I. Siniosoglou, V. Argyriou, S. Bibi, T. Lagkas, and P. Sarigiannidis, "Unsupervised Ethical Equity Evaluation of Adversarial Federated Networks," The 16th International Conference on Availability, Reliability and Security. ACM, Aug. 2021. doi: 10.1145/3465481.3470478.

This is the authors' preprint version and not the final published version. The published version is available by ACM at the following link: https://dl.acm.org/doi/10.1145/3465481.3470478

Unsupervised Ethical Equity Evaluation of Adversarial Federated Networks

ILIAS SINIOSOGLOU, University of Western Macedonia, Greece
VASILEIOS ARGYRIOU, Kingston University, United Kingdom
STAMATIA BIBI, University of Western Macedonia, Greece
THOMAS LAGKAS, International Hellenic University, Greece
PANAGIOTIS SARIGIANNIDIS, University of Western Macedonia, Greece

While the technology of Deep Learning (DL) is a powerful tool when properly trained for image analysis and classification applications, some factors for its optimization rely solely on the training data and their environment. In an effort to tackle the problem of knowledge bias created during the training process of a Deep Neural Network (DNN) and specifically Adversarial Networks for image augmentation, this work presents an entirely unsupervised methodology for discovering the unfairness level of Deep Learning (DL) models and in extend, its wrongly accumulated or biased classes. Fdi, the proposed evaluation metric for quantizing the level of unfairness of a model is introduced, along with the method of weighting the model's knowledge and producing its weakest aspects in a data-agnostic way.

1 INTRODUCTION

In the field of image processing and artificial intelligence (AI), the discrimination, unfairness or biased direction of the data, indicates their tendency to be unevenly distributed, leaning in fact towards a specific category subset[12–14, 16, 30]. This implies that, in the case of Machine Learning (ML) or Deep Learning (DL) the performance of the given algorithm will heavily depend on the statistical distribution of the corresponding dataset. Therefore, the fairness, or lack thereof, of a given set of data, intended to be used in ML or DL implementations, will in extend decide the quality of the produced model and its expected outcome, e.g. data classification and augmentation, anomaly detection, recommendation systems, and so on. Furthermore, the same fairness evaluation problem for federated solutions could be applied to other machine learning, cybersecurity and computer vision tasks, such as anomaly detection for UAVs and other systems [3, 11, 20, 21], pose, face and behaviour analysis [1, 2, 4, 10].

In the latest years, and with the introduction of the Generative Adversarial Networks (GANs) [6, 25], DL has skyrocketed to new experimental and implementation heights, finding solutions by utilizing the powerful augmentation

advantages of adversarial approaches. In particular, the GAN architecture, is based on a pair of neural opposing networks, a) the Generator G and b) the Discriminator D, that play and adversarial game with each other [5, 7]. G usually takes a random noise input and tries to mimic the real data of the respective DL problem. On the contrary, D is a network trained to distinguish between the fake data produced by G and the real samples of the given dataset. The GAN architecture aims in training both networks, so that the Generator can produce realistic samples that the Discriminator can't differentiate from the real ones and vice versa. After the training of a GAN network it is common to use the G component to solve the subsequent problem. This methodology relies on the fact that G is trained to produce realistic data and as so establishes an active knowledge-driven baseline to tackle the problem.

As it is slowly becoming apparent, the ethical outcomes or the fairness of the produced model plays a major role in its consequent performance [8, 17, 23, 24, 29]. It is often difficult, though, to evaluate a model's fairness in wide scale applications like in the novel field of Federated Learning (FL), which undertakes the decentralized training of DL models in distributed networks, aimed at data-agnostic training procedures in order to ensure the privacy and security of the remote data [15]. Since in this case the training data and its aspects, like their heterogeneity and distribution, are unknown, the need for a way to measure the fairness of the produced models becomes critical. As of yet, some work on reducing the bias of the data accumulated and in extend generated by adversarial networks has been made. In particular, the authors of [28] present the FairGAN, a fairness-aware GAN architecture that learns to be fair during training. One setback with this solution is that an additional network, i.e. an extra D module, is introduced in the training process, that opposes the disparate accumulation of data by relying on a conditional protected attribute pointing at a certain group in the data. This means that the education of the models has to be further parameterized on the remote end point in the case of FL. The same principle is also explored in [9] where they employ an additional D to discriminate unfair bias towards protected categories and they evaluate their proposal by analyzing the model's bias-variance dilemma to prove its performance against benchmark fairness-oriented datasets. In another work, the authors in [26] present an adversarial representation learning methodology, ensuring the fairness models used by third parties. They, map known fairness evaluation metrics like, a) demographic parity, b) equal opportunity and others to the adversarial training process. They document their method with experimental results that prove the utility of their proposal. This method, though, imposes on the training process of the proposed model, making it again oriented. on the correct parametrization of the model during training in a data-dependent manner.

It is also important to mention the metrics currently used for optimization of fairness problems. They are divided into three main categories, a) Pre-Process, b) In-Process and c) Post-Process, utilizing measures like i) Normalized prejudice index, ii) Disparate impact, iii) Equalized ods, etc., which are based mostly on training set and process oriented solutions. The wide range of the aforementioned metrics and techniques are well documented and described in [18]. The aforementioned metrics cannot be compared with the content and outcomes of this work, since they present an entirely different approach on the fairness problem which relies, not on the model, but on the data and so they are directly incompatible for the comparison.

This work is mainly oriented in classifying and ranking DL models' fairness in an unsupervised manner by comparing pre-trained models in a distributed fashion. Specifically, this paper presents a fairness measuring mechanism that relies only on the DL models, without any prior interaction and knowledge of the training procedure or the data used in the aforementioned process. The produced methodology uses minimal parametrization and relies only on the given pre-trained models. This work aims in producing the means to measure DL models rank of unfairness and present a description of the unfair representations in the given model, i.e. the classes that were not properly accumulated. The

presented paradigms are evaluated on benchmark datasets in order to facilitate their general use. In retrospect, the contributions of this paper can be summarized as follows:

- *Producing an unsupervised data-agnostic methodology for fairness characterization of DL models:* The produced methodology was developed in respect to systems that do not have access to the training data or procedure of a given model and must evaluate its fairness for subsequent use, as in the case of federated architectures.
- *Introducing a Fairness evaluation metric:* An evaluation metric is developed in order to help characterize the fairness of DL models in data-agnostic meta-use.

The rest of this paper is organised as follows. Section 2 explains the methodology used for the solution of the fairness discovery problem, while Section 3 presents the evaluation of the introduced methodology. Finally, Section 4 concludes this work.

2 METHODOLOGY

The proposed methodology assumes that several pre-trained adversarial models are available and aims to evaluate their fairness for a classification problem in a federated system. The available models correspond to the ones provided by the workers in the distributed architecture. As it is shown in Figure 1 the suggested approach follows four main axis, namely, a) Data Clustering, b) Class Ranking , c) Biased Class Identification and d) Fairness Factor Calculation.

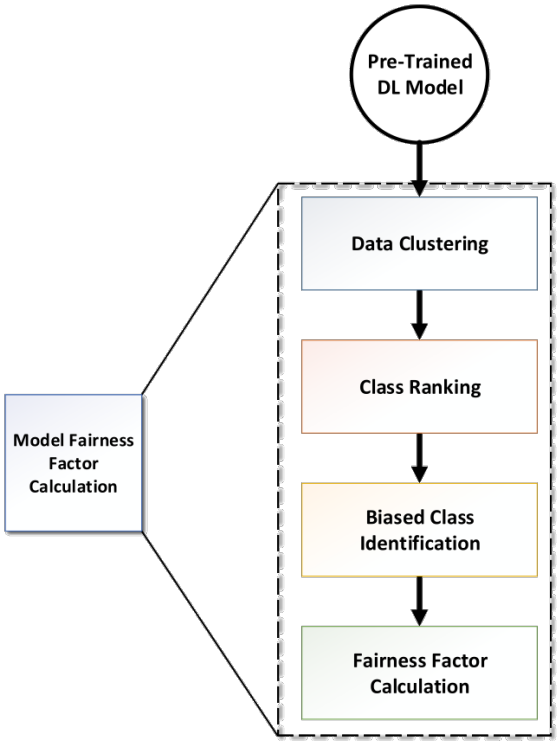

Fig. 1. Fairness Factor Calculation Pipeline

2.1 Data Clustering

For the purpose of effectively measuring the distribution of knowledge of the given DL models, the need for data on which analysis will be performed, occurs. Since, it is assumed, that there is no access to the training data, the evaluation data are augmented by the models in two main data augmentation techniques. This is done to establish a knowledge baseline and because of the intended use of this method (distributed learning).

In addition to the data generated by the given under evaluation DL models G_i from latent representations, we also deform the latent space on random directions [27] that correspond to semantic manipulations (shifts). In order to perform these shifts and control the generated data, two components are trained for each model G_i, first a matrix AER^{dxK}, where d is the dimensionality of the latent space of G_i and K corresponds to the number of directions in the latent space. Then a reconstructor R which obtains an image pair $(G_i, G(z + A(ae_k)))$ and outputs $R(I_1, I_2) = (k', a')$. 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 is perfomed using the a minimisation process on the following loss function:

$$\min_{A,R} \mathop{E}_{z,k,a} L(A, R) = \min_{A,R} \mathop{E}_{z,k,a} [L_{cl}(k, k') + cL_r(a, a')] \tag{1}$$

where $L_{cl}$is the cross-entropy and L_r is the mean absolute error, while we selected experimentally $c = 0.25$.

Following the data generation process, the obtained data I_i^D of the models G_i are aggregated and normalized to a given range. Next, a dimentionality reduction algorithm (e.g. tSNE) is applied transfering the data into a space I_i^d, with $d << D$ and the best size/information loss trade-off. Both the normalization and dimentionality reduction are data-dependent so no threshold can be extrapolated for a wide variety of produced data. The samples are also transformed to a one-hot encoded vector per sample. Then the data are fed to a K-means clustering instance. Before the clustering, though, the data are shuffled sufficiently to reduce the clusterer's bias. A separate index named $dset$ keeps track of the correspondence between the shuffled samples and the dataset they came from. The number of clusters can either be selected manually or be found using the elbow method. In this work, the Kneedle algorithm [22] was leveraged to dynamically find the number of clusters C for the clustering process.

2.2 Class Ranking

Subsequent to the fitting of the K-means algorithm with the data, the produced labels are correlated with the original data. To finalize the analysis of the models, two more parameters are computed, a) the population matrix and b) the mean standard deviation distance matrix of the samples. In order to calculate the population matrix Pm of the samples per cluster of each dataset, each sample is assigned to the matrix following the intersection of the thee arrays. Then Pm_m is simply the amount of elements in each Pm_i^j where i signifies the dataset index and j denotes the label index of the samples. The mean standard deviation distance matrix Dm of the distances between the samples of a cluster is also calculated. The class ranking is produced by (2):

$$R_i = argsort(Pm_i) \tag{2}$$

where R is the rank matrix of the data and R_i denotes the rank vector of i^{th} model based on the population of each cluster in matrix Pm.

2.3 Biased Class Identification

Since the classes have been identified and ranked, the threshold on which the biased classes end, and the rightly accumulated classed begin, must be calculated. This is achieved by the following process for each R_i vector. First, the vector is logarithmically normalized. Subsequently, it is smoothed using a Savitzky-Golay Filter. This filter was chosen because of the unpredictable nature of the population vector, i.e. the uneven gaps between populations that produce noise that will obscure the calculation of the threshold point. Equation (3) describes the smoothing process of the population vector.

$$R_{smoothed_i^j} = \frac{1}{h}\left[\sum_{k=-\frac{p_i-1}{2}}^{\frac{p_i-1}{2}} a_k log(R_{i_{j+1}})\right] \tag{3}$$

Here $R^i_{smoothed}$ is the denoised i^{th} population vector point, h is the normalization factor, p_i denotes the number of points of the vector, a_k is the smoothing coefficient and $R_{i_{j+1}}$ is the corresponding j^{th} element of vector R_i. Then, the Kneedle algorithm is again employed to find the saturation point sp_{kneed} of $R^i_{smoothed}$, while a relative saturation point sp_{rel} is calculated by finding the first local maxima of $R^i_{smoothed}$. The absolute saturation point sp_{abs} is computed using the following formula (4):

$$sp^i_{abs} = \frac{sp^i_{kneed} + sp^i_{rel}}{2} - 1 \tag{4}$$

where $sp^i_{kneed} = Kneedle(R^i_{smoothed})$ and $sp^i_{rel} = \left|\nabla^2(R^i_{smoothed})\right|_0$ signifying the first peak found in $R_{smoothed^i}$. Since sp^i_{abs} is the index of the saturation point of the sorted population vector and also denotes the number of certain biased clusters. In the Fig. 2 , blue shows the sp^i_{kneed}, red shows the sp^i_{rel} and green shows the sp^i_{abs} class predictions.

2.4 Fairness divergence indicator (Fdi) Calculation

Finally, the last step is to calculate the Fairness factor of each model. This was done by first finding the difference between the tangent of the curve of biased clusters $atan_b$ and the tangent of curve the rest of the clusters $atan_r$ in respect to the x-axis, i.e. a line with slope 0. In Fig 2 the inclining curve of points depicts the increase of the population of each i^{th} cluster j in matrix Pm. The more balanced the model and the distribution it produced, in respect to the other models, the more the angle of tangent of the curve would tend to zero itself. Thus, the bigger the incline, the more unbalanced a model is. By subtracting $atan_b$ which produces a factor of divergence from the rest of the data and $atan_r$ which shows the factor of the dataset's tendency to be balanced, we achieve a measurable divergence indicator di. The final Fairness factor is deducted by multiplying the divergence indicator with the percentage of biased clusters and the mean standard deviation of the samples in the sp_{abs} point's cluster. Equation (6) presents the Fairness factor calculation formula. Let

$$atan_b = arctan(\triangledown\left(\sum_{k=1}^{sp_{abs}} R^k_{smoothed}/sp_{abs}\right))$$

and

$$atan_r = arctan(\triangledown\left(\sum_{k=sp_{abs}+1}^{p} R^k_{smoothed}/(sp_{abs}+1-p)\right))$$

where p is the number of points in $R_{smoothed}$. If the divergence indicator is

$$di = |atan_b - atan_r| \tag{5}$$

Then,

$$F_f = log(\frac{sp_{abs}}{p} * di * Dm_j) \tag{6}$$

Here, F_f formulates the Fairness factor and Dm_j denotes the mean distance between the samples of the j^{th} cluster produced by the model. Finally, it was found that the higher value of the $|Fdi|$ the fairest the model is.

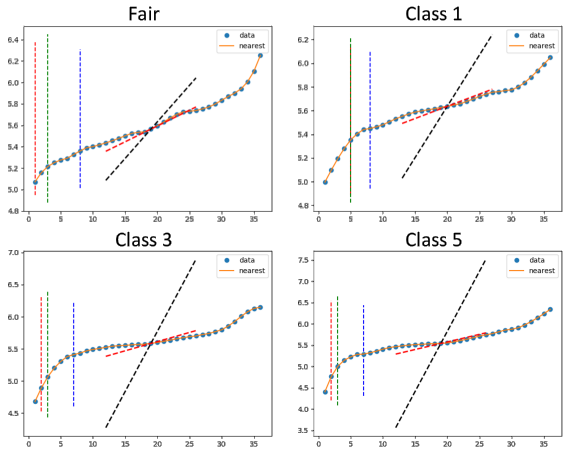

Fig. 2. Population curves predicted on Digit MNIST.

3 EVALUATION

In order to evaluate the proposed methodology and Fairness evaluation metric, two benchmark datasets were utilized, namely, the i) Digit MNIST and ii) Fashion MNIST datasets, introducing morphologically different benchmark image data. The data were chosen based on the fact that they do not contain any natural biases than can affect the validation of the presented methodology, like, outliers, elements that diverge from the goal of the dataset like color augmentation and others. The experiments were run on a Linux workstation consisting of 16Gb RAM memory, an i7 intel core processor and an NVIDIA GTX 1080 8Gb GPU. For the sake of the experiment four DCGAN networks [19] were trained on four different distributions of each dataset. Specifically, for each dataset, one model was trained on all of the classes equally, and the others were trained on 10%-20% of the corresponding 1st, 3rd and 5th classes, respectively. Each model was made to generate 10000 samples using the simple Self Data Generation technique but also using latent direction shifts produced by the second method, Augmented Generation based on Latent Deformation Random Directions. This resulted in four aggregated generated datasets, the i) Simple Digit MNIST, the ii) Simple Fashion MNIST , the iii) Augmented Digit MNIST and the iv) Augmented Fashion MNIST. Attributes used for the training and data generation are described in Table 1. It is clear that the training of the models was sub-optimal, which was preferred to simulate the uncertainly of a model of unknown training environment. Table 2 summarizes the results of the proposed method on the different datasets on different configurations, which are also described in Table1.

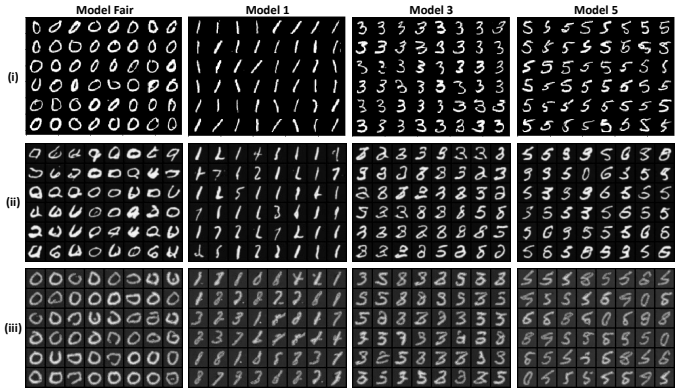

Fig. 3. Unfair Classes , i) original, ii) predicted on Digit MNIST , iii) predicted on Augmented Digit MNIST

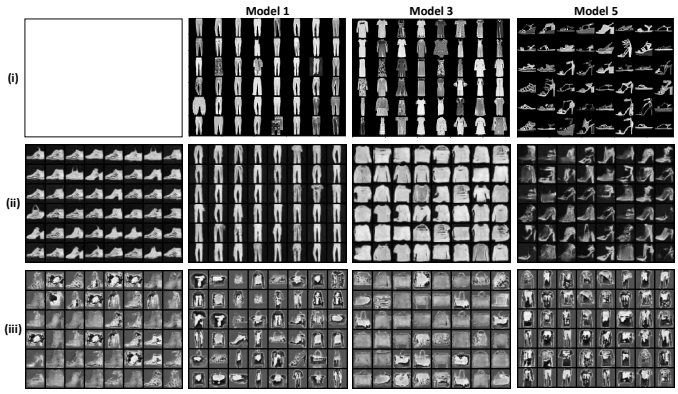

Fig. 4. Unfair Classes , i) original, ii) predicted on Fashion MNIST , iii) predicted on Augmented Fashion MNIST

Dataset	n_clusters	epochs	Latent Dimension	Batch Size
Digit MNIST	30	30	64	128
Fashion MNIST	25	35	64	128
Augmented Digit MNIST	26	10000	64	16
Augmented Fashion MNIST	27	1000	64	16

Table 1. Scenario Configuration

As can be seen in Fig 3 and 4 the Fdi method succeeds in finding the most unfair classes. For the sake of the results, only the most unfair class is depicted. In the first row of the two figures, the classes of the original dataset are depicted and on the rest the predicted most unfair class. As was mentioned, three of the four models were trained on a small percentage of a certain class of the original dataset. In this case, the Digit MNIST GANs were trained to be biased on the numbers 1, 3 and 5, while the Fashion MNIST models are biased towards the trousers, dress and sandals (classes 1,3,5) respectively. For the digits, the method produces the exact match of the biased classes. A note has to be made that the method found that class 0 of the digits is unfair. This is true because the proposed method produces the relative

Dataset	Model	Unfair Class	di	Fdi
Digit MNIST	Fair	3	0.00041	-12.9
	Unfair c1	5	0.00074	-10.6
	Unfair c3	3	0.00382	-9.7
	Unfair c5	3	0.00947	-8.7
Digit MNIST tSNE	Fair	3	1.66e-05	-14.18
	Unfair c1	5	0.0025	-8.85
	Unfair c3	7	0.00028	-10.9
	Unfair c5	2	0.015	-7.73
Fashion MNIST	Fair	2	0.0171	-8.56
	Unfair c1	3	0.0064	-8.72
	Unfair c3	2	0.0031	-10.33
	Unfair c5	2	0.0221	-8.25
Fashion MNIST tSNE	Fair	3	0.0011	-9.27
	Unfair c1	4	0.0089	-7.8
	Unfair c3	4	0.0037	-8.48
	Unfair c5	2	0.0155	-8.1

Table 2. Scenario Results

biased classes to the rest of the model. But in the case of the digits the same class was found to be biased in all the experiments of both the simple and additionally augmented data. This reveals that the model trained equally on all classes is in fact, though involuntarily, unfair towards 0. In contradiction, the Fdi indicates that this model is the less biased, as it has the bigger absolute value among the other models. Furthermore, as it can be seen in Table 2 and from the Fashion MNIST predicted unfair classes, the Augmented Fashin MNIST data were not suitable for this process due to its high saturation during augmentation. The method still did find a large portion of the biased samples.

4 CONCLUSION

The technology of Federated Learning is steadily but surely intruding in the modern distributed Deep Learning deployments. Its powerful effect on trained model in parallel to its ability to preserve the privacy of the data and data owners while training models with distributed knowledge makes this technique a major stepping stone for the modernization and optimization of contemporary smart systems. Though a lot of work has been performed in equalizing the fairness of the models through detecting the balance of the training data or trying to balance the model mid-training, little has been done to recognize unfair model in the FL environment moreover so in an unsupervised manner. To tackle this straggle, the presented work produced an novel unsupervised methodology for detecting the level of unfairness of a DL model and finds the weakest learned classes. Fdi is introduced, which is an evaluation metric for quantizing the level of unfairness of a model along with the method of weighting the model's knowledge and producing its weakest

learned points in a data-agnostic way. The results reveal that the proposed methodology can successfully identify a big portion of the unbalanced classes and present a way to measure that unfairness.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Vasileios Argyriou and Maria Petrou. 2009. Chapter 1 Photometric Stereo: An Overview. In *Advances in IMAGING AND ELECTRON PHYSICS*, Peter W. Hawkes (Ed.). Advances in Imaging and Electron Physics, Vol. 156. Elsevier, 1–54. https://doi.org/10.1016/S1076-5670(08)01401-8

[2] Vasileios Argyriou, Maria Petrou, and Svetlana Barsky. 2010. Photometric stereo with an arbitrary number of illuminants. *Computer Vision and Image Understanding* 114, 8 (2010), 887–900. https://doi.org/10.1016/j.cviu.2010.05.002

[3] Paolo Bellavista, Carlo Giannelli, Thomas Lagkas, and Panagiotis Sarigiannidis. 2018. Quality Management of Surveillance Multimedia Streams Via Federated SDN Controllers in Fiwi-Iot Integrated Deployment Environments. *IEEE Access* 6 (2018), 21324–21341. https://doi.org/10.1109/ACCESS.2018.2822401

[4] Pierre Bour, Emile Cribelier, and Vasileios Argyriou. 2019. Chapter 14 - Crowd behavior analysis from fixed and moving cameras. In *Multimodal Behavior Analysis in the Wild*, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe (Eds.). Academic Press, 289–322. https://doi.org/10.1016/B978-0-12-814601-9.00023-7

[5] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil Bharath. 2017. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* 35 (10 2017). https://doi.org/10.1109/MSP.2017.2765202

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (06 2014).

[7] Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon. 2017. How Generative Adversarial Nets and its variants Work: An Overview of GAN. *Comput. Surveys* 52 (11 2017). https://doi.org/10.1145/3301282

[8] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. 2020. FairFaceGAN: Fairness-aware Facial Image-to-Image Translation.

[9] Jin Young Kim and Sung Bae Cho. 2020. Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck. *CEUR Workshop Proceedings* 2560 (2020), 105–112.

[10] Dimitrios Konstantinidis, Vasileios Argyriou, Tania Stathaki, and Nikolaos Grammalidis. 2020. A modular CNN-based building detector for remote sensing images. *Computer Networks* 168 (2020), 107034. https://doi.org/10.1016/j.comnet.2019.107034

[11] Thomas Lagkas, Vasileios Argyriou, Stamatia Bibi, and Panagiotis Sarigiannidis. 2018. UAV IoT Framework Views and Challenges: Towards Protecting Drones as "Things". *Sensors* 18, 11 (2018). https://doi.org/10.3390/s18114015

[12] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[13] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByexElSYDr

[14] Lingjuan Lyu, Xinyi Xu, and Qian Wang. 2020. Collaborative Fairness in Federated Learning. arXiv:2008.12161 [cs.LG]

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data *(Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, Fort Lauderdale, FL, USA, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[17] Alejandro Pena, Ignacio Serna, Aythami Moralinproceedingses, and Julian Fierrez. 2020. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[18] Dana Pessach and Erez Shmueli. 2020. Algorithmic Fairness. arXiv:2001.09784 [cs.CY]

[19] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 [cs.LG]

[20] Panagiotis Sarigiannidis, Eirini Karapistoli, and Anastasios Economides. 2017. Modeling the Internet of Things Under Attack: A G-network Approach. *IEEE Internet of Things Journal* 4, 6 (2017), 1964–1977. https://doi.org/10.1109/JIOT.2017.2719623

[21] Panagiotis Sarigiannidis, Eirini Karapistoli, and Anastasios A. Economides. 2015. VisIoT: A threat visualisation tool for IoT systems security. In *2015 IEEE International Conference on Communication Workshop (ICCW)*. 2633–2638. https://doi.org/10.1109/ICCW.2015.7247576

[22] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. 2011. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *2011 31st ICDCS-W*. 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[23] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63, 4/5 (2019), 3:1–3:9. https://doi.org/10.1147/JRD.2019.2945519

[24] Shubham Sharma, Yunfeng Zhang, Jesús Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Ramazon Kush. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. 358–364. https://doi.org/10.1145/3375627.3375865

[25] Ilias Siniosoglou, Panagiotis Sarigiannidis, Vasileios Argyriou, Thomas Lagkas, Sotirios Goudos, and Maria Poveda. 2021. Federated Intrusion Detection In NG-IoT Healthcare Systems: An Adversarial Approach.

[26] N. T. Tran, V. H. Tran, N. B. Nguyen, T. K. Nguyen, and N. M. Cheung. 2021. On Data Augmentation for GAN Training. *IEEE Transactions on Image Processing* 30 (2021), 1882–1897. https://doi.org/10.1109/TIP.2021.3049346

[27] A. Voynov and A. Babenko. 2020. Unsupervised Discovery of Interpretable Directions in the GAN Latent Space. *ArXiv* abs/2002.03754 (2020).

[28] D. Xu, S. Yuan, L. Zhang, and X. Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. In *2018 IEEE International Conference on Big Data (Big Data)*. 570–575. https://doi.org/10.1109/BigData.2018.8622525

[29] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. 2020. Inclusive GAN: Improving Data and Minority Coverage in Generative Models. arXiv:2004.03355 [cs.CV]

[30] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. 2020. Hierarchically Fair Federated Learning. arXiv:2004.10386 [cs.LG]