Industries: A Machine Learning Approach.

# Topic area

Use Federated Learning to train a model and determine how existing frameworks compare to the one developed in this project, finally assessing whether key industries, such as the healthcare or the technological sector can benefit from it.

# **Research Objectives and topic alignment**

# Research Objectives

# Evaluate the performance of existing Federated Learning frameworks

By evaluating Federated Learning frameworks, such as *TensorFlow Federated, Flower, EasyFL, IBM Federated* and *FLGo,* this section assesses the performance of each framework architecture in conjunction with deep learning models.

# To develop a Federated Learning system

In some cases, existing FL frameworks tend to be difficult to implement, this objective bridge the gap, showing that a FL system can be built in a pragmatic and intuitive manner. Primary research as experimentation and the literature review have helped developing this system, primarily sections distributed machine learning and grey materials. This includes experimentation with deep learning models suitable for distributed learning, aggregation, and image processing. Their performance will be evaluated comparing their metrics to finally evaluate and deploy a Federated Learning system.

# Comparative analysis

This section compares existing framework and the developed Federated Learning system and discusses about the pros and cons that each can bring to key industries, especially healthcare and technology sectors.

# Topic alignment

The objectives are carefully designed to closely align with the topic, which is how the industry is addressing privacy when sharing data to train global models with third parties. By evaluating the performance of existing Federated Learning frameworks, the aim is to understand and evaluate the solution the industry gave to train global models without privacy being compromised.  
The second section developing a Federated Learning framework, focuses on experimentation of suitable deep learning models to construct a robust Federated Learning system. Primary research helped on finding grey materials to develop this system, there is a gap between existing FL frameworks and the usage of them. Often these are difficult to put in practice and abstract, however this developed FL system will address this gap in terms of usage and practicality.  
And the final objective compares existing and developed FL frameworks, the aim is to determine the benefits that each can bring to the industries when privacy is compromised by distributed models being trained.

# **Literature review**

# Introduction

This literature review comprises thirty sources, structured using methodology as the organizing principle. Throughout the research, five themes have been identified as follows: Federated Learning frameworks, Distributed Machine Learning, Federated Learning implementation (grey materials), commonly used datasets, and real-world Federated Learning scenarios.

The literature review is key to this proposal as it sheds light on many aspects of the FL. Starting with the frameworks, a total of fourteen sources have been reviewed. The criteria for selecting these sources were based on publication date and whether the framework is widely used. All included sources are recent, and the frameworks are extensively used by both industry and researchers. Some of the aspects considered include how user-friendly these frameworks are, which machine learning algorithms they use, and how the FL is carried out (centralized, decentralized, or vertical).

For the second section, a total of five sources have been reviewed. This section aligns with the second research objective and introduces the idea that a potential FL system can be built within a distributed file system *(Hadoop).* Datasets stored across different directories will emulate the clients. Using *MapReduce* or *Spark* as a central server, datasets can be trained and aggregated into a global model.

In the third section, two YouTube videos serve as grey materials. These demonstrate in a straightforward manner how to deploy a FL system, maintaining client privacy with ease while training the global model.

There are many FL datasets primarily used for research purposes, while those used within the industry are kept private for obvious reasons. The purpose of this section is to identify the most popular FL datasets based on their usage in FL frameworks. These datasets will also serve as part of the experimentation in this project. A total of four sources have been reviewed in this section.

The last section, which has reviewed a total of five sources, has been instrumental to the third research objective. It enhances understanding of how real-world industries deploy their FL systems.

# Federated Learning frameworks

Zhuang *et al.,* (2022) developed *EasyFL* designed a low code platform to assist beginners and researchers to experiment and prototype FL artefacts. It offers practical functionalities such as handling heterogeneity, simulation, comprehensive tracking, optimization of distributed training, and seamless deployment. While numerous FL platforms have been developed by institutions and companies, these are difficult to implement. *TensorFlow Federated (TFF)* is deployable but does not optimize distributed training. On the other hand, *FATE* supports deployment but is not user-friendly, presenting high entry barriers. In contrast, *EasyFL* is user-friendly and facilitates efficient experimentation along with seamless deployment. It also supports diverse training methods, including standalone, distributed, and remote. Also, Beutel *et al.,* (2021) presented a user-friendly framework, *Flower*, an open-source framework designed to enhance the capabilities and practical implementation of FL, across different environments. Flower addresses the challenges associated with the deployment and scalability of FL. Flower architecture allows simulations in real-world scenarios, making it a great tool for federated learning scenarios. The framework is agnostic supporting different machine frameworks like TensorFlow and PyTorch. Comparing Flower to other federating learning frameworks such as *TFF, PySyft, FedScale*, and *LEAF*. Flower stands out by supporting actual deployment on real devices rather than just simulation, unlike *TFF* and *LEAF*, which mainly focus on simulated environments. Continuing with open-source platforms, Ziller *et al.,* (2021) introduced *PySyft* a multi-language library that facilitates secure and private ML. It was developed by the *OpenMined* community with the objective of making FL data science more accessible through Python bindings and user-friendly interfaces. *PySyft* uses libraries like *PyTorch* and TensorFlow with additional capabilities. Comparing it with other frameworks like *TFF* and *PaddleFL*; *PySyft* offers detailed building blocks, allowing developers to implement FL efficiently. Also compared to Flower that supports heterogeneous client environments and offers tools for mobile and edge devices, claiming and advantage over *PySyft* in these aspects.

Wang *et al.* (2023) developed *FLGo* a platform designed to streamline the process of cross-application FL research and enhance shareability among developers. It is a lightweight FL framework aiming to be a customizable solution to suit different applications and data heterogeneity. *FLGo* addresses the gap that exists in current FL frameworks which often make the FL deployment very complex.  
Some of the key Features, are benchmarks and algorithms, customization, experimental tools, and high degree of shareability. Compared to other frameworks it stands out in, system heterogeneity, high-level API, multi-architecture support, asynchronous operations and customization and flexibility. As a conclusion *FLGo* has been developed with the intention of making FL more accessible to a broader range of developers by simplifying customization and enhancing its shareability. It also aims to bridge the existing gap with conventional machine learning and FL. Another innovative framework *OpenFL* (Anthony et al., 2021) created by Intel Labs and the University of Pennsylvania, OpenFL supports decentralized machine learning models. It allows organizations to train models using data locally without any transfer, and that operates by distributing a global model across various nodes while each organization trains its model locally. Model updates are sent to an aggregator to enhance the global model. This framework is compatible with popular ML frameworks like TensorFlow and PyTorch. In comparison to other frameworks, it stands out due to its open-source nature, TensorFlow Federated or PySyft focus more on academic research applications while OpenFL is focused on real-world applications.

Kraska *et al.,* (2015) designed *MLbase* Some of the key features are declarative language, optimization, and distributed runtime. Among its functions are supporting tasks like classification, regression, and complex procedures like collaborative filtering and exploratory data analysis. This functionality extends to a variety of domains including healthcare, music recommendation, and social media analysis. Compared to systems like Google Prediction API, Weka, and *SystemML*, *MLbase* is unique integrating a distributed system with a focus of optimizing machine learning tasks, rather than merely supporting them. This framework is in continuous improvement as takes user feedback to enhance model accuracy and utility over time. Also, it has a high-level API, for developers to deploy their artefacts. Similarly, Iterative Clustering Algorithm (*IFCA*) was created by Ghosh *et al.,* (2022) where users are divided into clusters based on the similarity of their objectives or learning tasks. This model hopes to enhance efficiency by leveraging the similarities among users within the same cluster to perform more effective learning. It differs from traditional FL models in the way that *IFCA* provides personalization by learning distinct cluster models with similar data distributions. It avoids computational overhead of centralized clustering by distributing the clustering process.

Fan *et al.,* (2023) developed presents *FATE-LLM*, an industrial-grade FL framework designed to address the challenges associated with training large language models (*LLMs*) such as *GPT* and *BERT*. The main challenge is the high computational resources required and the need for high-quality data. This framework offers tools to facilitate the training of large language models in a distributed manner without compromising privacy. It compares to other FL frameworks by its specific focus on large language models. TensorFlow Federated or *PySyft*, are more generalized, *FATE-LLM* focuses on *LLMs*, it offers advanced techniques like federated intellectual property protection and parameter efficient fine tuning. The framework leverages advanced techniques such as knowledge distillation and model quantization to optimize training.

According to Cheng et al., (2019) *SecureBoost* offers unique contributions such as a lossless privacy preserving framework that builds tree-boosting models across multiple parties. Desing based on gradient tree boosting methods while keeping all training data local, without the need for a trusted third party. It has encryption to ensure data privacy is maintained throughout the learning process. It compares to other models due to its lossless nature, meaning that it achieves the same level of accuracy as centralized models without the need of centralized data. It uses vertical partition data whereas popular Google’s FL framework focuses on horizontal partitioned data. In terms of performance *SecureBoost* outperforms tree boosting methods like *GBDT* and *XGBoost*.

Solanki *et, al.,* (2022) delve into how *TensorFlow Federated* *(TFF),* an open-source framework, is utilized for machine learning on decentralized data. It has been designed for research and experimentation. Some of the key features are *TFF* enables FL through low-latency models with less power consumption. The framework uses two layers, the FL learning API and the federated core (FC) API. The FL API allows developers to implement training and evaluation on existing TensorFlow models through a high-level interface. The FC API integrates *TensorFlow* with distributed communication operators focusing on computations across distributed systems like mobile phones, tablets, and sensors. Comparing *TFF* to other frameworks, it offers a unique well integrated structure others do not provide this level of integration. *TFF* allows experimenting with new algorithms is not tied to predefined algorithms. A different study that showcases a fairness-aware federated learning algorithm designed to group fairness while computing (Salazar et al., 2023). It incorporates a fairness-aware momentum to compute the global model by considering client model fairness level. Some of the key points are fairness-aware aggregation, momentum term, group fairness and real-world experiments. Compared to existing FL frameworks like *FedAvg* and *FedMom*, *FAIR-FATE* utilizes a momentum-based approach to address fairness specifically. Unlike some methods that require local debiasing strategies for each client, *FAIR-FATE* operates without that requirement. Other frameworks have attempted to aggregate models based on fairness, but *FAIR-FATE* seems the one achieving this. As a conclusion, *FAIR-FATE* is an effective approach for increasing fairness in federated learning models, showing significant improvements in various heterogeneous scenarios.

Chen *et al.,* (2022) introduced a benchmarking framework to evaluate personalized federated learning (*pFL*) algorithms across various non-i.i.d. data distributios and applications. The framework is developed by rese at Alibaba and addresses significant gaps in the standardized testing and systematic analysis of *pFL* methods which suffer from inconsistent dataset usage, diverse evaluation metrics, and non-uniform simulation settings. Some key aspects are a comprehensive coverage *pFL-Bench* has 12 datasets with different data types like images, text, and graphs, all with uniform distributions to mimic real client distributions. Extensible codebase that integrates 20 state of the art *pFL* methods and systematic evaluation, the framework is executed in a containerized environment, ensuring consistency across studies. Unlike traditional FL benchmarks that focus on global model performance, *pFL-Bench* focuses on personalized models. Compared to other *pFL* frameworks like *FATE* or *TensorFlow Federated*, which offer support for extensive personalization, pFL-Bench provides a detailed evaluation of personalized models. In summary *pFL-Bench* offers a solution by introducing rigorous, repeatable, and scalable benchmarks for personalized federated learning methods.

A framework developed by the industry is *FLINT* (Wang et al., 2023), a device-cloud collaborative platform designed to facilitate the seamless integration of federated learning with existing centralized systems. This paper motivation lies on the challenges that growing computation and storage capacities of user devices pose to distributed training. *FLINT* is designed to integrate with LinkedIn’s existing ML platforms, bridging the gap between centralized training and federated deployments. One of the challenges is the heterogeneity of devices, which can often lead to inconsistencies in model performance, *FLINT* addresses this with a framework that models different capabilities and network conditions. *FLINT* provides a robust framework for simulation, evaluation, and deployment, it was also inspired by frameworks such as Google’s *GFL*, Apple’s FL system, and Meta’s *PAPAYA*, it integrates some features from these systems. A different FL framework developed by the industry is *IBM FL* (Ludwig et al., 2020) This framework is designed to facilitate the implementation of federated learning across diverse enterprise environments. It helps users to model without centralizing training data, addressing the key issues of privacy. It supports the integration of Deep Neural Networks and traditional machine learning methods. It also provides tools for design and deployment of federated jobs minimizing the learning curve. *IBM FL* is different from other existing FL frameworks by its focus on enterprise needs, including secure deployment, failure tolerance, and rapid model specification. Also, when compared to other existing frameworks *IBM’s* solution is tailored for multi-cloud or hybrid cloud environments where data privacy is critical. It supports both federated learning systems using a central aggregator and more decentralized models. In conclusion, it provides an effective bridge between traditional centralized data processing and the emerging needs of decentralized, privacy preserving machine learning applications.

# Distributed Machine Learning

Galakatos *et al.,* (2017)provide a detailed overview of how machine learning algorithms are optimized and implemented in a distributed environment to improve performance, increase accuracy, and scale to large datasets. Distributed machine learning enables more informed decision making by processing large amounts of data spread across multiple nodes. Historically, before distributed computing systems like MapReduce smoothed the landscape, developers designed and delivered complex services including data distribution, parallelization, synchronization, and fault tolerance development was slow and prone to errors frequently modern distributed machine learning systems alleviated this burden by automating low-level processing tasks they could focus on the use of ML in a distributed environment and turning single-threaded algorithms into parallel ones, a complex step that requires a deep understanding of the algorithm supervised and unsupervised learning are the main categories two applications. For example, supervised learning might include algorithms such as linear regression, which can predict outcomes based on prior data, while unsupervised learning, such as k-means clustering, predicts data structure without predefined labels Distributed machine learning has many advantages, including the flexibility to deal with large data sets, improving model accuracy through increased computing power, and reducing the time required for data processing through task a distributed over many nodes. This scalable approach allows for handling complex learning tasks more effectively than could be achieved with single-node systems.

According to *Liu, et al.,* (2022) the changeover from distributed machine learning to federated learning systems addresses major concerns on data privacy, security, and conformity to legal requirements mostly regulations. The need to shift is necessitated by the difficulties of centralizing sensitive information because of the laws protecting private life as well as risks that might arise out of transferring raw content through networks. FL provides a way out through allowing localized machine learning model training on user’s gadgets thus sharing only updates or gradients instead of data itself. This not only observes strict data protection laws like GDPR within EU and CCPA in USA but also reduces chances for security breaches during transit. This periodization involves among others creation protocols for safe training which will require minimum exposure of data. To secure personal while being aggregated together for processing federation algorithms should have advanced encryptions methods like advanced encryption techniques, differential privacy, and secure multi-party computation to protect data during the aggregation and training processes. Besides it is expected that there will be variety security enforcement points within FL Systems since they manage different parts during federated training process such as data handling, model training and compliance checks. Moving into of federated Learning systems is considered as huge leap forward in ML sector with specific emphasis placed on technologies that protect confidentiality without compromising availability, this is especially important fact-oriented areas like healthcare finance where importance laid upon data privacy so much more than anything else.

Verbraeken *et al.,* (2020) also examine the critical aspects of the transition from traditional distributed machine learning frameworks to integrated learning frameworks. It highlights how these changes address critical challenges such as data privacy, security, and efficient use of distributed data sources without the need to centralize sensitive data. Growing regulatory requirements around data security require techniques such as differential privacy, secure multiparty computing, and homogeneous encryption to keep data private and secure during the training process, the research includes various architectures and algorithms developed to provide federated learning systems performance has been improved and ways to reduce the objectives of this development. FL has to inherent overcome challenges such as high latency, limited bandwidth, and disparate data distribution among them participated in. Overall, the study presents integrated learning as a robust solution that not only ensures compliance with data security and privacy regulations but also provides a flexible approach and effective machine learning in distributed environments is easy. These changes represent a significant shift towards more sustainable ethical practices in the use of machine learning technologies.

Distributed deep learning effectively leverages Hadoop to efficiently manage the complexities of big data processing (Dev, 2017) which is necessary due to the heavy computing requirements of deep learning Hadoop's architecture support complex data processing across multiple servers. It has basic features like Hadoop Distributed File System *(HDFS),* *MapReduce*, and *YARN* (Yet Another Resource Negotiator). *HDFS* ensures data availability and flexibility by distributing and refreshing data across all nodes in a cluster. MapReduce simplifies data processing by dividing tasks into smaller blocks (Mapper) and combining the results (Reducer), which is particularly suitable for iterative deep learning algorithms This approach reduces the computational burden on individual machines down and speed up the process. *YARN* improves resource utilization by providing dynamic resource allocation throughout the cluster, which improves workflow and system efficiency. Using Hadoop for distributed deep learning offers many advantages including scalability, as it can process large amounts of data across thousands of servers; fault-tolerant, automatically recovering from node failures to maintain system operation and data integrity; the cost of implementing inventory hardware. And the ability to run multiple tasks at the same time increased speed. Additionally, Hadoop’s ability to efficiently handle multiple sources of data increases its flexibility. Integrating Hadoop into a distributed deep learning framework supports efficient training of complex models by leveraging the need for speed and scalability in today’s data-driven applications. These policies not only increase computing performance but also enhance data security and reliability, which is important for them.

A different approach by *Xing et al.,* (2015) the *Petuum* framework, which is designed to facilitate grand-scale distributed ML, applying sophisticated ML algorithms efficiently to technical-sized problems involving large scales and data *Petuum* is innovative in its approach, it goes beyond traditional bulk synchronous processing as well It goes beyond the special graph-based executions commonly used in other systems *Petuum* addresses the challenges associated with data- and model-parallel processing in large MLs using an optimization-focused approach. Most ML schemes and iterative convergence algorithms are inherently fault tolerant. This unique perspective provides an integrated system design, which can implement dynamic scheduling based on the scheduling requirements of *Petuum* ML systems and bounded-error network synchronization Features like these enable the platform to execute ML tasks quickly and large than traditional models. A key advantage of *Petuum* is the ability to handle detailed models and large amounts of data it manages hundreds of billions of parameters over petabytes of data Its focus on optimization provides the ability to streamline tasks, consume data and meeting model-centric needs through fault-tolerant recursive meetings Boundary-error synchronization adapts to changing computational requirements, reducing the burden often associated with rigid consistency models Overall, *Petuum’s* architecture dramatically improves the speed and scope of ML systems, making better use of distributed resources. This makes *Petuum* a powerful tool for researchers and practitioners tackling complex ML operations on large datasets.

# Federated Learning implementation

The dataset, Chest X-Ray Images, is utilized in a model implemented with a multi-layer convolutional network using *Keras*. The platform AI4EOSC will be used (www.youtube.com, 2024). The task involves using the images to implement a *NN* to predict whether new X-rays are normal or indicate pneumonia. The initial dataset is divided among three clients that simulate hospitals wishing to collaboratively develop a global model without sharing patient data. For each client, the training data is divided using a random split of 75% for training and 25% for testing. Using the *AI4* FL server, the model is deployed. The Flower package is used to build the model. The use of *Jupyter Notebook* is recommended; once the server is running, hardware is configured by selecting the number of CPUs, disk memory, and RAM. The next step is federated configuration, which includes the number of rounds, evaluation metric, number of clients, and federated aggregation strategy. On the client side, three instances are opened and running in the cloud. It is important to pass the *AI4* FL server key into each client's *Python* script. After this, by running *$ python3* *Client.py*, each client will be initialized and wait for the third one to start and trigger the entire process. On the server side, accuracy is calculated. This is a great example, and the approach differs from how frameworks typically deploy federated learning; in summary, this example is close to a real-world implementation case.

This other practical implementation (www.youtube.com, 2022) case uses Azure Machine Learning, and the dataset used is the same as discussed above. The experiment begins by training the model with a classical approach and then comparing this result with the federated one. Three clients are present, representing hospitals in the US, Europe, and Asia; these are computer instances in Azure. The Nvidia Flare framework is used in this model. A *Jupyter Notebook* is utilized as a controller, sending instructions to the clients, and tracking accuracy. In conclusion, this demonstration, along with the one above, is fundamental to understanding how FL is implemented. Unlike the first section, 4.2, which emulates the clients within a *Jupyter Notebook*, these two examples use virtual machines that closely emulate actual hospitals. They have different IP addresses and connection protocols must be used to connect and train the model.

# Commonly used datasets

Luo et al., (2021) introduced a real-world image dataset specifically designed to address the challenges associated with non-IID and unbalanced data distributions that are common in federated settings. This dataset comprehends images captured by 26 street cameras, across 900 images categorized into 7 object types. Method focus is the study of implementing and benchmarking two major object detection algorithms, *YOLO* and *Faster R-CNN*. As a result, a non-IDD and imbalanced dataset was created, ideal to test FL models under real-world conditions. This dataset offers a unique resource compared to other benchmark datasets used in FL research. *MNIST* or *CIFAR* are balanced and homogeneous whereas the real-world image dataset provides a real challenge due to its realism, leading to better FL models after training. Also, the use of *YOLO* and Faster R-CNN differs from simpler models or more synthetic setups often used. In summary realistic approaches can enhance FL modelling. Similarly, Lai et al., (2024) used versions of *MINST* and *CIFAR*, *Fashion-MNIST, CIFAR-10 and CIFAR-100*. These datasets are chosen for their varying levels of complexity and the ability to simulate non-IDD conditions effectively. Other frameworks employ the Federated Averaging (FedAvg) algorithm, not performing well with non-IID data, and other counterparts do not pay attention to aggregating dissimilar client updates that can lead to poor global modelling. Pfitzer et al., (2021) discuss the use of various datasets in the context of FL for medical applications, emphasizing the importance of handling non-IID, unbalanced, and vertically split data. They highlight the challenges of training models across different distributions, which is common in healthcare due to varying patient demographics across hospitals. Key datasets mentioned include *MNIST* for handwritten digits, *CIFAR-10* for object recognition, and keyword spotting datasets, which are used to illustrate the performance of FL algorithms under non-IID conditions. The authors detail how these datasets help in understanding the impacts of data distribution on FL model accuracy and training efficiency. Zhang, D. et al., (2021) used in his paper a dataset from the Radiological Society of North America (RSNA), includes 5,786 chest X-ray images primarily sourced for a Kaggle competition aimed at advancing medical image analysis for pneumonia detection. This dataset is valuable in federated learning environments, allowing for the development and testing of models across different institutions while maintaining data privacy. Its real-world application, especially in training models to recognize pneumonia from X-rays, highlights its relevance and popularity in healthcare-focused machine learning tasks.

# Real world Federated Learning scenarios

Kareem *et al.,* (2023)deployed a FL network with multiple institutions to participate in collaborative ML training without sharing patient data. The research uses a dataset that contains 5856 chest X-ray images, labelled as pneumonia 73% and non-pneumonia 23%. Several CNNs like AlexNet, DenseNet, RestNet-50, Inception and VGG19 are adapted to this federated setting. Each device trains a model locally, and model parameters are shared to a central server for aggregation. This study represents a practical application of FL in healthcare. A different application in the healthcare system (Lee and Shin, 2020) included modified MNIST, Medical Information Mart for Intensive Care-III (*MIMIC-III*), and electrocardiogram (ECG) dataset. The *MNIST* dataset was split across 10 clients. *MIMIC-III* and ECG were used to evaluate in-hospital mortality predictions and ECG classification. Implementation showed high performance across different setups, in more complex imbalanced and skewed scenarios, FL maintained a solid performance. These results are significant as they demonstrate FL’s applicability in real world clinical settings.

Going into the automotive industry Zhang *et al.,* (2021) authors study the integration of Machine Learning/Deep Learning (ML/DL) techniques in real-world systems, for autonomous driving. They focus on steering wheel angle prediction and introduce an end-to-end FL approach combined with an asynchronous model aggregation protocol. This method has significant improvements over common FL approaches that use synchronous protocols. This paper makes three primary contributions to the field of FL, first uses FL to validate an industrial case in autonomous driving, second, proposes a new method for processing real-time streaming data within the FL- paradigm and third it empirically evaluates its effectiveness over traditional and asynchronous FL methods in terms of accuracy and performance metrics.

In the telecommunication sector the application of FL in mobile networks (Kang et al., 2020) particularly focusing on the reliability of data updates from mobile devices (workers). In FL, mobile devices collaboratively train a global model without sharing their local data, which helps in preserving user privacy. However, the system is vulnerable to unreliable data updates, either due to intentional data poisoning or unintentional factors like low-quality data from energy constraints or mobility. To tackle these issues, the article introduces a reputation metric and a reliable worker selection scheme utilizing a consortium blockchain. This blockchain helps in managing the reputation of workers effectively, ensuring that only reliable updates contribute to the model training.

Finally in the field of cybersecurity (Alazab et al., 2021) the use of FL allows devices to collaboratively learn a model without centralizing data. It is especially useful in real-time applications, where cyber-attacks are common. Significant challenges in implementing FL for cybersecurity are identified, including inference, backdoor, adversarial, and free-riding attacks, with current solutions and potential future enhancements discussed. The paper also highlights real-world applications of FL in fields like finance and healthcare, illustrating its practical benefits and effectiveness in improving system performance while ensuring data privacy. The survey concludes by proposing future research directions, emphasizing the development of more secure, efficient, and transparent FL systems.

# Conclusion

There is a parallelism between the literature review and the objective definition; without the first, the second cannot be established, and vice versa. It unlocked many concepts inherent to FL that were not obvious initially.

To define objective one, section 4.2 determined that *TensorFlow Federated, Flower, Easy, IBM Federated and FLGo,* will be evaluated due to their extensive usage and popularity across the industry and the research fields.

For Objective Two, developing an FL system, all remaining sections have been instrumental, starting with Section 4.3, Distributed Machine Learning. Since this was FL at an early stage, within Hadoop different nodes served as clients and MapReduce aggregated the model; it could be a proof of concept before the actual data project. Section 4.4 showed visually how FL can be deployed, with virtual machines acting as clients and models being aggregated in cloud environments. It also highlighted that a key sector for FL is healthcare, and a popular dataset for its implementation is the Chest X-Ray Pneumonia dataset. This aligns with Section 4.5, where the Pneumonia dataset appears, alongside popular versions like MNIST and CIFAR.

Regarding objective three, section 4.6. showcased the importance of the healthcare system within the FL paradigm and another key industry which is the technology sector.

Finally, the literature review not only helped in defining the research objectives but also enriched the understanding of the subject matter. Sources are informative of different FL methods and approaches to undertake this project.

# **Proposed sampling strategy, population, sampling method and type**

The sampling method used in this proposal is non-probability, and the sampling type is judgmental. The sampling and method type are the same across all populations identified, which are:

Primary population for objective one: FL frameworks such as *TensorFlow Federated*, *Flower*, *EasyFL*, *IBM Federated*, and *FLGo.*

For objective two, developing a FL system, a secondary dataset population was identified, within the FL dataset scope the Chest X-ray Pneumonia and MNIST.

Lastly, the third objective has identified key industries within the FL paradigm, such as healthcare and technological sector.

Regarding the sampling method is necessary to address that bias is inherent to this method and to mitigate it, the following considerations were taken:

* The frameworks selected are widely used and accepted by the FL community, researchers, and private sector.
* The datasets, Chest X-ray and MINST are typically used in research papers and both are included in as part of *TensorFlow Federated datasets.*
* The technological and pharmaceutical sectors, the first created the concept of FL and the second is using it heavily to preserve patients’ privacy.

Given the reasons above chosen samples are representatives of their entire populations.

# **Proposed primary research methodology**

The primary research methodology for this proposal is experimentation. The intention is to determine the cause-and-effect relationship of model performance accuracy depending on the FL system used, whether traditional FL frameworks or the FL system developed in this project.

Experimentation is selected after research objectives were set. Initially, FL frameworks need to be evaluated; secondly a FL system is developed; and thirdly a comparative analysis between the two is conducted with the aim of measuring model performance accuracy. Model performance accuracy will be the independent variable, while the dependent variables will be the FL frameworks and the FL system. Finally experimentation, will measure the changes in the independent variable by using on of the two dependent variables.

# **Ethical and legal/regulatory aspects sourcing and analysing the data. Strategies. Addressing these aspects according to the European Commission Ethics and Data Protection handout.**

Starting with data sourcing, the datasets used to train the models are *MNIST*, which is publicly available in *TensorFlow* datasets (TensorFlow Datasets, 2010), and the pneumonia chest x-ray dataset, which is publicly available through the Radiological Society of North America (www.rsna.org, 2018). The *MNIST* dataset is licensed under *"CC BY 4.0 DEED Attribution 4.0 International"*, and the chest x-ray dataset has its own terms of use and attribution. Neither dataset imposes usage limits; however, certain clauses become applicable when there is an intention to modify or redistribute the datasets. This project does not intend to modify or redistribute the datasets.

Secondly, the legal or regulatory aspect of the data analysis does not address findings in pneumonia x-ray detection for the healthcare sector or number recognition for the telecommunications sector. The focus is solely on classifying and predicting based on dataset labels and comparing the performance of legacy Federated Learning (FL) frameworks with the FL project-developed system. It also examines whether any outcomes could benefit the aforementioned sectors. Given these considerations, the data analysis in this project does not cross any legal or regulatory boundaries.

According to the *Ethics and Data Protection handout* (HAYES and KUYUMDZHIEVA, 2021)there are five sections that apply to this research proposal:

1. Section II. Identifying and addressing ethics issues in your research proposal.

In this research proposal, there is a potentially vulnerable group: the pneumonia patients. However, the outcome of the experiment is solely to assess x-ray pneumonia detection. The intention is to measure model performance in image classification, and this dataset was chosen because it is widely used in FL scenarios.

1. Section III. Pseudonymisation and anonymisation.

One of the reasons for selecting these datasets is that none of them contain personal data, making it impossible to link to and identify any individuals.

1. Section IV. Data protection by design and default

The principle of data protection by design will be implemented throughout the entire project development, until the final results are released. Therefore, any inconclusive findings will not breach any ethical standards until they are verified upon release.

1. Section V. Informed consent to data processing.

The terms and conditions of the pneumonia x-ray detection dataset require users to inform the Radiological Society of North America of its use.

1. Section VII. Use of previously collected data.

Both datasets have been previously collected and are publicly available; consent is given to any researcher to conduct their research. Therefore, the data collected do not raise ethical concerns. As stated in the EU handout, Box 4, these datasets can be identified as open source.

# **Bibliography**

Alazab, M., R M, S.P., M, P., Reddy, P., Gadekallu, T.R. and Pham, Q.-V., 2021. Federated Learning for Cybersecurity: Concepts, Challenges and Future Directions. *IEEE Transactions on Industrial Informatics*, pp.1–1. doi: <https://doi.org/10.1109/tii.2021.3119038>.

Anthony, G., Gruzdev, A., Foley, P., Perepelkina, O., Sharma, M., Davidyuk, I., Trushkin, I., Radionov, M., Mokrov, A., Agapov, D., Martin, J., Edwards, B., Sheller, M., Pati, S., Narayana Moorthy, P., Wang, S.-H. and Bakas, S., n.d. *OpenFL: An open-source framework for Federated Learning*. Available at: https://arxiv.org/pdf/2105.06413.pdf. [Accessed Date].

Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P.P.B. and Lane, N.D., 2021. *Flower: A Friendly Federated Learning Research Framework.* Available at: https://arxiv.org/abs/2007.14390.

Chen, D., Gao, D., Kuang, W., Li, Y. and Ding, B. (2022). *pFL-Bench: A Comprehensive Benchmark for Personalized Federated Learning*. Available at :https://doi.org/10.48550/arXiv.2206.03655.

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D. and Yang, Q., 2019. *SecureBoost: A Lossless Federated Learning Framework*. Available at: https://doi.org/10.48550/arxiv.1901.08755

Dev, D. (2017). *Deep Learning with Hadoop*. Packt Publishing Ltd, pp. 49-55.

Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L. and Yang, Q., 2023. *FATE-LLM: A Industrial Grade Federated Learning Framework for Large Language Models.* Available at: https://doi.org/10.48550/arXiv.2310.10049

Galakatos, A., Crotty, A. and Kraska, T. (2017). Distributed Machine Learning. *Encyclopedia of Database Systems*, [online] pp.1–6. doi: https://doi.org/10.1007/978-1-4899-7993-3\_80647-1.

Guan, H., Yap, P.-T., Bozoki, A. & Liu, M., (2024). Federated learning for medical image analysis*: A survey. Pattern Recognition*, pp.110424–110424. Available at: <https://doi.org/10.1016/j.patcog.2024.110424>.

HAYES, B. and KUYUMDZHIEVA, A. (2021). *Ethics and data protection*. [online] Available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection\_he\_en.pdf.

Kang, J., Xiong, Z., Niyato, D., Zou, Y., Zhang, Y. & Guizani, M., (2020). Reliable Federated Learning for Mobile Networks*. IEEE Wireless Communications*, 27(2), pp.72–80. Available at: <https://doi.org/10.1109/mwc.001.1900119>.

Kareem, A., Liu, H. and Vladan Velisavljevic (2023). A federated learning framework for pneumonia image detection using distributed data. 4, pp.100204–100204. doi: https://doi.org/10.1016/j.health.2023.100204.

Kraska, T., Talwalkar, A., Duchi, J., Griffith, R., Franklin, M. and Jordan, M. (2015). MLbase: *A Distributed Machine-learning System*. Available at: <https://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper118.pdf>. [Accessed Date].

Lee, G. and Shin, S.-Y. (2020). Performance Assessment of Federated Learning on Clinical Benchmark Data (Preprint). *Journal of Medical Internet Research*. doi:https://doi.org/10.2196/20891.

Ghosh, A., Chung, J., Yin, D. and Ramchandran, K., (2022). An Efficient Framework for Clustered Federated Learning. *IEEE Transactions on Information Theory,* 68(12), pp.8076–8091. Available at: https://doi.org/10.1109/tit.2022.3192506. [Accessed Date].

Lai, Y.-H., Chen, S.-Y., Chou, W.-C., Hsu, H.-Y. and Chao, H.-C. (2024). Personalized Federated Learning with Adaptive Feature Extraction and Category Prediction in Non-IID Datasets. *Future internet*, 16(3), pp.95–95. Available at: https://doi.org/10.3390/fi16030095.

Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H. and Dou, D. (2022). From distributed machine learning to federated learning: a survey. *Knowledge and Information Systems,* 64(4), pp.885–917. doi: https://doi.org/10.1007/s10115-022-01664-x.

Ludwig, H., Baracaldo, N., Thomas, G., Zhou, Y., Anwar, A., Rajamoni, S., Ong, Y., Radhakrishnan, J., Verma, A., Sinn, M., Purcell, M., Rawat, A., Minh, T., Holohan, N., Chakraborty, S., Whitherspoon, S., Steuer, D., Wynter, L., Hassan, H. and Laguna, S., 2020*. IBM Federated Learning: an Enterprise Framework White Paper V0.1.* Available at: <https://doi.org/10.48550/arXiv.2007.10987>.

Luo, J., Wu, X., Luo, Y., Huang, A., Huang, Y., Liu, Y. and Yang, Q. (2021*). Real-World Image Datasets for Federated Learning*. Available at: https://doi.org/10.48550/arXiv.1910.11089.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Agüera y Arcas, B., (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, FL, USA. JMLR: W&CP volume 54.

Pfitzner, B., Steckhan, N. and Arnrich, B. (2021). Federated Learning in a Medical Context: A Systematic Literature Review. *ACM Transactions on Internet Technology*, [online] 21(2), pp.1–31. doi:https://doi.org/10.1145/3412357.

Salazar, T., Fernandes, M., Araujo, H. and Abreu, P.H. (2023). FAIR-FATE: *Fair Federated Learning with Momentum.* Available at: https://arxiv.org/abs/2209.13678 [Accessed 25 Apr. 2024].

Solanki, T., Rai, B.K. and Sharma, S., (2022). Federated Learning Using Tensor Flow. In: *EAI/Springer Innovations in Communication and Computing,* pp.157–167. Available at: https://doi.org/10.1007/978-3-030-85559-8\_10 [Accessed Date].

TensorFlow Datasets (2010). *mnist | TensorFlow Datasets*. [online] TensorFlow. Available at: https://www.tensorflow.org/datasets/catalog/mnist.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T. and Rellermeyer, J.S. (2020). A Survey on Distributed Machine Learning. *ACM Computing Surveys*, 53(2), pp.1–33. doi: https://doi.org/10.1145/3377454.

Wang, E., Kannan, A., Liang, Y., Chen, B. and Chowdhury, M., (2023). *FLINT: A Platform for Federated Learning Integration*. Available at: <https://doi.org/10.48550/arXiv.2302.12862>. [Accessed Date].

Wang, Z., Fan, X., Peng, Z., Li, X., Yang, Z., Feng, M., Yang, Z., Liu, X. and Wang, C. (2023*). FLGo: A Fully Customizable Federated Learning Platform*. Available at: <https://doi.org/10.48550/arXiv.2306.12079>.

www.rsna.org. (2018). *RSNA Pneumonia Detection Challenge (2018)*. [online] Available at: https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018.

www.youtube.com. (2024). Tutorial: *Federated Learning in AI4EOSC.* [online] Available at: https://www.youtube.com/watch?v=FrgVummLNbU [Accessed 10 May 2024].

www.youtube.com. (2022). *Federated Learning with Azure Machine Learning*. [online] Available at: https://www.youtube.com/watch?v=aTj4AqbCWEA [Accessed 10 May 2024].

Xing, E.P., Ho, Q., Dai, W., Kim, J.K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A. and Yu, Y. (2015). Petuum: A New Platform for Distributed Machine Learning on Big Data. *IEEE Transactions on Big Data*, 1(2), pp.49–67. doi: https://doi.org/10.1109/tbdata.2015.2472014.

Zhang, D., Ren, F., Li, Y., Na, L. and Ma, Y. (2021). Pneumonia Detection from Chest X-ray Images Based on Convolutional Neural Network. *Electronics*, 10(13), p.1512. doi: https://doi.org/10.3390/electronics10131512.

Zhang, H., Bosch, J. and Olsson, H., (2021). Real-time End-to-End Federated Learning: *An Automotive Case Study*. Available at: https://arxiv.org/pdf/2103.11879. [Accessed Date].

Zhuang, W., Gan, X., Wen, Y. and Zhang, S., (2022). EasyFL: A Low-code Federated Learning Platform For Dummies. *IEEE Internet of Things Journal*. Available at: <https://doi.org/10.1109/JIOT.2022.3143842>.

Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B., Bluemke, E., Nounahon, J.-M., Passerat-Palmbach, J., Prakash, K., Rose, N., Ryffel, T., Reza, Z.N. and Kaissis, G., (2021). PySyft: A Library for Easy Federated Learning*. Federated Learning Systems*, pp.111–139. Available at: https://doi.org/10.1007/978-3-030-70604-3\_5. [Accessed Date].