Post-Processing Fairness Evaluation of Federated Models: An Unsupervised Approach in Healthcare

Ilias Siniosoglou[†], Vasileios Argyriou[‡], Panagiotis Sarigiannidis[†], Thomas Lagkas[§], Antonios Sarigiannidis[¶], Sotirios K. Goudos[‖] and Shaohua Wan[**]

***Abstract*—Modern Healthcare cyberphysical systems have begun to rely more and more on distributed AI leveraging the power of Federated Learning (FL). Its ability to train Machine Learning (ML) and Deep Learning (DL) models for the wide variety of medical fields, while at the same time fortifying the privacy of the sensitive information that are present in the medical sector, makes the FL technology a necessary tool in modern health and medical systems. Unfortunately, due to the polymorphy of distributed data and the shortcomings of distributed learning, the local training of Federated models sometimes proves inadequate and thus negatively imposes the federated learning optimization process and in extend in the subsequent performance of the rest Federated models. Badly trained models can cause dire implications in the healthcare field due to their critical nature. This work strives to solve this problem by applying a post-processing pipeline to models used by FL. In particular, the proposed work ranks the model by finding how fair they are by discovering and inspecting micro-Manifolds that cluster each neural model's latent knowledge. The produced work applies a completely unsupervised both model and data agnostic methodology that can be leveraged for general model fairness discovery. The proposed methodology is tested against a variety of benchmark DL architectures and in the FL environment, showing an average 8.75% increase in Federated model accuracy in comparison with similar work.**

***Index Terms*—Fairness, Adversarial networks, Federated, Image synthesis, Image classification**

I. Introduction

Modern medical fields including the healthcare/medical and paramedical fields, medicine, and their respective research equivalent, are increasingly adopting AI-enabled means, such as Machine Learning (ML) and Deep Learning (DL) to optimize their operations. In these fields, the use of DL and ML pushes for innovative and evolved solutions, but also facilitates the operation of quality-of-life services. Specifically, in the medical sector, to successfully train AI models requires the utilization of large quantities of information, especially in the case of commercial deployment, produced by medical equipment that are sensitive in nature. As of late, the privacy of data in the AI fields has been the keen interest of both national and global entities that strive to fortify their security. To this end, the technology of Federated Learning (FL) is adopted in order to decentrally train and optimize DL models remotely. FL removes the need to transfer and process the sensitive data to a central system but only processes the produced models, thus ensuring their security and privacy aspect [1].

One of the primary challenges in Federated Learning is model development and optimization. Specifically, a big effort is being given in finding way to tackle low-quality models, developed and employed, that occur due to bad quality data from the distributed clients. This problem is dependent to the heterogeneous data, large population and pervasive uncertainty [2] and other such factors that can be found in largely distributed machines. These factors introduce unfairness or skewness to the data. This natural data bias is symptomatic of their tendency to be unevenly distributed, leaning towards a specific subset of the classes that is accumulated by the models during training [3]. This indicates that the statistical distribution of the relevant dataset will have a significant impact on the performance of ML and DL models. Thus, the apparent unbalanced bias and the lack of fairness that is identified in the utilised data intended for use in ML or DL implementations determines the quality of the resultant model and its desired output, such as data categorization and augmentation, anomaly detection, decision support, and so on. This shortcoming affects the widespread deployment of the aforementioned algorithms, their performance on the task at hand, and the long-term viability of the system in which they are utilized [4]. This topic is becoming increasingly relevant in medical applications, where even little miscalculations can have disastrous consequences, even resulting in the loss of human lives, necessitating the need for a solid solution. There is minimal evidence showing the stability and robustness of a participating worker's model since FL is a distributed scheme that inherits the difficulties of conventional DL practices. This indicates that an unfair or biased model who participated in the FL training might have a detrimental influence on the overall success of the training, leading to disastrous consequences. Because the training data and its features, such as heterogeneity or distribution, are unknown in the FL process, and because the data in practical scenarios is defined as non-IID, the need for a method to test the fairness of the created local models becomes crucial II-B. In [5] The suggested methodology enforces demographic parity and equalised chances on the local model using in-processing

[†] I. Siniosoglou and P. Sarigiannidis are with the Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece - E-Mail: {isiniosoglou, psarigiannidis}@uowm.gr

[‡] V. Argyriou is with the Department of Networks and Digital Media, Kingston University, Kingston upon Thames, United Kingdom - E-Mail: vasileios.argyriou@kingston.ac.uk

[§] T. Lagkas is with the Department of Computer Science, International Hellenic University, Thessaloniki, Greece - E-Mail: tlagkas@cs.ihu.gr

[¶] A. Sarigiannidis is with Sidroco Holdings Ltd, 1077, Nicosia, Cyprus - E-Mail: asarigia@sidroco.com

[‖] S. K. Goudos is with the Physics Department, Aritostle University of Thessaloniki, Thessaloniki, Greece - E-Mail: sgoudo@physics.auth.gr

[**] S. Wan is with School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China - E-Mail: shaohua.wan@ieee.org

methods and a limited optimization problem. A potential flaw in the method is that the technique does not address the issue of data inequalities between customers, which may impair the model's impartiality. The method introduced in this work intends to solve the problem of acquiring high-quality models for the FL training and optimization procedure by enforcing a process and evaluation criteria, such as a fairness metric, that evaluate the fairness of a model trained on the edge. This methodology, which follows a completely unsupervised manner, intends to help the model selection procedure that is realised during FL. This is done to produce higher quality models than stochastically selecting clients [6] to participate in FL which usually results in needing a high number of clients [7], as shown in the experimental results.

On the federated fairness measuring and optimization aspect, some work on reducing the fairness saturation of the learned data and in extent generated by DL networks has been performed. In [8], the authors produce the FairGAN network, a fairness-aware GAN architecture that learns to be fair during training. It achieves this by augmenting an additional D module that opposes the biased accumulation of data by relying on a corresponding conditional protected attribute pointing at a certain group in the data, $s \rightarrow P_{data}(s)$. One main drawback of this method is that the additional network is introduced in the training process, meaning that the training of the models needs an extra configuration on the remote endpoint in the case of FL, but also requires more intensive resource allocation during the active FL training. The same principle is also explored in [9], where the authors employ an additional D to discriminate unfair bias towards protected categories. They evaluate their work by analyzing the model's bias-variance dilemma to prove its performance against benchmark fairness-oriented datasets. In an effort to map the bias measurement problem and provide an adequate solution, [10] presents a method for evaluating the fairness of a GAN deep network. This work is based on the property of GANs to correlate their overfitting with their classification accuracy. This work shows the relationship between the fairness of the GAN's biased data and augmented data and their distribution. The authors utilize statistical sampling to measure this fairness, based on the under-evaluation network. A setback of this approach is that it relies on the knowledge of the training procedure of the network. Moreover, the considered assumptions of the methodology are not representative of a wide variety of augmentation networks, except GAN networks. In another work, the authors in [11] present an adversarial representation learning methodology, ensuring the fairness models used by third parties. They apply well-known fairness evaluation measures, such as a) demographic parity, b) equal opportunity, and others, to the adversarial training process to establish a discriminatory baseline. They evaluate their strategy with experimental findings that demonstrate the usefulness of their concept. This strategy, however, imposes on the proposed model's training process, refocusing it on the accurate parameterization of the model during training in a data-dependent way. In [12], the authors propose FairFed, a novel algorithm for fairness-aware aggregation in federated learning, was proposed to enhance group fairness in machine learning models. FairFed's empirical evaluation demonstrated that FairFed provides fairer models, particularly under highly heterogeneous data distributions across clients. One potential flaw in the suggested method is that the fairness metric utilised to adaptively alter the aggregate weights of various clients depending on their local fairness metre may be insufficient or may not capture all aspects of fairness relevant to the challenge at hand. This might lead to model update biases and, eventually, lower-quality global models. In, [13], a set of metrics is defined, namely, a) the Fano inequality and b) the structural similarity index that measures group and individual fairness. The issue is however that the notion of individual fairness is task-specific, limiting the proposed fairness measures. Specifically, both measures depend on in-task and data-specific aspects that might not be applicable in every application. Furthermore, the suggested method for attaining individual fairness necessitates that each device calculates and exchanges its inequality value with the FL server which might create privacy concerns. Lastly, in [14], the authors present a post-processing fairness measuring method that evaluates trained Federated models. In the paper they experiment with a processing pipeline that leverages latent GAN deformations to quantify the model's predictions. They produce an unsupervised ranking of each model's learned classes by clustering the deformation predictions. The method quantifies the fairness of each model based on the performance of the rest of the models using Fdi, a custom fairness metric.

It is also deemed necessary to mention currently utilized metrics for the goal of model optimization tackling the fairness problems. As is defined in [15], currently used fairness metrics can be categorized in three major categories, namely, a) Pre-Process, b) In-Process, and c) Post-Process, in respect to the stage they are measured, i.e, before, during or after the training process. These use measures like i) Normalized prejudice index [16], ii) Disparate impact [17], iii) Equalized odds [18], and so on, which rely on statistical measures targeting the training data or how they are used by the training process. The problem with these metrics and the reason for omitting them from the comparative results of this work is that these are mainly statistical-based fairness evaluation metrics aiming at the pre-training testing (i.e., the data to be used in the training) and as so diverge greatly from the scope of the proposed methodology.

Finally, a problem that current research is emphatically trying to solve is the interpretability and transparency of Machine and Deep Learning models. Until recently, AI models were seen and used a Black-Box, mostly due to the complex, non-linear, manner in which they learn and in extent make decisions. In the healthcare sector, were decisions concerning patients must be made using concrete justifications, the need to be able to understand the reasoning behind a decision made by an AI model, becomes critical [19][20]. This sector falls under the organized research of Explainable Artificial Intelligence (XAI) [21]. This work delves in this field by producing interpretable visualizations of what the models have learned, as discussed in Section III.

Based on the aforementioned notions and problems this work, striving to solve the biased skewness of the learned knowledge in Federated models, offers the below contributions:

- Presents a novel quantification pipeline oriented in measuring the fairness or bias of DL models without knowing the model's architecture or the training data.
- Introduces and processes micro-Manifolds for latent knowledge clustering and discrimination for model optimization
- Presents a latent knowledge visualization method for trained DL models adding to the explainability of the produced models
- Designs and presents a novel and end-to-end evaluation pipeline that organises the data, enforces bias, and measures the holistic performance of the DNNs in a comparative manner

In particular, this work tries to fill the gap in the fairness measuring methodology up to now, by producing and evaluating a strictly unsupervised pipeline that can measure and visualise the volumetric properties of each DNN. This also aims to help the Federated process by evaluating the aggregated models, post-process, and identifying stragglers that possess and offer low-sample and low-quality data to the FL process. The proposed methodology is designed to produce a quantifiable metric to assess and describe the fairness of each DNN.

The remainder of this work is structured as follows. The tools used to implement the suggested strategy are described in Section II. Section III outlines the approach developed to solve the fairness discovery problem, whereas Section IV offers the evaluation findings of the offered work. Furthermore, hyperparameter ablation is created in Section V. Concluding, Section VI brings this effort to a close.

II. Background

In this section, the necessary background for the tools and methods used in this work are presented. In particular, i) the GAN Deep Neural Network architecture is explained and ii) the Federated Learning environment is presented.

A. GAN Architecture

Two neural sub-networks, namely the Generator G and the Discriminator D, form the basis of the GAN architecture and compete with each other [22], [23] in an adversarial game. G usually takes as input random noise data and tries to produce similar-to-the-real data of the given use case. On the other hand, D is trained to identify the real samples and the fake data produced by G. The GAN architecture aims in training the two rivaling sub-networks, in a manner that G is able to generate realistic samples, which cannot be differentiated by D from the real data and vice versa. Equation 1 shows the relation between both sub-networks.

$$\begin{aligned} \underset{G}{min}\underset{D}{max} V(G,D) = \underset{G}{min}\underset{D}{max} \mathbb{E}_{x\sim p_{data}}[log(D(x))] + \\ \mathbb{E}_{z\sim p_z}[log(1-D(G(z)))] \end{aligned} \tag{1}$$

where G accumulates from space Z noise z and maps it to the space X which is used by D to input x. The probabilistic distribution of spaces X and Z are denoted by $p_{data}(x)$ and $p_z(z)$, respectively.

Following the training from the Discriminator module, an intermediate model is produced for the Anomaly Detection procedure. The specific model is placed within the Discriminator between the input and a latent layer in advance of the network's output sequence. The model is used to reduce the dimension of the input samples into a specific latent space.

B. Federated Learning

Federated Learning is a distributed stochastic learning and privacy-preserving approach that enables the orchestration, distribution, learning, and aggregation of Deep Learning models over several cloud devices or edge nodes [24]. It works by stochastically dispersing a central Deep Learning model among a specific corpus of nodes in order to train locally the on-device acquired data. As a result, the models are returned to the centralized system and input into a process known as Federated Averaging [1], which aggregates the edge-calculated weights with the central model.

Specifically, the central server distributes a global model w^0_{Global} along with training instructions to a Federated population $P_f \in [1, N]$ where $N \in \mathbb{N}^*$, each holding a set of local data $D_{i\in N}$ and local models w^i_l. The distributed models are subsequently trained on the local data D_i and then the weights w^i_{Global} are send back to the central system to be aggregated through the Federated Averaging (2), or similar, process in order to produce an updated global model w^k_{Global} [25].

$$w^k_G = \frac{1}{\sum_{i\in N} D_i} \sum_{i=1}^{N} D_i w^k_i \tag{2}$$

Here w^k_G is the global model at the k_{th} training iteration and w^k_i denotes the Federated population i_{th} model at that iteration.

C. Unsupervised Latent Direction Discovery

The unsupervised Latent Direction Discovery [26] is an exploratory process through which different directions in the latent space of a model are identified. Traversing these directions in a GAN network can change its output based on its knowledge, adding/subtracting or transforming elements, called semantic manipulations. For example, let's assume a GAN network trained on dataset containing a large number face images. The latent space of this model contains information about the different characteristics of a face, such as expressions, hair color and skin tone, eye color, and so on. By traversing a latent direction in the latent space of this model it is possible to input an image of a person with a neutral expression and produce an image of that person smiling, based on information accumulated of persons with smiling expressions. This methodology utilises a specialized GAN network, aimed at mapping the directions in a model's latent space. This mapping results in a number of latent directions and corresponding semantic manipulations (shifts) in those directions. Using these will the output of the model in question can be manipulated. To do this, two components are trained for each model G_i, first a matrix $AER^{d_i K_i}$, where d_i is the dimensionality of the latent space of G_i and K_i corresponds to the number of directions in the latent space. Then, a reconstructor R_i, which obtains an image pair $(G_i, G_i(z + A(ae_k)))$ and outputs $R_i(I_1, I_2) = (k', a')$. Here e_k is a unit vector and a is a scalar, while k' and a' are the prediction of a direction index k, and a prediction of a shift magnitude a, respectively. The learning process is performed using a minimisation process on the following loss function:

$$\underset{A,R}{min} \underset{z,k,a}{E} L(A,R) = \underset{A,R}{min} \underset{z,k,a}{E} [L_{cl}(k,k') + cL_r(a,a')] \tag{3}$$

where L_{cl} is the cross-entropy and L_r is the mean absolute error, while we experimentally found $c = 0.25$ to be optimal. After the training is completed, R has produced k latent directions.

III. Methodology

This section delves into the implementation of the proposed methodology of this work. Our work relies on the methodology developed in [26] which undertakes the design and development of the Unsupervised Latent Direction Discovery technique, which was analyzed in the Background section. The proposed methodology is divided into four main pillars, namely, *a) Pre-trained Model Preparation* (only for evaluation purposes), *b) Latent Direction Discovery*, *c) Latent micro-Manifold Processing* and lastly, *d) Model Ranking*, Figure 1. Overall, this methodology aims to assess the fairness level of deep learning models and augment them without prior knowledge of training data, hyperparameters, or training/evaluation environment. The methodology is defined in the shape of a pipeline that can be used sequentially to measure the fairness of each individual model. In particular, the process starts by exploring the disentangled latent knowledge of the model through Latent Direction Discovery in order to create a perspective of what each model has learned. Next, sample points in the disentangled directions are processed and clustered to micro-manifolds of latent knowledge to create a quantifiable volume of the model's knowledge. Finally, by analysing the properties of this volume, significant information can be obtained, which leads to the fairness evaluation of each model.

A. Pre-trained Model Preparation

As stated before, the approach discussed in this study focuses on finding the fairness level of prediction and augmentation Deep Learning models with no prior knowledge of the training data, hyperparameters, or training/evaluation environment. A default training method is offered in this section to robustly explain the technique and provide a global view of this work. The Unsupervised segment, which would be used in an FL system, is fully documented in section III-B and below. Two different DL model architectures were employed in this work since we refer to two sorts of DL use cases, namely a decision-based and specifically a classification scenario and a data augmentation/generation scenario. To that end, a GAN network and a rudimentary DNN classifier were built and tested following the given method.

1) Generative Network Architecture: One of the two Deep Learning architectures selected for this work is the GAN architecture. GAN networks are widely used in a variety of scientific fields because of their powerful ability to accumulate and generate data spaces but also due to their versatile aspect and adaptability to a plethora of applications, such as anomaly detection, regression, classification, dataset generation and so on. To implement and test the proposed methodology the DCGAN (Deep Convolutional GAN) [27] scheme was selected due to its simplicity and wide application in the medical sector but also because it can be abstracted to extend subsequent GAN architectures. In the presented pipeline and to measure the Fairness of the GAN models only the Generator module is used while the rest of the network is treated as a black box.

2) Classifier Network Architecture: The second architecture that is investigated is the simple DNN image classification scheme. This network is made to accumulate a batch of preprocessed images and outputs a vector of length $C \in \mathbb{N}^*$ in the form of a probability vector containing the correlation of the input to each predicted class. The classifier is trained on the benchmark data selected for this work, section IV, and is integrated into the generated pipeline by leveraging an augmentation head as is described below.

To better justify the proposed methodology, a more advanced classification DL model was selected. The ResNet18 architecture [28], a branched scheme of the ResNet architecture [28], which is one of the most powerful series of deep neural networks with high performance on most benchmark datasets, was leveraged. Like the simpler DNN model, ResNet18 models were trained on the given data and the resulting trained models were integrated to this work's pipeline.

B. Model Latent Knowledge Discovery

In the proposed methodology, the models under revision go through the process of Latent Direction Discovery. This methodology produces K directions and S shifts to deform the output of that model. A subset of those K directions are chosen and based on S shifts along those directions are uniformly implemented upon N number of random noise samples to produce a $K \times S \times N$ matrix M of latent deformations of the original DL model's knowledge. This is performed to assert a knowledge baseline of the model while traversing the different directions within its latent space, Figure 2. The larger the value of K, S, and N the higher the probability of capturing a higher percentage of the latent knowledge of the DL model.

Up to this point, and utilizing simply the GAN's G module as an example, the procedure has flowed smoothly with the supplied model. However, an issue occurs when evaluating the second tested case, the classification model, because a classifier, by definition, is not an augmentative model and does not create data, or at least data in a comprehensible manner. This raises the question of how to employ it in the Latent Deformation process. Two approaches were investigated in order to overcome this challenge. The initial approach would to adapt the classification model to the Latent Deformation process would be to extract an intermediate model up to a

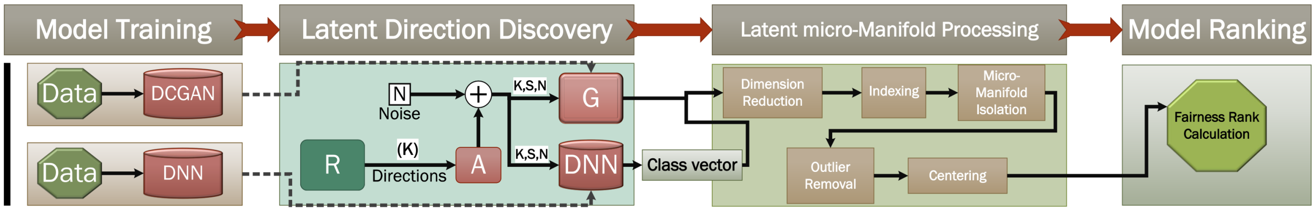


Fig. 1: Methodology pipeline

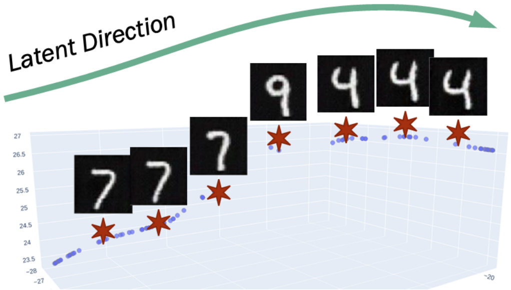


Fig. 2: Latent Direction Transversal Effect

Latent layer L_l. That layer's output can then be translated into a two-dimensional matrix describing image data and given to the pipeline. Of course, the generated image would not be translatable to any human visible pattern, but it would be transmitted to the pipeline as information. One disadvantage of this strategy is that portions of the model's layers would be removed, potentially resulting in information loss. The second solution, which is also chosen for this study, is to add a passive output reshaping unit to the classifier. This reshaping head would make the model produce image data in the same way as an augmentative model would. Because the additional component operates in a tunneling fashion, all of the information stored in the classifier is kept without introducing any learnable parameters to the model. Fig.1 also depicts the coupling of the classifier with the passive output reshaping head. We shall use the terms "model" and "generator" interchangeably throughout the rest of the paper to describe either classification or augmentation models.

C. Latent micro-Manifold Processing

The generated data M are subsequently aggregated and then scaled and normalized to a specific range. Next, a dimensionality reduction algorithm (e.g. tSNE) is applied, transferring the data of D dimensions to a space M^d, where $d << D$, trying to accommodate a balance between size loss and information loss. Since normalizing and reducing the data to a specific dimension (Dimensionality Reduction) are reliant on the training data, no threshold can be specified that can be used in a wide variety of cases. In this work, we project the data to a three-dimensional space for illustration purposes. However, before performing the dimensionality reduction, the data are sufficiently randomized to eliminate any biases caused by the nature of this technique. Two indexes K_{index} and N_{index} are kept in parallel in order to reassemble the shuffled data to its original order pointing to the direction and noise they belong to. The structure of the data can be seen in Fig. 3.

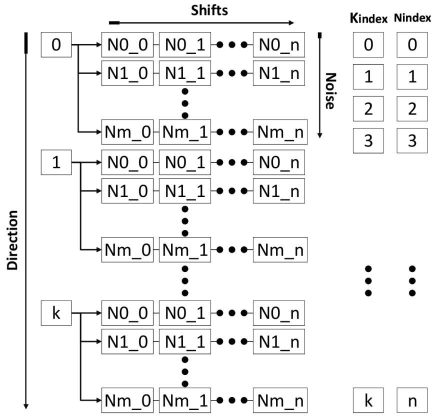


Fig. 3: Structure of generated samples

After indexing and reassembling the data, small clusters that were created in the overall data manifold due to the deformed latent knowledge are isolated. These small cluster define the micro-Manifolds described in this work. In this instance, it was observed that micro-Manifolds are formed by the points corresponding to the different noise inputted in the deformation network, which is deformed S times along K directions. This results in the creation of N micro-Manifolds in the overall information manifold, formed from a set of kxs elements, where $k \in K$, $s \in S$. Fig. 5 depicts two different micro-Manifolds, a two- and a three-dimensional. Following, the outliers from each micro-Manifold are eliminated in order to reflect the clean latent information around each noise n. As mentioned in [26], outliers are created because, except for the interpretable deformations, like, zooming, rotating, augmenting features, and so on, there exist also transformations that deform, saturate or even transform an object into another. In the digit MNIST dataset, for example, a latent direction might convert a digit into an unrecognizable/different shape or even rebuild it as another digit, Figure 2. These directions are dropped because they have deemed outliers in relation to each noise n. The elbow approach is used to determine the threshold at which outliers are recognized. In this work, the Kneedle algorithm [29] was leveraged to dynamically calculate the outlier removal threshold. In this work, we denote the outlier removal threshold θ which is calculated using the elbow method. In particular, the distances, calculated by Eq. 4, between the consecutive points in a latent direction for each direction in a cluster are inputted and the system removes any information furthest than the biggest gap in the direction.

This indicates an information disparity in the corresponding direction, removing data from the cluster centre.

$$d = \sqrt{\left|b_i^2 - b_{i-1}^2\right|} \tag{4}$$

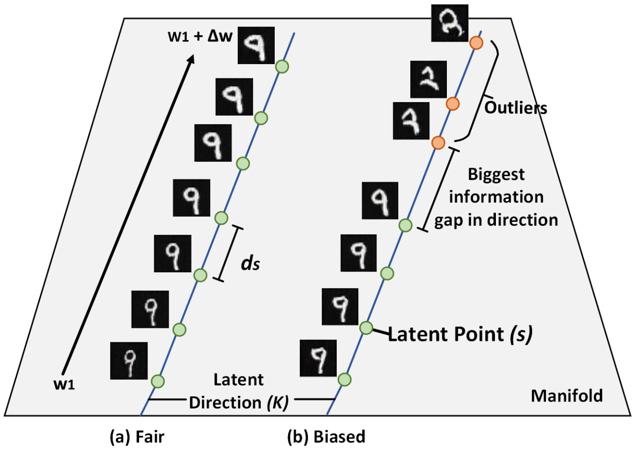


Fig. 4: Latent Direction Quality: (a) Fair, (b) Biased

Where b_i denotes the i^{th} element of a direction.

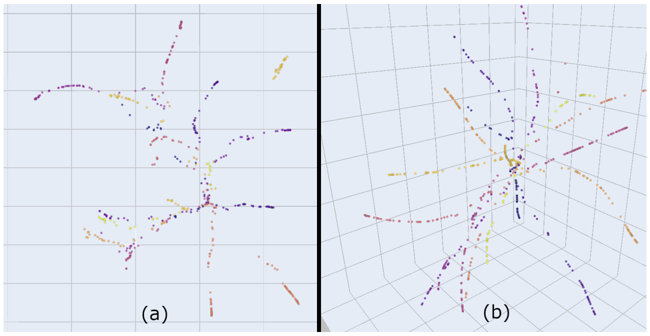


Fig. 5: (a) 2D micro-Manifold, (b) 3D micro-Manifold without outliers

Micro-Manifolds are also centered, by simply removing the mean as seen in Eq. 5.

$$M^i_{zero} = M^i - \overline{M^i} \tag{5}$$

Here, M^i_{zero} denotes the i^{th} noise index, with no outliers. M^i denotes the i^{th} element of the original matrix, and $\overline{M^i}$ denotes its mean average. Finally, the micro-Manifolds are normalized in a predefined range. The described process can be seen in Fig. 6.

D. Model Ranking

The final phase in the proposed methodology is the calculation of the Fairness Rank of the DL model. This is achieved through the volumetric analysis of the discovered and processed micro-Manifolds. Specifically, the distance among neighboring points in each direction of a noise sample is calculated and then it is averaged along the different directions within the same noise vector. This computes the density, or rather the sparsity of the micro-Manifold. To calculate a volumetric representation of the model's knowledge, the average of the mean of all the micro-Manifolds is calculated, Eq. 6. Through this process, we can quantify the knowledge or lack thereof in a deep learning model.

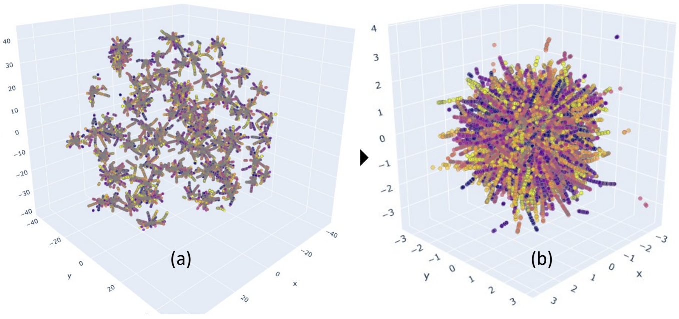


Fig. 6: (a) Manifold, (b) No outliers and centered

$$\overline{\rho} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{S-1}\sum_{s=1}^{S-1} d_s\right)\right] \tag{6}$$

where $\overline{\rho}$ is the mean density and d_s denotes the distance between pair of elements s in direction K. The same is done to calculate the mean standard deviation of the distances in Eq. 7,

$$\overline{\sigma} = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K} std(d_k])\right] \tag{7}$$

where d_k describes the distance vector of a certain direction K. This process is also depicted in Figure 4, were w_1 denotes the model weight and $w_1 + \Delta w$ denotes the transversal of the weight in direction K by Δw.

The final Fairness Rank can be calculated by using the metrics collected up to this point, seen in Eq. 8. This process server is also a secondary purpose. By taking advantage of the information of the density in the different directions in the noise vectors against the outliers discovered in that direction the quality of each direction can be actively inspected. Measuring the density of a direction in space N will reveal the quality distribution of that direction in the latent space, thus differentiating between useful and fewer directions in the respective latent space.

$$F_f = \frac{\overline{\rho\sigma}}{\frac{\overline{\mu*}}{\overline{\mu}} + 1} \tag{8}$$

Here $\overline{\rho}$ is the mean manifold density, $\overline{\sigma}$ denotes the mean standard density deviation while $\frac{\overline{\mu*}}{\overline{\mu}}$ denotes the fraction of mean manifold outliers $\overline{\mu*}$ over the mean population $\overline{\mu}$.

IV. Evaluation

In this section, we present the evaluation environment, the leveraged datasets, and the various metrics and results produced by the proposed techniques.

A. Evaluation Environment

To realize the outlined methods, a mid-range evaluation system was utilized. The simulations were performed on a workstation utilizing the Linux OS, and relying on 16GB of RAM memory, an i7 Intel core processor, and an NVIDIA GTX 1080 8Gb GPU. Since the resource allocation and experiment times are relative to the evaluation environment, they are not presented here. Nevertheless, it was observed that the developed approach is computationally expensive for three reasons. First, the outlined method requires the training of the GAN network that maps the latent space of the evaluated models, which is a computationally arduous task. Furthermore, the methodology produces a large number of samples and performs dimensionality reduction on the data in order to map the latent knowledge of the Deep Learning models. Dimensionality reduction is also a method that needs expensive computations. This adds to the resource requirements of the developed method. This although should not be considered a drawback that hinders the Federated application, since the proposed methodology is not an online task but can be performed asynchronously.

B. Simulation Data

To substantiate the proposed approach, the need for targeted sets of data arises. Thus, it was deemed necessary to utilize benchmark data, both from the field of medical application, but also widely applicable datasets that are used to validate DL models. For this work, two datasets were selected, namely, the a) DigitMNIST [30] and the b) MedMNIST [28] (Medical Mnist) benchmark datasets. The Digit MNIST dataset comprises multiclass data containing images of the handwritten digits $0-9$ for general DL use, while the MedMNIST dataset contains a series of multi-modal sub-datasets covering diverse data scales and purposes, respectively, like classification, regression, anomaly detection, and so on. For the purpose of the experiment, the DigitMNIST and the PathMNIST sets were chosen for their homogeneity of sample numbers and multiclass features. Fig. 7 depicts samples from these two datasets.

At this point, a notice has to be made regarding the reasoning behind the use of the DigitMIST data collection. The reason for using this particular dataset involves two factors. First, this methodology does not aim in solving a data-oriented problem, like classification or regression, but rather tries to optimize the model itself without knowing its purpose by fining badly trained models. The second factor involves the dataset's property of being by its nature balanced and does not contain natural biases (like class overlapping) and, thus, we can control the bias distribution to establish the correct validation of our methodology. These two notes combined with its the wide use of the chosen data constitute the reasons for its use.

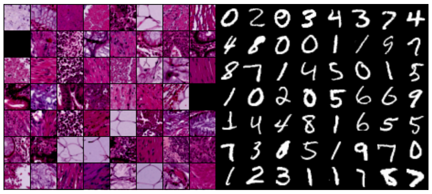


Fig. 7: Samples of: (left) PathMNIST, (right) DigitMNIST

Concerning the DCGAN network and the classifier network, both were trained with the i) DigitMNIST fairly trained and biased on the separate class 1/3/5, respectively and ii) PathMNIST fairly trained, and biased on the separate class 1/3/5, respectively. The biased networks were trained on 10-30 percent of the bias-induced class. After the training process and the Latent direction discovery, the models were made to augment a total of 100000 samples each. To do that, $K = 10$ random directions, $S = 100$ random shifts per direction, and $N = 100$ random noise were selected, creating a manifold of $[10x100x100]$ for each case, respectively. Two different dimensionality reduction algorithms were used for the experiments, namely, the i) tSNE [31] and b) PCA [32]. In the performed experiments, data were normalized leveraging a Min-Max Scaler to the range of $[0, 1]$ and reduced the manifolds to the dimensions 2 and 3. For the outlier removal, an empirical custom threshold was used based on the specific manifold at hand with the best size/information loss trade-off or it was dynamically calculated through the elbow method. All the parameters used for each manifold, as well as the produced results on the performed experiments, are presented in Tables II, III, IV and V. After that, the findings of the experiments are examined. Each experiment evaluates the performance of the proposed technique against key data aspects, such as, a) Data bias (fair, % of a class), b) Dimensionality reduction (e.g., tSNE, PCA), c) Reduced Dimension (e.g., 2D/3D), d) Dataset (DigitMNIST/PathMNIST), e) DL Architecture (GAN, DNN). Note that during the indexing of the manifolds, some values are dropped as invalid. These values occur due to the stretch of the deformation component [26] when drawing the latent directions, resulting in invalid values (like out of index or nan values), leading to some differences in the final population of the manifold. All of the values reported in the experiments were estimated relative to the resultant population of each experiment. The notations used in the tables describing the results are shown in Table I. Furthermore, Fig.8 presents the outline of the experiments realised in the premise of this work. In particular, Fig.8 presents the different aspects of the experiments, e.g., Data bias, Dimensionality reduction, and the network architectures tested against the proposed methodology as described above.

C. Evaluation Experiments

The experiments described in Table II were produced using different combinations of thresholds, dimension reductions, and dimension reduction algorithms and, of course, data distributions for each clustered of correlated cases (models). A cluster of experiments is defined as the arrangement of experiments with the same dataset, dimension reduction algorithm, and dimension of manifolds. The elbow approach was used to dynamically determine all of the thresholds in Table II (averaging the produced thresholds for each cluster for uniformity). The initial experiments were carried out using the

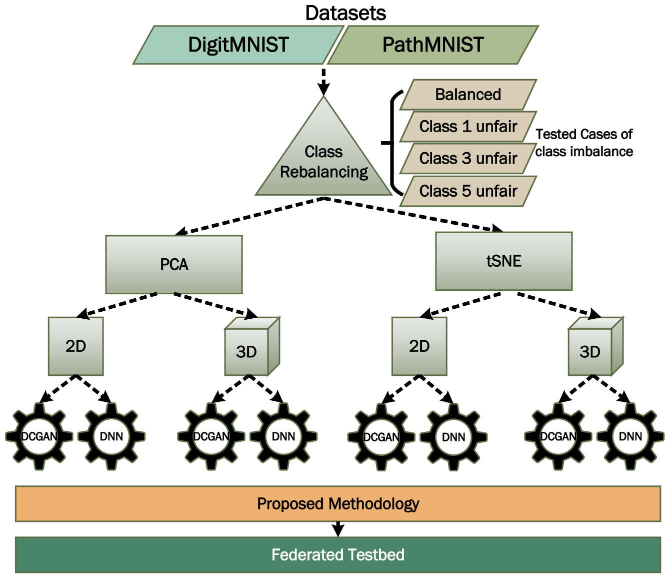
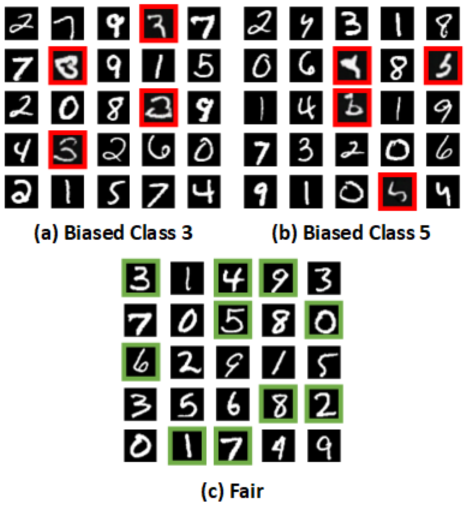


Fig. 8: Experiment Outline

TABLE I: Result notations

Notation	Meaning
DR	Dimensionality Reduction Algorithm
D*	Reduced Dimensions
θ	Outlier Removal Threshold
$\overline{\rho}$	Mean Density
$\overline{\sigma}$	Mean Density Standard Deviation
$\overline{\mu}$	Mean Manifold Population
μ	Manifold Population
$\overline{\mu*}$	Mean Outlier Number
$\mu*$	Outlier Number
$\mu*/\mu$	Outliers / Manifold Population
F_f	F factor
$log(F_f)$	F factor logarithmic
W*/*/.../*	Workers with id *
acc	Accuracy
rec	Recall
pre	Precision
f1s	F1 Score
BC	Biased Class

tSNE method, which produced 3D manifolds and leveraging a θ value of 0.8. It can be shown that in the case of the biased classes, the mean outlier number increases, indicating that the manifolds have a greater percentage of decorrelated values. The Fairness factor increases as the model becomes more biased. By reviewing the visual results in Figure 9, by examining the digits, generated by each model, namely, a) the model with bias on class 3, b) the model with bias on class 5, and c) the fair model which was selected by the proposed methodology, it can be seen that the method finds the more balanced and higher quality model. In particular, in the first and second case, it can be seen that the generated digits are severely saturated in the respective biased classes. On the other hand, the fair model produces higher-quality samples. By selecting the fair model, the final aggregated FL model will be able to generalise much faster, producing better results. In the second cluster having 3D PCA reduced data, there seems to exist a contradiction. According to F_f the fairest model is a biased one. In fact, the fair model has the lowest score. The same can be seen in the fourth clustered, featuring 2D PCA reduction. PCA seems to produce sparser micromanifolds with big value distances between a steady number of consecutive direction values. Taking these experimental results under consideration it is shown that the PCA method is not suitable for calculating the fairness of the models with the proposed Fairness Factor. Nevertheless, the micromanifold representation of PCA shows a robust knowledge representation so further experimentation should be performed. The same also applies to the reduction of the data to 2D manifolds as the clusters featuring 2D data reduction show inconsistent results. This is a result of excessive data loss due to $D >> d$.


(a) Biased Class 3

(b) Biased Class 5

(c) Fair

Fig. 9: Generated Digit Quality of Evaluated DNNs: (a) biased class 3, (b) biased classes 5 (c) fair *[Selected]*

Table III shows the experiments performed on DCGAN models trained on the DigitMNIST dataset. The tSNE algorithm is subsequently used to reduce the data in three directions. A θ value of 2, which was experimentally selected, was used to remove the outliers in the manifold. As would be the case in a Federated Environment, the DL models were initialized with the same weights and then trained on the provided data distributions. Two fair models and two unbalanced models were trained for the experiment. As can be seen from the results, the two fair models hold lower F_f values than the unbalanced ones. Moreover, comparing these results with the ones from the first cluster of Table II it can become apparent that the models trained on the unfair towards class 3 dataset tend to be worse than the rest of the unbalanced ones. This demonstrates that in the DigitMNIST data collection, class 3, has a significant influence on the resultant trained model.

Finishing with the DCGAN model experiments, Table V reveals the performance of the models on the PathMNIST medical dataset. An interesting note in the results of this

TABLE II: Experiments on DCGAN-DigitMNIST on different dimensionality reductions

Dataset	**Context**	**DR**	**D***	θ	$\overline{\rho}$	$\overline{\sigma}$	$\overline{\mu}$	μ	$\overline{\mu*}$	$\mu*$	$\mu*/\mu$	F_f	$log(F_f)$
DigitMNIST	fair	tSNE	3	0.8	11.4557	6.7832	436.61	43661	36.6188	1137	0.0839	71.6934	$\underline{4.2724}$
	biased class 3				15.4486	8.1926	481.67	48167	41.2518	1142	0.0856	116.5794	4.7586
	biased class 5				12.4335	7.4635	467.9	46790	39.5952	1018	0.0846	85.5574	4.4492
DigitMNIST	fair	PCA	3	1	1.9368	0.746	465.38	46538	51.0222	1024	0.1096	1.3022	0.264
	biased class 3				1.9332	0.7455	465.29	46529	47.7444	1024	0.1026	1.3071	0.2678
	biased class 5				1.6344	0.7109	467.33	46733	46.2385	1274	0.0989	1.0574	$\underline{0.0558}$
DigitMNIST	fair	tSNE	2	1.7	12.6381	9.6439	859.78	85978	48.4452	1035	0.0563	115.3799	4.7482
	biased class 3				12.8862	9.9498	846.02	84602	46.2543	1118	0.0547	121.5689	4.8005
	biased class 5				11.2267	9.4876	863.29	86329	53.1292	1166	0.0615	100.3398	$\underline{4.6086}$
DigitMNIST	fair	PCA	2	1.9	1.5154	0.7588	879.97	87997	55.4833	1097	0.0631	1.0817	0.0785
	biased class 3				0.7553	0.7553	880.11	88011	43.8596	1097	0.0498	1.0911	0.0872
	biased class 5				1.3508	0.6795	877.74	87774	43.8151	1340	0.0499	0.8743	$\underline{-0.1343}$

TABLE III: DCGAN-DigitMNIST results with same initialized weights

Dataset	**Context**	**DR**	**D***	θ	$\overline{\rho}$	$\overline{\sigma}$	$\overline{\mu}$	μ	$\overline{\mu*}$	$\mu*$	$\mu*/\mu$	F_f	$log(F_f)$
DigitMNIST	fair 1	tSNE	3	2	9.5497	6.0059	890.02	89002	57.8095	1016	0.065	53.8568	$\underline{3.9863}$
	fair 2				10.6185	5.5618	875.05	87505	46.3709	1158	0.053	56.0861	4.0269
	biased class 3				14.2306	9.2214	867.22	86722	52.1042	1016	0.0601	123.7888	4.8186
	biased class 5				12.3917	7.6142	862.28	86228	51.014	1161	0.0592	89.0835	4.4896

experiment is the notably lower $\overline{\sigma}$ of the fair model in contrast to the unbalanced models. The PathMNIST data, in contrast to DigitMNIST, is defined by more complicated illustrations that feature a variety of patterns versus handwritten digits with just white edges and vertices on a black background. The results demonstrate that the directions predicted by the deformator on the imbalanced models are significantly sparser than those for the fair model, validated by the change of value $\overline{\sigma}$. Simply, the micro-Manifolds describing the knowledge in the fair model are thicker with smaller gaps between samples, which gives the model stability. Again, as shown in the reported results, the fair model has the lowest F_f.

Table IV presents the output of the DNN classifier on the DigitMNIST dataset. The classifier was trained to predict the class of an input picture as a probability vector, with the greatest value indicating the class to which the sample belongs. As previously stated, a reshaping unit was added to convert the probability vector to an image matrix in order for the classifier to be consistent with the suggested technique. As can be seen, the models associated with the fair and biased classes 3 compete for the lowest F_f. The results are contained in Table VII, which holds the Federated experiments with the DNN model on the DigitMNIST dataset. Four distinct configurations were implemented in the Federated context, namely, a) all models balanced, b) one worker is unbalanced in class 1, c) in class 3, and d) class 5 respectively. The aforementioned models only have 0.2 difference in their accuracy scores which justifies the close F_f values. This implies that with a further optimized θ value instead of the statically chosen $\theta = 2$, the F_f would better reveal the fair model. In the same manner, Table VI presents the experiments realised using the widely used ResNet-18 classification architecture. As can be seen, the proposed algorithm is able to differentiate the differently biased models of the ResNet algorithm, clearly outlining the correctly trained and balanced model.

In Table VII we compare the results of the proposed methodology against the work performed in [14]. The results depict the accuracy, precision and f1 score measured by the resulting model that the federated process produces. The results show the quality of the models after the selection process occurring by using the two different methods. In retrospect, both methods strive to quantify the knowledge accumulated by a model in order to measure how well it has learned that data. In particular, in [14], fairness is tested by augmenting a large number of samples, clustering them and then quantifying the GAN models' prediction, measuring their tendency to follow a specific distribution. In contrast, our study visualizes and evaluates the volumetric properties of small manifolds (Fig.6) created around noise points when augmenting the model's input using permutations of that noise point. This reveals gaps in the learned classes. In Table VII it can be seen that the models chosen by the proposed method of this work, in the different bias cases, produce better quality models than the ones selected by the method introduced in [14]. Specifically, we can see a steady increase in the quality of the models with an average of 8.75% increase in accuracy, 9.69% in precision, and 8.77% in F1-Score, respectively.

V. Ablation

To distinguish the micro-Manifold centered changes over the change of key hyperameters, we ablate parameter θ, seeing the scaling over the mean density of the produced clusters Table VIII. The cascading effect of the outlier omission is visible in the experiments on the DigitMNIST dataset. As the θ value increases, fewer outliers are omitted pointing to an increase of the standard deviation and density, as well as the population of the manifolds. Its effects are also visible in the calculated fairness.

The methodology proposed in this work is oriented in an unsupervised way to discover the latent knowledge of Deep Neural Networks in order to aid the Federated Learning process produced high-quality models. Even though in this paper we present an experimentally driven approach to produce quantization and visualization of the latent knowledge of

TABLE IV: Classification DNN-DigitMNIST results

Dataset	Context	DR	D*	θ	$\overline{\rho}$	$\overline{\sigma}$	$\overline{\mu}$	μ	$\overline{\mu*}$	$\mu*$	$\mu*/\mu$	F_f	$log(F_f)$
DigitMNIST	fair	tSNE	3	2	5.7822	0.8073	918.15	91815	57.875	1007	0.063	4.3912	1.4796
	biased class 1				5.7935	0.836	911.6	91160	50.9093	1250	0.0558	4.5872	1.5233
	biased class 3				5.7828	0.7931	911.69	91169	44.4167	1016	0.0487	4.3731	1.4755
	biased class 5				5.9052	0.8724	911.66	91166	55.9496	1161	0.0614	4.854	1.5798

TABLE V: Experiments on DCGAN-PathMNIST

Dataset	Context	DR	D*	θ	$\overline{\rho}$	$\overline{\sigma}$	$\overline{\mu}$	μ	$\overline{\mu*}$	$\mu*$	$\mu*/\mu$	F_f	$log(F_f)$
PathMNIST	fair	tSNE	3	2	5.7822	0.8073	918.15	91815	57.875	1007	0.063	4.3912	1.4796
	biased class 1				9.1794	7.0835	899.4	89940	68.5	1015	0.0762	60.4199	4.1013
	biased class 3				9.5536	6.9159	928.37	92837	48.0938	1030	0.0518	62.8174	4.1402

TABLE VI: Experiments on classifier ResNet18-PathMNIST

Dataset	Context	DR	D*	θ	$\overline{\rho}$	$\overline{\sigma}$	$\overline{\mu}$	μ	$\overline{\mu*}$	$\mu*$	$\mu*/\mu$	F_f	$log(F_f)$
PathMNIST	fair	tSNE	3	2	20.9837	9.59	919.77	91977	47.5585	1370	0.0517	190.5213	5.1345
PathMNIST	biased class 1	tSNE	3	2	18.4441	11.1881	932.53	93253	49.6804	1202	0.0533	195.9168	5.2777
PathMNIST	biased class 3	tSNE	3	2	23.753	10.8686	929.95	92995	49.9543	1310	0.0537	245.0005	5.5013
PathMNIST	biased class 5	tSNE	3	2	20.6939	10.8702	910.4	91040	47.9232	1183	0.0526	213.6966	5.3646

TABLE VII: Federated model (DNN classifier) results on DigitMNIST, *BC:Biased Class

Method	$log(F_f)$				Fdi [14]			
No. Workers	W0/1/2/3	W0/1/2	W0/1/2	W0/1/2	W0/1/2/3	W0/1/2	W0/1/2	W0/1/2
Biased Worker	None	W3	W3	W3	None	W3	W3	W3
BC*	None	1	2	3	None	1	2	3
acc	0.9728	0.9363	0.9521	0.9446	0.9638	0.8542	0.8857	0.8451
acc (%)	*+0.92%*	*+8.76%*	*+6.97%*	*+10.53%*				
pre	0.9728	0.9445	0.9548	0.9517	0.9615	0.8528	0.8786	0.8433
pre (%)	*+1.17%*	*+9.70%*	*+7.98%*	*+11.39%*				
f1s	0.9728	0.9376	0.9525	0.9456	0.9677	0.8588	0.8792	0.8487
f1s (%)	*+0.52%*	*+8.40%*	*+7.69%*	*+10.24%*				

TABLE VIII: Ablation of θ on DCGAN/DNN - DigitMNIST

DCGAN				**DNN**
Parameter	θ =0.8	1.7	2	2
$\overline{\rho}$	13.1126	12.25	14.1268	5.8159
$\overline{\sigma}$	7.4797	9.6937	8.2989	0.8272
μ	46206	85636.33	86168.17	91327.5
F_f	91.2767	112.4295	117.4763	4.5513
$log(F_f)$	1.9603	2.0508	2.0699	0.6581

neural networks, some limitations are still in need of resolution. First, as mentioned in the result evaluation section, the methodology proposed is quite computationally intensive. This means that a suitable system runs the proposed pipeline. Of course, since in Centralized Federated Learning the aggregator, in most cases, is a central server, this issue can be resolved. This though is scaled based on the number of models that need to be examined before the Federated Fusion process. Furthermore, even though, in this work we examined classification and augmentative architectures, the field of Deep Learning is rapidly expanding and so more network architectures need to be examined to keep the compatibility of the proposed method to state-of-the-art standards. Nevertheless, as described in the manuscript, the proposed methodology is unsupervised also in the aspects of network architectures, meaning that evaluating a network architecture not presented in this work is feasible following the methodology described in section III-A. Future work will strive to resolve these limitations.

VI. Conclusion

In this work, we address the issue of bias in Deep Neural Networks (DNNs) for Non-IID Federated Learning Training in the healthcare context, and to a lesser extent in general, in a totally unsupervised data/model agnostic approach. The FL procedure dictates the aggregation of the weights of remotely trained models, producing a global model containing the mutual knowledge of the distributed devices. This technique annexes security to the training process while keeping the data private and the data owners anonymized. Though often, the remote devices do not contain balanced data that sometimes lean towards only one sub-class of that data. This can protrude negatively to the global model, sometimes saturating the whole aggregation process, which in fields like healthcare can have catastrophic results. To tackle this phenomenon, this work undertakes the proposition of an unsupervised ethical equity or fairness methodology to identify defective locally trained models. To this end, a deformation mechanism is employed that stretches the latent knowledge of the model in random shifts along several directions, creating thus the saturation of the learned knowledge along a number of latent axes. These deformations create micro-clusters of the learned information that this work examines with respect to its sparsity and volumetric properties to define the quality of knowledge that the model has obtained in the form of a Fairness factor (F_f). To substantiate the proposed methodology, the pipeline is tested against benchmark datasets and different DL architectures in both Generative and Decision Support architectures. The results of the proposed methodology is simulated and tested on a variety of Federated settings using the benchmark data, showing promise in measuring the fairness or bias of the to-be Federated models. The work proposed in this paper is tested against a variety of benchmark Deep Learning architectures, namely, i) GANs, ii) DNNs, iii) ResNet-18, while also in the FL environment, showing an average 8.75% increase in Federated model accuracy in comparison with similar work.

Future work will involve the testing of more DL models, used for various problems, as well as further experimentation with the hyperparameters involved in the process to produce more solid and robust results aiming at a unified ethical equity exploration for Federated training of DL models.

VII. Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957406 (TERMINET).

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. Fort Lauderdale, FL, USA: PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: http://proceedings.mlr.press/v54/mcmahan17a.html

[2] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*. USENIX Association, Jul. 2021, pp. 19–35. [Online]. Available: https://www.usenix.org/conference/osdi21/presentation/lai

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019.

[4] S. Hwang, S. Park, D. Kim, M. Do, and H. Byun, "Fairfacegan: Fairness-aware facial image-to-image translation," 12 2020.

[5] A. Carey, W. Du, and X. Wu, "Robust personalized federated learning under demographic fairness heterogeneity," 12 2022, pp. 1425–1434.

[6] A. Suresh, F. Yu, H. McMahan, and S. Kumar, "Distributed mean estimation with limited communication," 11 2016.

[7] M. Mohri, G. Sivek, and A. Suresh, "Agnostic federated learning," 01 2019.

[8] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 570–575.

[9] J. Kim and S. Cho, "Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck," *CEUR Workshop Proceedings*, vol. 2560, pp. 105–112, 2020.

[10] M. Hu and J. Li, "Exploring bias in gan-based data augmentation for small samples," *ArXiv*, vol. abs/1905.08495, 2019.

[11] N. T. Tran, V. H. Tran, N. B. Nguyen, T. K. Nguyen, and N. M. Cheung, "On data augmentation for gan training," *IEEE Transactions on Image Processing*, vol. 30, pp. 1882–1897, 2021.

[12] Y. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," 10 2021.

[13] A. Mashhadi, A. Kyllo, and R. Parizi, "Fairness in federated learning for spatial-temporal applications," 01 2022.

[14] I. Siniosoglou, V. Argyriou, S. Bibi, T. Lagkas, and P. Sarigiannidis, "Unsupervised ethical equity evaluation of adversarial federated networks," in *The 16th International Conference on Availability, Reliability and Security*, ser. ARES 2021. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3465481.3470478

[15] D. Pessach and E. Shmueli, "Algorithmic fairness," 2020.

[16] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," 09 2012, pp. 35–50.

[17] J. Forde, A. Cooper, K. Kwegyir-Aggrey, C. Sa, and M. Littman, "Model selection's disparate impact in real-world deep learning applications," 04 2021.

[18] C. Tran, F. Fioretto, and P. Van Hentenryck, "Differentially private and fair deep learning: A lagrangian dual approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9932–9939, 05 2021.

[19] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, 06 2018.

[20] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2017, pp. 1–6.

[21] A. Barredo Arrieta, N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. García, S. Gil-López, D. Molina, V. R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," 10 2019.

[22] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, 10 2017.

[23] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How generative adversarial nets and its variants work: An overview of gan," *ACM Computing Surveys*, vol. 52, 11 2017.

[24] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. McMahan, T. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," 02 2019.

[25] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y. C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[26] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," *ArXiv*, vol. abs/2002.03754, 2020.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[28] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," *arXiv preprint arXiv:2010.14925*, 2020.

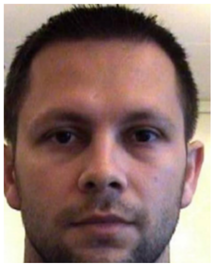
[29] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[30] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[31] S. Arora, W. Hu, and P. K. Kothari, "An analysis of the t-sne algorithm for data visualization," in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 06–09 Jul 2018, pp. 1455–1462. [Online]. Available: http://proceedings.mlr.press/v75/arora18a.html

[32] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar joseph, "A review of dimensionality reduction techniques for efficient computation," *Procedia Computer Science*, vol. 165, pp. 104–111, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

Ilias Siniosoglou received his Diploma degree (5 years) from the Dept. of Electrical and Computer Eng., University of Western Macedonia, Greece, in 2020. He is now a Ph.D.student in the same department. His main area of research is Deep and Federated Learning on Next Generation IoT platforms, primarily focusing on optimization, deployment and scalability methodologies. Currently, he is working as a research associate at the University of Western Macedonia in national and European funded research projects, including (a) H2020-DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid, (b) H2020-SU-DS-2018 (SU-DS04-2018), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm, (c) MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation) and (d) H2020-ICT-2020-1 (ICT-56-2020) TERMINET: nexT gEneRation sMart INterconnectEd ioT.



Vasileios Argyriou (Member, IEEE) received the B.Sc. degree in computer science from the Aristotle University of Thessaloniki, Greece, in 2001, and the M.Sc. and Ph.D. degrees in electrical engineering working on registration from the University of Surrey, in 2003 and 2006, respectively. From 2001 to 2002, he held a research position with Aristotle University, working on image and video watermarking. He joined the Communications and Signal Processing Department, Imperial College, London, in 2007, where he was a Research Fellow working on 3D object reconstruction. He is currently a Professor with Kingston University London, working on computer vision and AI for crowd and human behaviour analysis, computer games, entertainment, and medical applications. Also, research is conducted on educational games and on HCI for augmented and virtual reality (AR/VR) systems.

Prof. Panagiotis Sarigiannidis (Member, IEEE) received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2007, respectively. He is the Director of the ITHACA Lab, the Co-Founder of the 1st spin-off of the University of Western Macedonia: MetaMind Innovations P.C., and an Associate Professor with the Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece. He has published over 260 papers in international journals, conferences, and book chapters, including IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE INTERNET OF THINGS, IEEE TRANSACTIONS ON BROADCASTING, IEEE SYSTEMS JOURNAL, IEEE Wireless Communications Magazine, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, IEEE/OSA JOURNAL OF LIGHTWAVE TECHNOLOGY, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE ACCESS, and Computer Networks. He has been involved in several national, European and international projects. He is currently the Project Coordinator of three H2020 Projects, namely a) H2020- DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid, b) H2020-LC-SC3-EE-2020-1 (LC-SC3-EC-4-2020), EVIDENT: bEhaVioral Insgihts anD Effective eNergy policy acTions, and c) H2020-ICT-2020-1 (ICT-56-2020), TERMINET: nexT gEneRation sMart INterconnectEd ioT, while he coordinates the Operational Program MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation) and the Erasmus+ KA2 ARRANGE-ICT: SmartROOT: Smart faRming innOvatiOn Training. He also serves as a Principal Investigator in the H2020-SU-DS-2018 (SU-DS04-2018), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm and in three Erasmus+ KA2: a) ARRANGE-ICT: pArtneRship foR AddressiNG mEgatrends in ICT, b) JAUNTY: Joint undergAduate coUrses for smart eNergy managemenT sYstems, and c) STRONG: advanced firST RespONders traininG (Cooperation for Innovation and the Exchange of Good Practices). His research interests include telecommunication networks, Internet of Things, and network security. He participates in the editorial boards of various journals.

Thomas Lagkas (Senior Member, IEEE) received the B.Sc. degree (Hons.) and the Ph.D. degree in wireless networks from the Department of Computer Science, Aristotle University of Thessaloniki, in 2002 and 2006, respectively, and the M.B.A. degree from Hellenic Open University, in 2012, and the postgraduate certificate on Teaching and Learning from the University of Sheffield in 2017. He is an Assistant Professor with the Department of Computer Science, International Hellenic University. He has been a Scholar of the Aristotle University Research Committee, as well as a Postdoctoral Scholar of the National Scholarships Institute of Greece. His research interests are in the areas of IoT communications with more than 110 publications at a number of widely recognized international scientific journals and conferences. He is a Fellow of the Higher Education Academy in U.K. Moreover, he actively participates in the preparation, management, and implementation of several EU funded research projects.

Antonios Sarigiannidis Antonios Sarigiannidis received the B.S. and M.S. degrees in computer science from the Department of Informatics, Aristotle University, Thessaloniki, Greece, in 2007 and 2009, respectively. Also, he received the PhD degree from the same department in 2016. His research interests include telecommunication networks, internet of things, cybersecurity, and smart grid. He is author or coauthor of more than 20 journal paper, conference papers, and chapter books. He has been involved in many national and European R&D projects. He received the CCNA Routing and Switching certification from Cisco in 2018. He is the co-founder of Sidroco Holdings Ltd, Nicosia, Cyprus.

Sotirios K. Goudos (Senior Member, IEEE) received the B.Sc. degree in physics, the M.Sc. degree in electronics, and the Ph.D. degree in physics from the Aristotle University of Thessaloniki in 1991, 1994, and 2001, respectively, the master's degree in information systems from the University of Macedonia, Greece, in 2005, and the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2011. He joined the Department of Physics, Aristotle University of Thessaloniki in 2013, where he is currently an Associate Professor. He is the Director of the ELEDIA@AUTH Lab Member with the ELEDIA Research Center Network. He has authored the book Emerging Evolutionary Algorithms for Antennas and Wireless Communications (Institution of Engineering & Technology, 2021). His research interests include antenna and microwave structures design, evolutionary algorithms, machine learning, wireless communications, and semantic Web technologies. He is currently serving as a Chapter/AG Coordinator for IEEE Greece Section. He is the founding Editor-in-Chief of the Telecom (MDPI). He is currently serving as an Associate Editor for IEEE ACCESS and IEEE OPEN JOURNAL OF THE COMMUNICATION SOCIETY. He is also a Member of the Editorial Board of the International Journal of Antennas and Propagation, the EURASIP Journal on Wireless Communications and Networking, and the International Journal on Advances on Intelligent Systems. He is also a member of the topic board of the Electronics, IEICE, the Greek Physics Society, the Technical Chamber of Greece, and the Greek Computer Society.

Shaohua Wan (Senior Member, IEEE) received Ph.D. degree from School of Computer, Wuhan University in 2010. He is currently a full Professor with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China. From 2016 to 2017, he was a visiting professor at the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. His main research interests include deep learning for Internet of Things. He is an author of over 150 peer-reviewed research papers and books, including over 40 IEEE/ACM Transactions papers such as TII, TITS, TOIT, TNSE, TMM, TCSS, TOMM, TETCI, PR, etc., and many top conference papers in the fields of edge intelligence