IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 May 2023, accepted 9 July 2023, date of publication 17 July 2023, date of current version 25 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296143

APPLIED RESEARCH

Implementing Privacy-Preserving and Collaborative Industrial Artificial Intelligence

RICARDO SILVA PERES, (Member, IEEE), ALEXANDRE MANTA-COSTA, AND JOSÉ BARATA, (Member, IEEE)
NOVA School of Science and Technology, NOVA University of Lisbon, 2829-516 Caparica, Portugal
UNINOVA-Centre of Technology and Systems, 2829-516 Caparica, Portugal
Associated Laboratory of Intelligent Systems (LASI), 2829-516 Caparica, Portugal

Corresponding author: Ricardo Silva Peres (ricardo.peres@uninova.pt)

This work was supported in part by the FCT/MCTES (UNINOVA-CTS) under Grant UIDB/00066/2020.

ABSTRACT Despite the growing connectivity and availability of sensor data boosted by the Industry 4.0 paradigm, data scarcity remains one of the biggest challenges for the widespread adoption of industrial AI, particularly regarding failure or defect data required for automated quality inspection solutions. Data sharing among different stakeholders is difficult due to data privacy, ownership, and cybersecurity concerns. Traditional centralized Machine Learning is already an essential part of several sectors, including smart manufacturing; however, it may lead to several issues regarding security and performance over time. In response to these problems, federated learning can be seen as a potential catalyst and enabler for the adoption of collaborative industrial artificial intelligence. It is an innovative and decentralized approach to machine learning, focused on collaboration and data privacy, shifting the information flow from the data to the models. Although considerable research efforts have been put forth towards this topic, there are still many unresolved issues concerning the application of federated learning approaches in smart manufacturing scenarios, including the lack of a common holistic reference framework and implementation. In this light, the present work aims to provide the foundation for a common framework for collaborative Industrial AI in smart manufacturing, addressing not only the collaboration aspect but also the main challenges of data privacy, ownership, and security. Additionally, we present and make publicly available a curated dataset focused on a manufacturing quality inspection scenario to further promote research on this topic.

INDEX TERMS Artificial intelligence, federated learning, Industry 4.0, manufacturing, privacy-preserving AI.

I. INTRODUCTION

As a consequence of the paradigm shift to Industry 4.0 and the subsequent push for digitalization and connectivity, the industrial world is now more interconnected than ever. Developments in fields such as the Internet of Things [1], Cyber-Physical System (CPS) [2] and Industrial Artificial Intelligence (AI) [3] elevated the industrial playing field to meet the requirements of flexibility, agility and reliability imposed by the current market demand.

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

However, despite the added level of connectivity and the subsequent growth in the availability of sensor data from machines, the volume of defect or failure-related data per asset remains relatively low [4]. This is only natural, as not only do manufacturers generally tend to optimise against the occurrence of defects, but also due to operational variation, assets may not present all degradation, failure or defect patterns during their lifetime. Hence, data availability represents a major challenge for AI solutions and their industrial adoption, particularly those based on deep learning, which requires very large volumes of data to be effective.

Some approaches based on generative models have been proposed [5], [6], [7] to address scenarios in the manufacturing domain for which data are scarce and difficult to obtain with adequate quality. However, even for such approaches a sufficiently large amount of initial data is required.

Another promising alternative to this is the promotion of collaborative Industrial AI approaches (*i.e.,* aggregating operational data from multiple stakeholders), if only the mechanisms to ensure data privacy, data ownership and cybersecurity issues are ensured, along with the potential for common benefit to be attained as the motivational factor.

The present work is focused on providing a foundation for this vision, with the main contributions being summarised as:

1) Contextualisation of Federated Learning (FL) in the scope of Industry 4.0 and collaborative Industrial AI;
2) Conceptual framework for collaborative FL in manufacturing;
3) Test case implementation using an automated quality inspection application;
4) Curated dataset made publicly available to promote research on the topic and reproducibility of the results.

The remainder of this paper is organised as follows: Section II provides a brief overview of key concepts and related work from current literature. Section III introduces the proposed conceptual framework for FL in a collaborative AI context. After this, Section IV describes the implementation for the example test case using an open-source framework. Afterwards, Section V presents and discusses the results, followed by Section VI where the conclusions are summarised, along with a brief discussion of future research directions.

II. RELATED WORK

In this section, a brief overview of the current literature is presented, addressing the main foundational concepts and approaches related to the proposed framework.

A. PRIVACY-PRESERVING ARTIFICIAL INTELLIGENCE

As we progress with the implementation Industry 4.0 and beyond, organizations become increasingly reliant on AI to drive innovation and improve efficiency by transforming data into business value. Simultaneously, there is also a growing number of connected machines generating data that is sensitive or proprietary and, therefore, cannot be shared with other organizations. This form of collaboration could represent a potential new business opportunity for manufacturers, providing an additional way to leverage the larger and larger volumes of data into a business advantage, so long as the privacy of their customers and the proprietary nature of their data are assured.

These topics have drawn the attention of several European governing bodies, with the European Commission setting up a High-Level expert group on AI, which has identified privacy as one of the seven key requirements for trustworthy AI [8].

From a technical standpoint, one possible way to achieve this is through the implementation of FL [9], [10]. It is an approach for privacy-preserving AI, as it enables machine learning models to be trained collaboratively on data distributed across multiple client devices, machines or even organizations, without requiring the data itself to be exchanged.

while FL makes considerable progress in terms of addressing important issues of data privacy, recent research has shown that there are still a number of risks present even in federated learning scenarios [11], [12], particularly with regard to reverse-engineering attacks that can directly extract sensitive information about the datasets from the model. As such, in spite of being one of the key components of privacy-preserving AI, FL should be employed in combination with other mechanisms that complement its shortcomings.

One such case is Differential Privacy (DP) [13], a method through which a certain amount of noise is added to the training data, in a way that not only makes it more difficult to extract information about individual data points in the dataset but also preserves the utility of the resulting model. Several recent studies can be found in the literature combining DP with FL as a formal privacy guarantee [14], [15], [16].

B. FEDERATED LEARNING IN INDUSTRIAL APPLICATIONS

Being a relatively recent technology, the FL domain is in ongoing development. Its application to industrial scenarios is still mostly conceptual and several challenges are yet to be mitigated. In an effort to create a FL solution to a smart industry scenario, Franco et al. [17] proposed a self-adaptive divided framework to cope with the existing industrial automation systems architecture as well as the Machine Learning (ML) procedures used in the industry. The focus of this work was the improvement of accuracy and efficiency of a FL framework. In order to achieve these improvements, the authors proposed modelling a multi-assignment optimisation problem and subsetting the datasets through the involved devices. Model accuracy is also the focus of the work of Liu et al. [18]. However, the effort to achieve the desired accuracy relies on a Gradient Compression Mechanism where local gradients are compressed to reduce the number of gradients exchanged between the clients and the central server.

In [19] the authors proposed a framework for an industrial scenario based on a combination of FL and Transfer Learning. This framework was focused on adapting the base industrial knowledge to each smart device's needs and applying this combination to locally specific data. The data division was made through every device, making each one a client of the FL. This approach leads to every device's specific adaptation of the global model suited to its particular needs.

While the aforementioned research efforts discuss FL frameworks applied to industrial scenarios, they are focused mostly on efficiency and adaptability to smart manufacturing. Data privacy problems are only addressed by using a FL system. In addition, with the information presented in the previous section, it is safe to assume that only the FL system is

not enough to protect data privacy it is also necessary further measures to safeguard each FL client's data.

With a focus on privacy preservation, Jiang et al. [20] proposed the use of membership proof in the FL system. Such membership proof is generated by cryptographic accumulators and is issued as a smart contract by the server on the blockchain. This technique ensures robustness to failures, verifiability of the clients and resistance against active adversaries.

Another similar approach can be found in the work of Hao et al. [21] which also uses encryption mechanisms to ensure privacy, while the communication between all the elements of the framework is encrypted. Despite the high level of encryption and consequently the high computational cost, the framework still achieves a decent accuracy in training on the MNIST dataset [22].

In Liu et al. [23], privacy and security of FL users were the primary concern. A blockchain-based FL framework was proposed, where attackers are recognised by the execution of smart contracts, defending the system against poisoning attacks. This study proved the trade-off between model performance and client protection, which although it ensures better protection against attacks, also under-performs compared to a simpler FL model.

The previous studies focus primarily on the security and privacy of FL models. Although an improvement in privacy is noticeable, a deprecation on accuracy is evident. In addition, all the above-mentioned frameworks create a FL system from scratch. However, applicability to real case scenarios is not a concern as data collection and preparation is not mentioned. In the testing stage of these frameworks, the datasets used were mainly MNIST [22] and CIFAR [24], which do not necessarily represent the full spectrum of challenges faced in an industrial scenario where data is complex and even may be untreated. Finally, beyond the literature, none of the frameworks were available for further improvement or testing, making it difficult to reproduce and implement.

III. COLLABORATIVE INDUSTRIAL AI FRAMEWORK

The main goal of the proposed FL framework is to facilitate collaboration among different stakeholders in Industrial AI applications. While doing so, it should still be robust and tailored to ensure the privacy of its stakeholders' data and modular to be easily adapted, extended and deployed to different industrial scenarios. Although not subjected to scalability testing, the framework's scope is fixed to a cross-silo scheme. According to the literature, the distribution in this scheme typically ranges from a few to tens of stakeholders [25], [26]. An overview of this framework is presented in Figure 1.

The upcoming subsections further explore each of the framework's nodes.

A. GLOBAL FEDERATION NODE (GFN)

In FL the server does not have direct access to the data from its clients. Instead, its role (and by extension the role

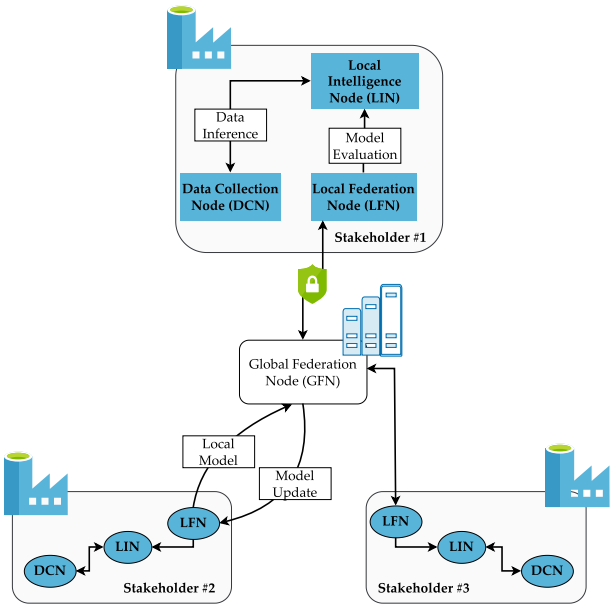


FIGURE 1. Conceptual framework for collaborative Industrial AI applications.

of the Global Federation Node (GFN)) is to manage the clients involved in the process, along with providing a secure mechanism by which the local model updates refined by the clients can be aggregated to form a consensus change based on a given strategy (e.g., federated averaging [27]). This secure mechanism includes encrypted communication points for the stakeholders to protect the shared model, which is then updated and propagated to the clients, after which the process can start anew.

This means that through this approach, the flow of information is shifted from moving data to the server to moving model parameters or gradients, providing a higher degree of privacy than traditional ML approaches. Furthermore, in the context of collaborative Industrial AI, such an approach provides the means to train models on data that would have been otherwise unavailable.

B. LOCAL FEDERATION NODE (LFN)

On the client's side, the Local Federation Node (LFN) is responsible for receiving updates from the global server, refining (training) the shared model using local data made available through the Data Collection Node (DCN) and sending said updates back to the server.

However, one issue arises when considering the more simplistic version of this interaction, as it requires the client to trust the server implicitly. In order to mitigate this problem, DP can be employed. With the usage of DP, each stakeholder can add an amount of noise to its local training, introducing plausible deniability to data and safeguarding each stakeholder from being associated with it. Also, this method

protects the participants from gradient/model inversion or attacks.

Despite the benefits of DP, a setback must be considered upon its usage. The introduction of noise to data negatively affects the performance of the model. As such, the trade-off between the level of noise introduced and the model performance must be accounted for.

C. LOCAL INTELLIGENCE NODE (LIN)

The Local Intelligence Node (LIN)'s main functionality is to operationalize the FL model. In other words, it receives the updated model from the LFN for deployment and uses it for inference with the run-time data fed by the Data Collection Node (DCN).

As such, its responsibilities are divided into two points of view. From a data point of view, the LIN is responsible for fusing data from multiple DCN sources, using the updated model for inference and safely storing the data in a secure database. On the other hand, from a model point of view, the LFN handles local model updates, evaluating their viability compared to the present model. Such a node can be implemented, for instance, by resorting to either agent-based or service-based approaches, as discussed later in Section IV.

D. DATA COLLECTION NODE (DCN)

Lastly, when looking through the lens of a CPS, the DCN can be seen as a key part of the digitalisation of a physical component or system, providing the means for run-time data to be collected and then used for training, evaluation and inference.

Multiple DCNs can be instantiated as needed, with data fusion being carried out either at the DCN level (*e.g.*, for a subsystem abstracted by a DCN, comprising multiple components which are in turn also abstracted by a DCN) and/or at the LIN level.

IV. QUALITY INSPECTION TEST CASE

The test case represents an adaptation of the scenario presented in [28], based on a real industrial test case (see Figure 2), which consists of images captured by a structural adhesive application station. An automated quality inspection system should detect two different defects in the adhesive cord during the process: discontinuity and excess. With prior knowledge of these defects, two classes of images were labelled as depicted in Figure 3. This scenario was chosen to establish a research guideline, where three stakeholders were simulated with data deficiency. Based on the experiments carried out in [29], this type of approach shows promise regarding the generalization beyond the bead shape included in the training set.

The used dataset comprised 372 images, 207 images of the *Discontinuity* class, and 165 belonging to the *Excess* class. This dataset is publicly available on Github.[1]

[1]https://github.com/AlexCosta157/CISP_FL

FIGURE 2. View of the industrial application that was simulated for the test case (adapted from [28]).

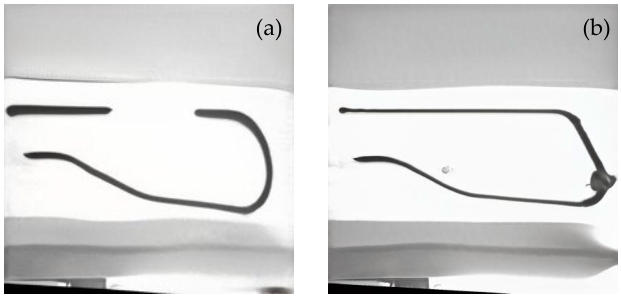


FIGURE 3. The two classes of images contained in the test case dataset, Discontinuity (a) and Excess (b).

This test case aims to serve the built framework with a real scenario to demonstrate the impact a multi-stakeholder collaborative approach based on FL can have compared to the performance (in terms of accuracy) of local, single-stakeholder solutions.

To this end, the dataset was split into three subsets attributed to each client considered in this case study. Two subsets were artificially manipulated to predominate one of the defects, and a third subset was built with minimal data. This division, illustrated in Figure 4, emulates a scenario where each manufacturer has limited access to data for a specific type of defect and an abundance of data from the other or limited data for both defects. Thus, it highlights how each manufacturer can benefit from a collaborative approach. A validation subset of 116 images (49 belonging to the excess class and 67 of discontinuity class) was used in all the model evaluations. A validation dataset should be attributed to each stakeholder in a real scenario. However, the same validation dataset was used to evaluate all stakeholder's data to have a common evaluation baseline.

In this test case, the data collection process was abstracted by a file containing the respective images in the dataset. As for the LIN, its implementation could pass for using an agent-based approach as exemplified in [30] or using a service-oriented approach as discussed in [31]. A service-oriented

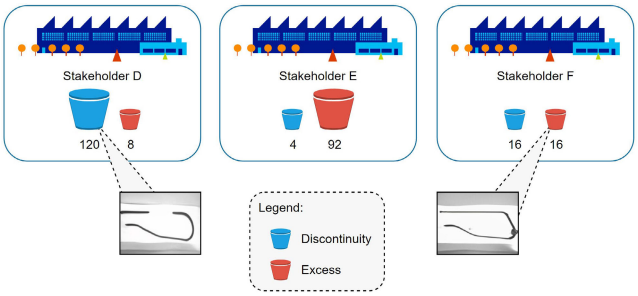


FIGURE 4. Distribution of the defect samples amongst the different stakeholders for the federated scenario.

approach was chosen, where a Flask API[2] was implemented in the LIN. The connected DCNs could request an image classification when one of the collected images was passed, being a classification provided as a response and the image and respective label saved in a secured database. On the other hand, the corresponding LFN could request a model update after a FL process, passing a newly updated model, which would be evaluated and updated if its performance was confirmed to be better than the currently implemented model.

For the implementation of the FL system, several general-purpose frameworks were taken into account. Among them, Flower [32] stands out for being scalable, agnostic to the underlying ML framework and easy to deploy. This was used to implement the GFN as a central server for client management and model aggregation, along with three LFNs responsible for the local refinement and evaluation rounds for each of the three clients using the locally held datasets. The LFN was implemented by extending Flower's client abstraction, while the GFN extends the server class using a gRPC transport layer, with both the LFNs and GFN being deployed on the same machine as this was sufficient for the scope of the test case. Nevertheless, the same implementation could be used for a distributed approach. In addition to the LFN's functionalities, the possibility to use differential privacy in the local rounds of the FL training was also implemented. The LFN configuration allowed stakeholders to customize their privacy in training their model. The model training process for this test case consisted in using the same baseline model architecture, MobileNetV2 [33] for all individual clients, with three implementations of the aggregation algorithm, federated averaging based on [27], federated optimization and federated optimization the adaptive optimizer YOGI [34]. Furthermore, the centralized and individual models were trained for 200 epochs with a batch size of 32. In contrast, the federated models were subjected to various combinations between local and global rounds, with a categorical cross-entropy loss function and an Stochastic Gradient Descent (SGD) optimizer applied to all models. The best-performing local and global round combinations for each algorithm were chosen for the differentially private training, and several noise levels at different learning rates were applied to them. Also,

[2]https://flask.palletsprojects.com/en/2.2.x/

TABLE 1. Test case training configurations.

Model	Configuration
Client 1	batch size: 32
Client 2	learning rate: 0.0001
Client 3	input shape: (224,224, 3)
	epochs: 200
FedAvg	batch size: 32
FedOpt	learning rate: 0.0001
FedYOGI	input shape: (224,224, 3)
	Local rounds: 5, 10, 20
	Global rounds: 50, 20, 10

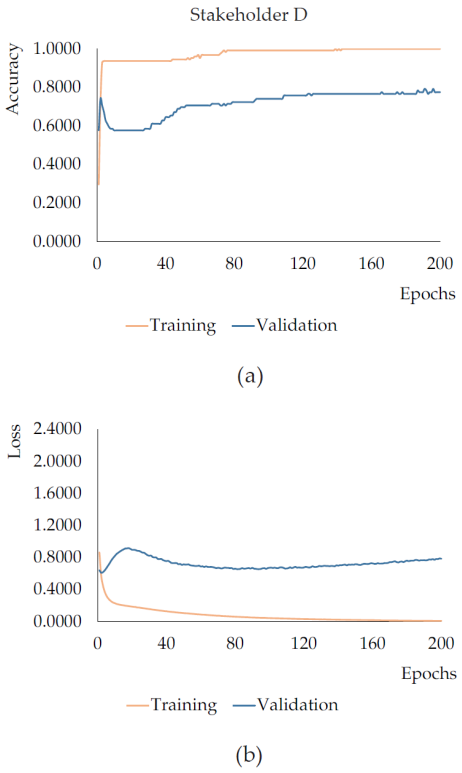


FIGURE 5. Comparison between the training accuracy (a) and loss (b) with the respective validation of stakeholder D.

the optimizer was changed to DP-SGD, allowing the addition of DP. Training configurations are summarised in Table 1, with missing parameters taking default values. The FL system parameters were chosen based on the recommendations in [35]. The results will be discussed in the next section.

V. DISCUSSION OF RESULTS

To analyze the results, it is interesting to start by observing the performance of the isolated clients emulating the case where one manufacturer has limited access to data on a specific type of defect. An example taken from Stakeholder D is shown in Figure 5. Here it can be seen that despite the large number of samples and epochs, the performance of Stakeholder D's model presents a significant difference compared to the validation results. This discrepancy can be most likely attributed to a lack of examples of the excess defect in the training

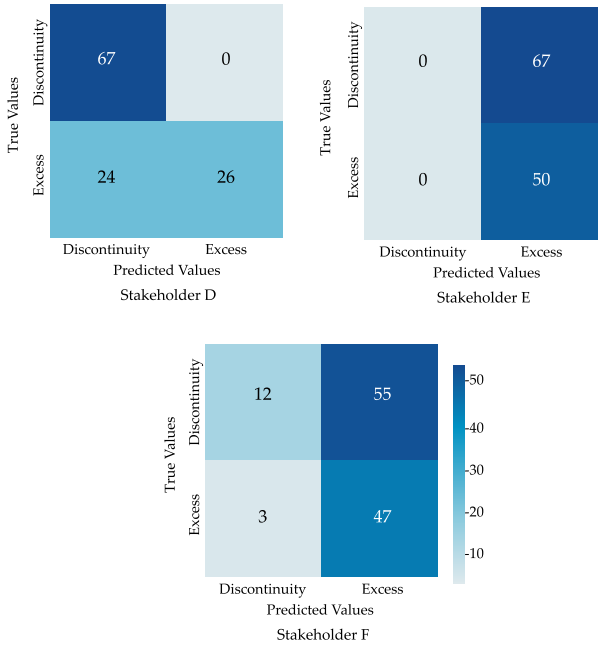


FIGURE 6. Confusion matrices of the local models of each stakeholder when tested with the validation data.

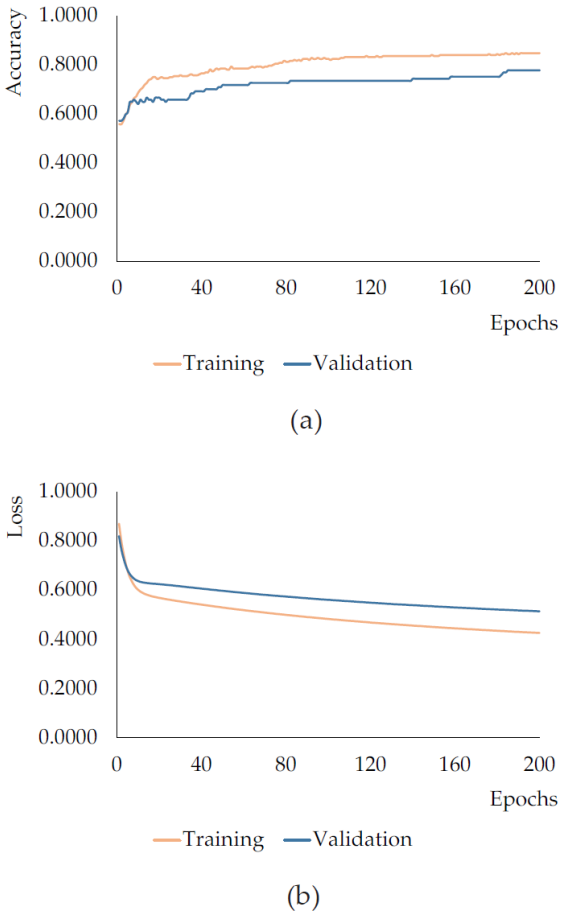


FIGURE 7. Centralized training accuracy (a) and loss (b) graphics.

data (which only contains eight samples of this defect in Stakeholder D's case). The effects of the shortage of data can also be identified in the other stakeholders, as Figure 6 depicts.

Another noteworthy performance to analyze is the model training with all data centralized. Although this approach in a real-world scenario would be challenging, it provides a baseline comparison for using FL. In Figure 7, it is possible to verify that the centralized training presents a similar accuracy to stakeholder D. However, the validation loss does not increase, contrary to the shown local model, which supports the data shortage assumption.

The visible under-performance of local training due to a lack of data is the main selling point for which a collaborative approach can provide considerable value, as it is possible to verify in the results from the FL experiments shown in Table 2.

As can be observed, the FL training was performed with three algorithms commonly adopted in the literature: Federated Averaging (Federated Averaging (FedAvg)), Federated Optimization (Federated Optimisation (FedOpt)) and Federated Optimization with the adaptive optimizer YOGI (Federated Optimization with the adaptive optimizer YOGI (FedYOGI)). In all the tested algorithms, several combinations of local and global rounds were made in order to verify the best combination. The best accuracy results of each aggregation algorithm are highlighted in Table 2, showing that in terms of accuracy, the FedOpt and the FedAvg algorithms achieved the same final accuracy topping the accuracy achieved by the FedYOGI algorithm.

However, it is important to evaluate further the best-performing models of the three algorithms to assess the best FL result. In Table 3, it is possible to verify a comparison between these models regarding the accuracy, loss, AUC and F1 Score.

From the comparison in Table 3, it is possible to verify that the best overall performance belongs to the FedOpt algorithm trained with 50 global and 20 local rounds. This model, along with one of the best accuracies achieved, provided the smaller value of loss, the best AUC and one of the top F1 scores.

When comparing the previously highlighted FL model with the centralized model (Table 4), it can be seen that the FedOpt model attained an improvement of 0.0239 accuracy (2.39%), 0.0502 AUC and 0.0239 in F1 score, with a marginal decrease of -0.552 in loss. From these results, it is possible to conclude that the federated model was fairly superior to the centralized one. Furthermore, as discussed in [35], this improvement could be limited by the scarce number of clients involved in the FL system, as this factor was shown to drastically affect the accuracy of most FL algorithms under test, with a higher number of clients leading to more significant improvements in accuracy. As such, further experimentation

TABLE 2. Accuracy values resulting from the FL trains with different combinations of local and global rounds.

		Algorithm								
		FedAvg			FedOpt			FedYogi		
		Global Rounds								
		10	25	50	10	25	50	10	25	50
Local Rounds	5	0.4655	0.5603	**0.8017**	0.5948	0.6034	0.7155	0.7069	**0.7845**	0.7759
	10	0.5776	0.6293	0.7672	0.5776	0.6896	0.7155	0.7240	0.7759	0.7759
	20	0.7759	0.7845	0.7759	0.6466	0.6810	**0.8017**	**0.7845**	0.7672	0.7672

TABLE 3. Evaluation metrics of the models with the best accuracy from each aggregation algorithm.

Algorithm		Accuracy	Loss	AUC	F1 Score
FedAvg	50 global rounds 5 local rounds	0.8017	0.5354	0.8345	0.8017
FedOpt	50 global rounds 20 local rounds	0.8017	0.4585	0.8801	0.8017
FedYogi	25 global rounds 5 local rounds	0.7845	0.4865	0.8496	0.7845
FedYogi	10 global rounds 20 local rounds	0.7845	0.4880	0.8425	0.7845

TABLE 4. Comparison between the FL train with the best performance and the centralized training.

Model	Loss	Accuracy	AUC	F1 score
Centralized	0.5137	0.7778	0.8299	0.7778
FedOpt	0.4585	0.8017	0.8801	0.8017

and assessment of the FL system parameters could further increase the benefit of this approach.

Besides the FL testing, the framework's general implementation should also be tested to ensure it fulfils the necessary performance requirements. The guidelines from the IEEE Std 3652.1-2020 [36] standard were followed for these tests, as it establishes the testing criteria that FL frameworks should perform.

A. COMPUTATION EFFICIENCY

The first criterion that should be tested is computation efficiency, divided into two categories: time consumption and memory consumption. Regarding time consumption, the framework was tested regarding the time each stakeholder took to perform each local training and the global model evaluation. These values were averaged and normalized according to Equation 1. The normalization was necessary due to the time consumption being proportional to the dataset size of each stakeholder.

$$T_{te\_norm} = \frac{T_{te}}{Num_{te}} \tag{1}$$

Table 5 depicts each stakeholder's normalized average training time for each number of performed local rounds. It is possible to conclude that the presented results are unexpected since the stakeholders with higher data sizes took less time than the ones with fewer data. This might be because all three stakeholders were trained in the same machine without dedicated hardware or batch size standardization. In particular,

TABLE 5. Normalized training time for each stakeholder.

Local Rounds	Normalized time		
	Stk D	Stk E	Stk F
5	0.1088	0.1493	0.3711
10	0.1974	0.2767	0.6950
20	0.3846	0.5265	1.3235

TABLE 6. Normalized average evaluation time for each stakeholder.

Stakeholder	Normalized average evaluation time
D	0.0083
E	0.0177
F	0.0136

a standardization for all stakeholders might not be suitable for the dataset size of stakeholder F since its batch size and dataset size are the same.

As the evaluation time is concerned, Table 6 shows the evaluation times normalized by the validation data. As mentioned before, the validation dataset used to evaluate the global model was the same for all stakeholders to provide a uniform global model evaluation. It is possible to conclude that the usage of the same dataset for evaluation in all three stakeholders was the cause of the similar evaluation time results.

In the category of memory consumption, the evaluation of the memory usage of the proposed framework was divided into two subsets, the memory used by the training and evaluation data, and the memory used by the framework's code. Each stakeholder's training dataset and the validation dataset were contemplated to evaluate the memory of the data. The latter was included in all three stakeholders as they used it in the evaluation rounds.

As for the training and evaluation data, stakeholder D had a dataset with 1.98 Mb, stakeholder E had 1.97 Mb, and stakeholder F's dataset had 1.3 Mb.

Table 7 depicts the consumed memory by each stakeholder and the server. The virtual environment memory consumed was common to all the entities as the used virtual environment possessed the same dependencies and libraries. Also, it is essential to note that the virtual environment was common to all the stakeholders as they were instanced in the same machine. The choice to include the virtual environment memory in three stakeholders was made to simulate a real scenario where all three stakeholders must include the same dependencies in their respective machines.

TABLE 7. Auxiliary memory consumed by each entity, measured in Mb.

Memory usage	Entity			
	Stakeholder D	Stakeholder E	Stakeholder F	Server
Code	26.73	26.73	26.73	0.015
Virtual environment	5457.92	5457.92	5457.92	5457.92
Total	5484.65	5484.65	5484.65	5457.93

B. PRIVACY AND SECURITY

Another necessary criterion to evaluate the framework is the privacy and security it provides to its stakeholders.

Regarding privacy, FL is not enough to ensure full privacy as it is vulnerable to, for instance, reverse-engineering attacks that can extract sensitive information about the data directly from the model. In order to enhance the stakeholder's privacy, differential privacy was added to their models. However, adding noise to data presents a trade-off between privacy and training accuracy. The FedAvg, with fifty global rounds and five local rounds and FedOpt, with fifty global and twenty local rounds, were chosen to test the model's accuracy depending on the amount of noise to be added and the variation of the learning rate. These algorithms were chosen for being the ones that provided the best accuracy in the model evaluation.

In Table 8, the influence of the noise addition on the model accuracy becomes clear. The noise levels tested in both algorithms were chosen empirically from several tests. The choice of the four noise values tested relied on the capability to improve the privacy of each local training, and as the results show, an evident decrease in accuracy is verified as the amount of noise is increased. It is also possible to verify the influence of changing the learning rate in each algorithm for the different noise values. In the FedAvg algorithm, the increased learning rate showed improvements in the model accuracy in constant noise values. On the other hand, in the FedOpt algorithm increasing the learning rate proved, in general, to worsen the accuracies when the noise value was constant.

The privacy budget spent on the differentially private training (ε) is the quantifier of DP, where the smaller the ε, the greater the privacy of the algorithm [37]. Figure 9 compares the two chosen algorithms and the values of added noise. Naturally, the value of ε decreases with the increased addition of noise. In other words, the training guarantees more privacy for the stakeholder. However, both algorithms have a visible difference in the ε values. As the aggregation algorithm does not influence the privacy budget, this difference is attributed to the number of local epochs applied to each study case. The FedOpt algorithm used twenty local epochs, and the FedAvg algorithm only used five, which provoked the evident discrepancy in the epsilon value computation. It is also important to denote that the learning rate does not influence the ε, so it was not considered for this evaluation.

Differential privacy introduces an arbitrary quantity of noise to the DL training process, which translates into an additional layer of privacy for the data used in each local training. As it was verified, the addition of noise also presents a setback in the model's performance, lowering its accuracy and increasing its loss. The existing trade-off between model performance and privacy depends on several factors, from which two were highlighted, the noise volume and the learning rate of the model. Although the learning rate does not directly influence the differential privacy process, it influences the training process and, consequently, the final model's performance. It was later verified that the model with the best accuracy/privacy ratio (FedAvg with noise value of 5 and learning rate of 0.15) was not the most viable for usage. From this verification, it can be inferred that a thorough validation is needed for differential privacy as it may be deeply involved with the model performance. In the presented scenario, if the noise value was lowered to 2, the results were proven closer to the FL model in the same conditions without DP (Figure 9).

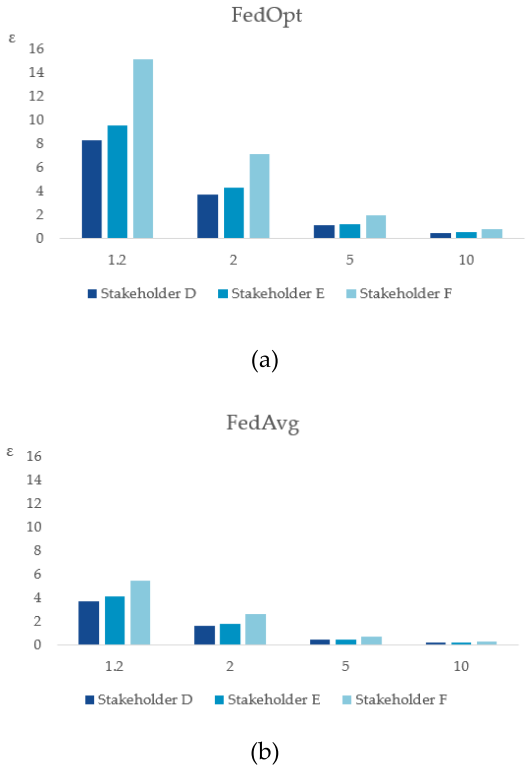


FIGURE 8. Resulting ε from the differentially private trains with different noise levels.

TABLE 8. Resulting accuracy of the differentially private FL training with the FedAvg and the FedOpt algorithms, when applied different learning rates and levels of noise.

		Algorithm							
		FedAvg				FedOpt			
Noise Level		1.2	2	5	10	1.2	2	5	10
Learning rate	0.0015	0.7759	0.5776	0.5776	0.5776	0.7845	0.7931	0.7672	0.7500
	0.1500	0.8017	0.7500	0.7500	0.6034	0.8017	0.7414	0.7069	0.6983

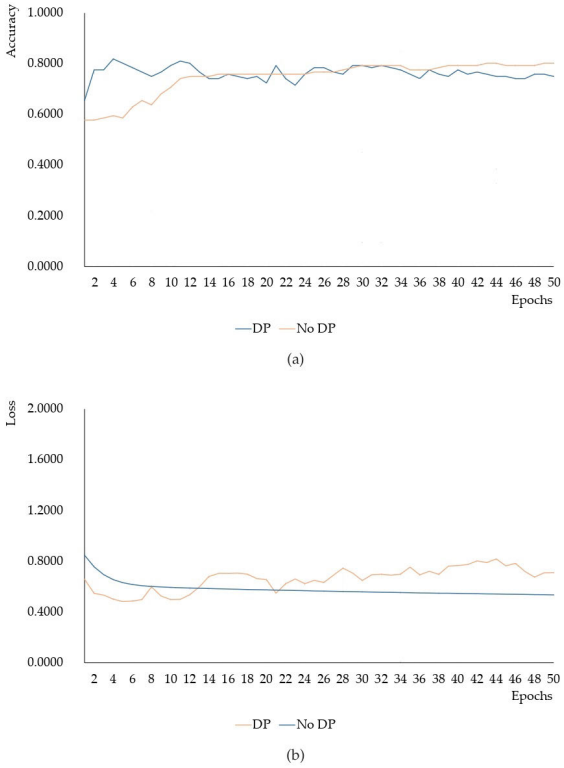


FIGURE 9. Comparison of the training accuracy (a) and loss (b) of the FedAvg algorithm with fifty global epochs, five local epochs with and without the usage of DP with a noise value of 2.

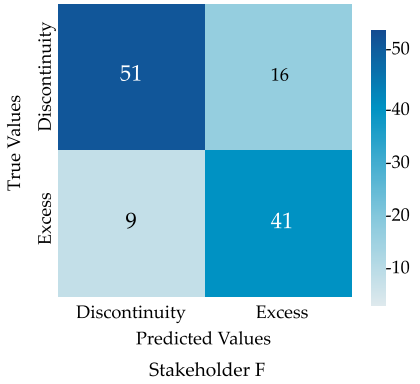


FIGURE 10. Confusion matrix of the model trained with the FedAvg algorithm with fifty global epochs, five local epochs with and without the usage of DP with a noise value of 2, when tested with the validation data.

It can be concluded that the usage of DP in FL brings an additional privacy layer to the framework with a performance cost. However, according to Figure 10, it is possible to verify that the results obtained from the FL process with DP showed an improvement in the label prediction than the local models, demonstrating a clear advantage of the collaborative approach.

As the framework is agnostic to the study case implemented, an absolute noise value cannot be recommended as it depends more on the ML scenario than on the framework configuration. However, the framework successfully supported the usage of DP in its trains. The ability to configure the training and DP parameters presents an advantage in protecting the privacy of the framework's clients.

The security testing was done considering the necessity of the framework to defend against attacks on each stakeholder's database and data transmission. For the database defence, it was decided for each stakeholder to encrypt the data before saving it in the database, being only decrypted when used for local training. As for the defence against channel monitoring, the SSL cryptographic protocol was chosen for encrypting the transmitted data between the stakeholders and the GFN.

Upon testing, the framework's security was demonstrated to be effective against the attacks, namely man-in-the-middle and read-write attacks. Although the security breaches in the whole system were assumed beforehand to evaluate a more profound attack, the framework proved to have implemented a security system that allows the data to remain secure within the framework's processes.

VI. CONCLUSION, LIMITATIONS AND FUTURE WORK

This paper addressed the topic of collaborative Industrial AI, with a particular emphasis on an automated quality inspection test case. A conceptual framework for FL in smart manufacturing was presented, addressing the main challenges of data privacy, ownership and cybersecurity.

Additionally, an initial implementation of the FL system was described as a reference example, focusing specifically on an automated quality inspection scenario. The results of the study were thoroughly discussed, including considerations about the potential impact of the federated solution and its parameters.

Based on this, the main contributions of this paper can be summarised as: i) the contextualisation of FL in the scope of collaborative Industrial AI; ii) a conceptual framework for FL in smart manufacturing; iii) a test case example implementation based on an automated quality inspection application; iv) a curated dataset of structural adhesive images with two

types of defects, made publicly available to promote research on this topic.

Even though FL represents a large step-up in terms of addressing critical issues of data privacy and security when compared to more traditional ML approaches, by itself it is still not sufficient to ensure full privacy as it has been shown to be vulnerable to for instance reverse-engineering attacks that can extract sensitive information about the data directly from the model.

The addition of differential privacy to the local training was also considered fruitful, enabling the data to remain confidential when training the models. However, this privacy technique also proved to affect the model performance negatively. With this factor in consideration, the framework allowed the stakeholders to configure the DP parameters and get the best performance/privacy ratio. The ability of the stakeholder to use (or not) DP in its training process gives the framework's clients the choice to increase their privacy with the trade-off of decreased accuracy.

The proposed work successfully implemented a framework which improved a manufacturing process using privacy-preserving federated learning. The datasets were kept in each stakeholder without information sharing among them.

Still, despite the promising results of the initial prototype, further work is required to improve the framework's robustness and fairness to production levels for all stakeholders. For instance, while the modularity of the proposed framework allows for an escalation of the number of stakeholders and nodes attached to them, more thorough scalability testing with a larger quantity of stakeholders is required. Also, homomorphic encryption could be applied to enable the computation of data in an encrypted state, boosting the security and privacy of the overall approach.

Lastly, a limitation of this initial implementation of the framework is that the adopted ML model only performs binary classification to distinguish between types of defects. This means that to be used in production, an upstream system capable of separating OK parts from defective ones is necessary. The model itself was trained with a limited amount of data containing only images of defective parts of each class, which was also a constraint of the use case itself. Nevertheless, both of these limitations are not related to the main focus of the present work, which is showcasing the capacity of the FL framework to enable a collaborative improvement of industrial AI processes in a limited-data context across stakeholders in a privacy-preserving manner. They are instead related solely to the ML model implementation and can be easily solved by training a different model (e.g., object detection or multi-class classifier for OK, Discontinuity, and Excess) and by increasing the dataset size/variety, given that the framework is agnostic to the type of ML task being employed.

REFERENCES

[1] L. D. Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[2] L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, and K. Ueda, "Cyber-physical systems in manufacturing," *CIRP Ann.*, vol. 65, no. 2, pp. 621–641, 2016.

[3] R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, "Industrial artificial intelligence in industry 4.0–Systematic review, challenges and outlook," *IEEE Access*, vol. 8, pp. 220121–220139, 2020.

[4] B. Bagheri, M. Rezapoor, and J. Lee, "A unified data security framework for federated prognostics and health management in smart manufacturing," *Manuf. Lett.*, vol. 24, pp. 136–139, Apr. 2020.

[5] F. Zhou, S. Yang, H. Fujita, D. Chen, and C. Wen, "Deep learning fault diagnosis method based on global optimization GAN for unbalanced data," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104837.

[6] W. Mao, Y. Liu, L. Ding, and Y. Li, "Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study," *IEEE Access*, vol. 7, pp. 9515–9530, 2019.

[7] D. Cabrera, F. Sancho, J. Long, R.-V. Sánchez, S. Zhang, M. Cerrada, and C. Li, "Generative adversarial networks selection approach for extremely imbalanced fault diagnosis of reciprocating machinery," *IEEE Access*, vol. 7, pp. 70643–70653, 2019.

[8] L. Floridi, "Establishing the rules for building trustworthy AI," *Nature Mach. Intell.*, vol. 1, no. 6, pp. 261–262, May 2019.

[9] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1175–1191.

[10] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proc. 29th Conf. Neural Inf. Process. Syst. (NIPS)*, Barcelona, Spain, 2016, pp. 1–10.

[11] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 911–926, 2020.

[12] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," 2018, *arXiv:1811.04017*.

[13] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 746–789, 1st Quart., 2020.

[14] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated learning with local differential privacy," in *Proc. 3rd ACM Int. Workshop Edge Syst., Analytics Netw.*, 2020, pp. 61–66.

[15] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam, "Local differential privacy-based federated learning for Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, Jun. 2021.

[16] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2587–2596.

[17] N. Franco, H. M. Van, M. Dreiser, and G. Weiss, "Towards a self-adaptive architecture for federated learning of industrial automation systems," in *Proc. Int. Symp. Softw. Eng. Adapt. Self-Managing Syst. (SEAMS)*, May 2021, pp. 210–216.

[18] Y. Liu, N. Kumar, Z. Xiong, W. Y. B. Lim, J. Kang, and D. Niyato, "Communication-efficient federated learning for anomaly detection in industrial Internet of Things," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[19] K. I. Wang, X. Zhou, W. Liang, Z. Yan, and J. She, "Federated transfer learning based cross-domain prediction for smart manufacturing," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4088–4096, Jun. 2022.

[20] C. Jiang, C. Xu, and Y. Zhang, "PFLM: Privacy-preserving federated learning with membership proof," *Inf. Sci.*, vol. 576, pp. 288–311, Oct. 2021.

[21] M. Hao, H. Li, X. Luo, G. Xu, H. Yang, and S. Liu, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, Oct. 2020.

[22] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[23] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato, and A. A. A. El-Latif, "A secure federated learning framework for 5G networks," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 24–31, Aug. 2020.

[24] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2009.

[25] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X20329848

[26] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, Jun. 2021.

[27] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[28] R. S. Peres, M. Azevedo, S. O. Araújo, M. Guedes, F. Miranda, and J. Barata, "Generative adversarial networks for data augmentation in structural adhesive inspection," *Appl. Sci.*, vol. 11, no. 7, p. 3086, Mar. 2021.

[29] R. S. Peres, M. Guedes, F. Miranda, and J. Barata, "Simulation-based data augmentation for the quality inspection of structural adhesive with deep learning," *IEEE Access*, vol. 9, pp. 76532–76541, 2021.

[30] R. S. Peres, J. Barata, P. Leitao, and G. Garcia, "Multistage quality control using machine learning in the automotive industry," *IEEE Access*, vol. 7, pp. 79908–79916, 2019.

[31] C. Briese, M. Schlüter, J. Lehr, K. Maurer, and J. Krüger, "Towards deep learning in industrial applications taking advantage of service-oriented architectures," *Proc. Manuf.*, vol. 43, pp. 503–510, Jan. 2020.

[32] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. Hei Li, T. Parcollet, P. P. B. de Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," 2020, *arXiv:2007.14390*.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[34] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, "Adaptive methods for nonconvex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

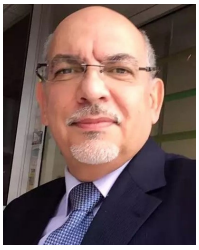
[35] A. Mulay, B. Gaspard, R. Naidu, S. Gonzalez-Toral, S. Vineeth, T. Semwal, and A. M. Agrawal, "FedPerf: A practitioners' guide to performance of federated learning algorithms," in *Proc. NeurIPS Workshop Pre-Registration Mach. Learn.*, 2021, pp. 302–324.

[36] *IEEE Guide for Architectural Framework and Application of Federated Machine Learning*, IEEE Standard 3652.1-2020, 2021, pp. 1–69.

[37] Z. Liu, J. Guo, W. Yang, J. Fan, K.-Y. Lam, and J. Zhao, "Privacy-preserving aggregation in federated learning: A survey," *IEEE Trans. Big Data*, early access, Jul. 15, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9830997/, doi: 10.1109/TBDATA.2022.3190835.

RICARDO SILVA PERES (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the NOVA University of Lisbon, Portugal, in 2015 and 2019, respectively, specializing in the application of AI in smart manufacturing. Since 2014, he has been a Researcher with the UNINOVA-Centre of Technology and Systems, working at the intersection of AI, robotics, and data science. He is currently an Invited Professor with the Department of Electrical Engineering, School of Sciences and Technology, NOVA University of Lisbon. He has participated in several national and international research projects, including FP7 PRIME, H2020 PERFoRM, H2020 OpenMOS, H2020 GO0D MAN, H2020 AVANGARD, and H2020 BIECO. He is the author of several publications in high-ranking international scientific journals and conference proceedings (peer-reviewed). His research interests include industrial AI, cyber-physical systems, and multi-agent systems. He has been a member of the IEEE IES Technical Committee on Industrial Agents, since 2018, and the IEEE Standards Association P2805.1/2/3 Edge Computing Nodes Working Group, since 2019.

ALEXANDRE MANTA-COSTA received the B.S. and M.Sc. degrees in electrical and computer engineering from the NOVA University of Lisbon, Portugal, in 2019 and 2022, respectively. Since 2021, he has been a Teaching Assistant with the NOVA School of Science and Technology. He is currently a Researcher with the UNINOVA-Centre of Technology and Systems, with a focus on the development of intelligent and predictive manufacturing systems. His research interests include artificial intelligence, deep learning, the IoT, and additive manufacturing.



JOSÉ BARATA (Member, IEEE) received the Ph.D. degree in robotics and integrated manufacturing from the Nova University of Lisbon. He is currently a Professor with the Department of Electrical Engineering, Nova University of Lisbon, and a Senior Researcher with the UNINOVA Institute. He has participated in more than 15 international research projects involving different programs (NMP, IST, ITEA, ESPRIT). Since 2004, he has been leading the UNINOVA participation in EU projects, namely EUPASS, Self-Learning, IDEAS, PRIME, RIVERWATCH, ROBO-PARTNER, and PROSECO. In the last years, he has participated actively in researching SOA-based approaches for the implementation of intelligent manufacturing devices (e.g., within the Inlife Project). He has published over 100 original papers in international journals and international conferences. His main research interests include intelligent manufacturing with a particular focus on complex adaptive systems, involving intelligent manufacturing devices. He is a member of the IEEE Technical Committee on Industrial Agents (IES), the Self-Organization and Cybernetics for Informatics (SMC), and the Education in Engineering and Industrial Technologies (IES). He is also a member of the IFAC Technical Committee 4.4 (Cost-Oriented Automation).