2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

A Survey on Bias Mitigation in Federated Learning

Bassey Ude, Olusola T. Odeyomi, Kaushik Roy, and Xiaohong Yuan
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, NC, USA
bassey.ude88@gmail.com, otodeyomi@ncat.edu, kroy@ncat.edu, and xhyuan@ncat.edu

***Abstract*—Federated learning (FL) enables collaborative model training while keeping data decentralized. However, system heterogeneity and statistical differences in decentralized data can introduce biases and unfairness. This paper surveys existing bias mitigation techniques in FL across various phases of the training process. We identify sources of bias and present a critical analysis of current fairness-aware FL algorithms, categorizing them as preventive (Pre-processing) or reactive (in-processing and Post-processing) based on when bias mitigation is applied. In addition, this paper reveals open challenges in balancing fairness and efficiency in FL, handling non-independent and identically distributed (non-IID) data, and ensuring privacy. This survey lays out the foundation for developing unbiased and privacy-preserving FL systems without discrimination in the future.**

***Index Terms*—Federated Learning, system heterogeneity, Bias mitigation, decentralized data, fairness.**

I. Introduction

Machine learning (ML) has experienced a rapid evolutionary trajectory since its inception, marked by the emergence of numerous methodologies uniquely suited to various applications and contexts. Today, these technologies have become inextricably woven into the fabric of our daily existence, seamlessly integrating with innumerable aspects of our lives. Consequently, an immense volume of data is constantly being generated, offering scores of opportunities for developing and implementing intelligent-driven learning services like disease diagnosis systems, search engines, and recommendation systems [1], which are significantly enhancing the decision-making processes [2]. To effectively utilize this voluminous data, it must be consolidated within a centralized infrastructure for training machine learning models. The process often results in substantial computational and storage costs while posing notable data privacy and security risks, necessitating sensitive data protection. New legal frameworks, such as the General Data Protection Regulation (GDPR), and the California Consumer Privacy Act (CCPA) brought about a paradigm shift in data storage, management, and usage [2], [3]. These regulations unequivocally mandate that data must be retained within their origin of jurisdiction, inevitably creating isolated data silos. Federated learning (FL) sprang up as a significant advancement in addressing the limitations of the traditional centralized machine learning models. This privacy-preserving distributed conceptual breakthrough offers a promising solution to the critical challenges associated with data privacy and model ownership in machine learning [4]–[6]. Over the years, many federated platforms have been developed and made available through open-source licenses. Such media include the Federated AI Technology Enabler (FATE) by WeBank, TensorFlow Federated (TFF) by Google and lots more. Industries like healthcare, telecommunications, finance, education, and urban computing are at the forefront of adopting and leveraging FL on a large scale, as evidenced by recent advancements and innovative implementations [2]. The FL ecosystem is shown in Figure 1.

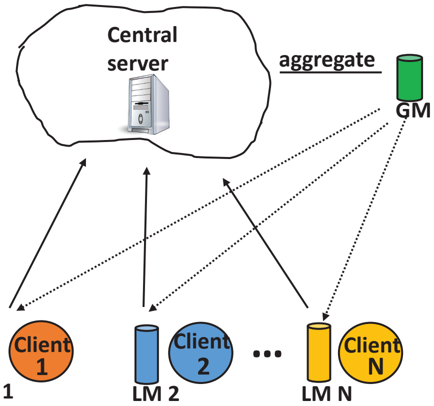

Fig. 1. The federated learning ecosystem, where LM represents the local model, GM represents the global model, and $1, \dots, N$ represents the clients.

Despite its advantages, FL exacerbates the problem of bias, through system heterogeneity, statistical heterogeneity, client selection, incentive mechanism design, and communication efficiency [7]. Bias is a complex and multifaceted issue that emerges at many stages of FL. When client selection protocols hinge on some characteristics, such as the device's computational capabilities [8], clients with more considerable assets are likely to have greater representation. Interestingly, bias can produce a pattern where client participation correlates with their socioeconomic status, racial background, or sexual orientation. For instance, let us say a bank is trying to ascertain a customer's ability to repay a loan before approving it. The system may rely on sensitive variables like race and gender or even skin color to learn predictions that may not be accurate.

Many fairness-conscious FL methods have been suggested from different viewpoints. Fairness in FL refers to the equitable treatment of various stakeholders in the FL process,

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

such as servers, the clients (i.e. data owners), and the final users of the FL model [9]. Several variables can influence fairness, such as client selection, model optimization, contribution evaluation, and incentive distribution [10]. Fairness-conscious FL methods strive for diverse fairness concepts, including performance distribution fairness, good-intent fairness, group fairness, selection fairness, contribution fairness, regret distribution fairness, and expectation fairness [11]. Nonetheless, most of these techniques prioritize achieving fairness objectives without considering their effect on FL's efficiency [12]. To attain a fair distribution of performance, choosing a more significant number of clients or raising the number of local epochs might be necessary, which in turn increases communication and computation expenses [13]. Also, attaining contribution fairness might necessitate assessing each client's influence on the overall model performance, potentially leading to increased convergence cost [14]. Adjusting or applying regularization to the loss function based on sensitive characteristics may be necessary to accomplish group fairness [15], which could decrease model precision [16]. As a result, it is imperative to ensure fairness in FL that guarantees equitable treatment and a high-quality global model [17].

Privacy issue introduces another dimension to addressing FL bias. Since safeguarding data privacy constitutes a primary driving force behind adopting FL, it becomes an essential factor impacting the development and implementation of bias detection and mitigation strategies. Therefore, efforts to rectify biases within FL must account for the constraints imposed by privacy requirements and explore solutions capable of addressing bias while adhering to these limitations.

This paper provides an extensive overview of bias mitigation techniques in FL. The main highlights of this review are outlined as:

- We give an in-depth analysis of how different sources of bias propagate through the federated learning pipeline and influence the model training.
- We present a critical analysis of current Fairness-Aware FL techniques and classify them based on their concepts of fairness.
- We also identify challenges and research directions in this area.

The subsequent sections in this paper are organized as follows: Section II discusses the sources of bias in federated learning. Section III outlines mitigation techniques currently deployed across various phases of the training process. Section IV concludes the paper and section V provides future research directions.

II. Bias Factors in Federated Learning Systems

Although bias may come from human sources, it presents technical issues. Owing to the inherent nature of FL, it presents unique challenges in achieving bias-free machine learning models. The paramount source of bias is attributed to the non-Independent and Identically Distributed (non-IID) feature of the distribution over the collaborating devices and becomes reflected in the parameters of the global model [18].

Let the local dataset at client k be represented as $D_k \sim P_k(x, y)$. With non-IID data, $P_k(x, y) \neq P_j(x, y)$ for $k \neq j$; hence, the data distribution varies across different nodes. Consequently, certain classes or outcomes y may be over-represented at some clients relative to others, thereby introducing skew in the global model's learning curve [19]. One example is developing a machine learning model to identify diseases using patient data from Hospitals A and B. Hospital A in a particular region might have a specialty in heart diseases, thus having more data related to many kinds of heart diseases. Conversely, Hospital B might specialize in cancer research, resulting in a richer dataset on diverse cancer types. The varying disease data distributions across these hospitals illustrate a non-iid scenario in FL, as the data is not distributed both independently and identically. Clients that have larger datasets often influence the learning process due to the variations in data quantities and quality across various nodes (hospitals in our case) [20], [21]. During the FL training process, the local model (LM) updates computed by each client may be skewed, and this distributional imbalance will be mirrored in the updates to the global model (GM). This can mathematically be represented as:

$$\Delta\theta_{\text{skewed}} = \eta \cdot \sum_k \gamma_k \cdot \frac{n_k}{n} \Delta\theta_k, \tag{1}$$

where $\Delta\theta_{\text{skewed}}$ is the skewed update to the global model parameters; η denotes the learning rate (a scalar that determines the step size in the direction of the gradient. It essentially scales the magnitude of the model update); $\sum_k$ is summing over all local datasets indexed by k. Each client's contribution to the global update is not just determined by the sheer volume of its data represented by $\frac{n_k}{n}$, but also by a skewness factor, $\gamma_k\cdot$. The skewness factor measures the extent to which a client's data distribution deviates from the overall data distribution. A value of $\gamma_k = 1$ proposes a typical distribution, while γ_k could deviate from 1, signifying skewness. $\Delta\theta_k$ is local model update parameter for the data D_k on client k. Consequently, the adjustments to the global model embody both the uneven distributions of data across clients and their respective data volumes. This nuanced understanding guarantees that the model continues to respond to the varying and potentially biased data environments within federated networks.

Systemic heterogeneity can also intensify bias. Different nodes (clients) in the system may have different statistical properties, computational capabilities, or other attributes that influence their capacity to contribute to the learning process [22]. Due to the remote nature of the ecosystem, clients with superior computational power or network connection shoulder a more significant contribution to the global model than the resource-constrained clients [23]. Mathematically, this manifests in the federated averaging as:

$$\Delta\theta = \eta \cdot \sum_k \left(\frac{C_k V_k A_k}{C_{tot} V_{tot} A_{tot}} \right) \Delta\theta_k, \tag{2}$$

where $\Delta\theta$ represents the change in the global parameters of the model; η is the learning rate; C_k, V_k, and A_k are the computational capacity, data volume, and availability of node k, respectively; C_{tot}, V_{tot}, and A_{tot} are the aggregate computational capability, data volume, and availability of all nodes. $\Delta\theta_k$ remains the local model update evaluated on the local data D_k at node k. $(\frac{C_k V_k A_k}{C_{tot} V_{tot} A_{tot}})$ represents the relative computational capability, data volume, and availability of each node. This biases the global model toward the distribution of clients with ample resources.

Another source of bias is the fusion algorithm used by the central server to aggregate the weighted average. The algorithm's design may vary, which invariably influences the final model. Some algorithms are designed to equally incorporate model updates from clients, while some perform weighted averages based on client size (clients with larger datasets influence the global model than those with smaller datasets) and can over-represent subsets of clients:

$$\Delta\theta = \eta \cdot F(w_k, \Delta\theta_k) + \beta \cdot B(F, D_k), \tag{3}$$

where F represents the fusion algorithm deployed by the central server. The term w_k denotes the weights associated with each client k, potentially based on criteria like data size. Critically, $B(F, D_k)$ captures the bias introduced by the fusion algorithm for the data at client k. This bias, adjusted by the factor β, accentuates the discrepancies between the ideal and actual aggregations by the fusion algorithm, This underscores the crucial part that the aggregation technique plays in potentially introducing biases to the global model update.

Data class imbalance may prejudice learning algorithms toward the predominant class, resulting in a bias toward the minority class. This becomes difficult to make sound judgments in situations such as fraud detection and medical diagnosis [24]. In FL, there are two types of data imbalance: *local imbalance* and *global imbalance*. The Local imbalance occurs when the class distribution of each client's dataset is misrepresented. One client might possess numerous Class A samples and limited Class B samples, whereas a different client could exhibit the reverse distribution [25]. This mismatch could lead to performance issues in scenarios involving FL where privacy requirements prohibit data redistribution. There is a global imbalance where all consumers in the federation favor particular classes. Even if each client's local dataset appears to be balanced, the preponderance of certain classes in the datasets of multiple clients may cause this skew [26]. To maintain fairness, FL must investigate data imbalance solutions.

III. Bias Mitigation Techniques

Bias can spring up throughout the FL training process. A number of authors have carried out studies on bias-mitigating techniques over the years. Their main goal ensures that sensitive attributes like gender or race are not the determining factor for the outcome of an FL model. According to Dr. Jennifer Chayes, the former managing director of Microsoft Research, "It is possible to design algorithms that are more fair than conventional human decision-makers. This is achieved despite the obstacles presented by training data with inherent bias" [27]. Depending on their application within a machine learning pipeline [28], the numerous FL bias mitigation techniques can vary as pre-processing, in-processing, and post-processing. These are further classified as **Preventive Techniques** (pre-processing), which seek to avert biased models from the training data; and **Reactive Techniques**, that uncover and correct bias when detected during and or after the training (i.e. in-processing and post-processing).

A. Pre-processing bias mitigation techniques

Bias mitigation of pre-processing begins with the data to be trained because the way data is used for training the learner decides the model's outcome. It involves analytically manipulating the data like imputing missing values, selecting extrapolative variables and aggregation [29]. Abay et. al. [30] explored the case of imbalance on sensitive features such as gender and race; particularly, where client datasets showed a non-uniform distribution of these features. They propose a unique pre-processing technique labeled *Reweighing*. Reweighing is a pivotal instrument for constructing fair machine-learning models by adjusting and assigning the weights of instances in the training set prior to training [31]. With this method, we gain access to the whole training dataset and calculate the weights on the ratio between predicted probability (P_{pre}) and the observed probability (P_{obs}) of the sample's sensitive features. Given the data privacy restrictions inherent to FL, which preclude direct access to the data, the authors proposed two modifications to the reweighing method that are applicable to FL settings: *Local reweighing*, where each client computes reweighing weights locally on its own dataset and uses them for its local training during the pre-processing. Here, there is no communication between clients and the aggregator to reveal the sensitive features and or data sample details [32]. Even when only a portion of the participating clients use it, this technique addresses bias effectively without affecting prediction accuracy. The authors also propose a differentially private global variation of the local reweighing called *global reweighing*, which relies on each client's willingness to share details of their sensitive features and their noisy sample counts with the aggregator and other participants. The differential privacy noise is introduced via a well-known privacy mechanism [33] and by adjusting the quantity of noise being injected, clients are able to regulate their data leakages toward bias mitigation. However, one major limitation of this technique is that it is not suitable for FL with dynamic participation settings. It recalibrates the global reweighing weights as the number and size of training sets change in the course of training which invariably reduces the accuracy of the global weights. Zhu, Y et al. [34] propose a novel method called M^{3}Fair to address this limitation. This technique extends the current reweighing technique to account for intersectional bias and to strike a balance between machine learning performance and fairness. They defined sensitivity levels as the sum of weights for numerous sensitive features within a sample and assign level weights to each sensitive feature. The weight coefficients for samples according to their sensitivity levels and labels are then cumulated and applied to the loss function of the models. In addition, the authors present a technique for identifying sensitive characteristics that consistently show tendencies of bias across four evaluation metrics namely: Disparate Impact (DI), Statistical Parity Difference (SPD), Average Odds Difference (AOD), and Equal Opportunity Difference (EOD) [35]. Using this technique, they select the sensitive feature for reweighing and assign level weights to them based on their bias tendencies. The M^{3}Fair technique was then applied to three datasets: Adult, Tongji Hospital COVID-19 dataset (TJH), and HM Hospitales Covid Data Save Lives (CDSL) [36], and its accuracy was compared to that of no mitigation and single-feature reweighing. Their findings indicate that M^{3}Fair achieved superior or equivalent fairness across all the metrics. It also exhibited encouraging results and could be applied in various domains, although there was a minor decrease in model performance. However, this method of bias detection relies on binarizing each feature by comparing its mean value, which may fail to convey the nuances of continuous or categorical features. More so, manually assigning level weights for sensitive features based on bias may add subjectivity and arbitrariness since M^{3}Fair solely examines dataset bias, not model or post-processing bias.

B. *In-Processing Bias Mitigation techniques*

In-process bias mitigation is performed at the algorithmic level, where learning algorithms are modified to bring about fairness. This technique comes to play in order to alleviate the limitations of the pre-processing techniques. The fundamental classifier's optimization problem is altered by introducing a discrimination-aware regularizer or bias mitigation constraints [37]. The primary goal of in-processing bias mitigation is to ensure systems are both fair and accurate. But this mitigation approach may be limited to particular machine-learning models and learning algorithms. An example is [38], where their proposed technique focuses on logistic regression models. Abay et. al. [30], proposed *Federated Prejudice Remover*, a technique that adds fairness-aware regularizer to the loss function. Every client in this ecosystem implements the prejudice remover algorithm [38] to perform the training of a less biased local model. The ultimate goal is to minimize the mutual information between the sensitive features and the predicted outcome. Each client computes the local gradient associated with the regularizer and sends it to the server, which in-turn aggregates and updates the global model accordingly [39]. Only the gradients of the regularizer are exchanged which preserves privacy. Howbeit, the major challenge is the selection of a reasonable coefficient for the regularizer which leads to degradation in accuracy and performance metrics. Another ubiquitous mechanism is the implementation of adversarial learning for bias mitigation across different applications [40] [41] [37]. This technique trains a separate adversarial network to predict protected demographic details based on biased labels. The adversarial learning then enables the fairness-oriented model to dissociate protected data from possible biases. Goodfellow et al. [42] proposed the Generative Adversarial Network (GAN) structure. Devised for generating images, the technique leverages multiple rival networks for model training that would delude one another. The adversarial network acts as a discriminator in a typical GAN setting. The fair network strives to reduce the probability of the discriminator predicting the protected feature accurately based on the model's output while concurrently preserving its own accuracy. Hence, adversarial learning contributes to reducing the influence that a protected feature may have on the model's predictions; thereby decreasing the inherent bias associated with the feature in the model's predictions. The strength of this technique is its vast applicability across divergent datasets and use cases. It also boasts accuracy as it upholds data integrity. Furthermore, it does not require presumptions regarding the dataset's distribution. Nevertheless, it demands access to the model parameters, rendering it impractical in scenarios with black-box models.

C. *Post-Processing Bias Mitigation techniques*

The post-processing approach is a set of techniques applied after the global model has been aggregated, but now the aim is to achieve balanced results. This adjustment to the model predictions is made under a specified fairness constraint. Like pre-processing, one core advantage of this technique is its ability to operate without access to the model specifications - adjusting only the results in place of the classifier or training data [43]. This characteristic facilitates its application in a black-box environment [44]. Intriguingly, as the post-processing techniques do not require entry into the input characteristics, they can be directly implemented in the joint distribution comprising the labels Y and the model predictions $\hat{Y}$, demonstrating the adaptability of this method. However, modifying outputs could potentially distort the accuracy of the entire model. For example, striving for balanced gender representation instead of focusing purely on qualifications may decrease the hiring of competent men – an effect sometimes termed positive bias or affirmative action. Although this process might affect the precision of the model, it fulfills the intended goal in the long run.

Pentyala et al. [45] propose *PrivFairFL*, a comprehensive framework that amalgamates federated learning, secure multiparty computation (MPC), and differential privacy (DP) to develop machine learning models to combat bias against certain demographics delineated by sensitive characteristics like as gender or race [46]. *PrivFairFL* introduces two distinct

TABLE I
FAIRNESS TECHNIQUES IN FEDERATED LEARNING

Category	Technique & Key Idea	Benefits	Limitations
Data Preprocessing	Reweighting: Adjust weights for balance Class Rebalancing: Correct class imbalance	Group fairness; efficiency	Limited flexibility; reduces data diversity
In-Processing	Adversarial Learning: Remove associations Regularization: Fairness penalty in optimization	Applicability; data integrity	Access to model; high computation
Post-Processing	Threshold Adjustment: Optimize thresholds Prediction Mod.: Achieve statistical parity	Model-agnostic; simplicity	Reduced accuracy; lack of transparency

bias mitigation strategies: *PrivFairFL-Pre* and *PrivFairFL-Post*, applied before and after the model training phase respectively. The *PrivFairFL-Pre* debiases the consolidated training dataset by assigning weights to the samples grounded in the sensitive attribute and/or class label values. The MPC computes the aggregated statistics of the label and sensitive attribute value distributions across the federation, ensuring the clients' sensitive features remain undisclosed. Further privacy enhancement is achieved by applying DP to the aggregated statistics, providing a solid privacy guarantee [47]. The MPC is essential for constructing Receiver Operating Characteristic (ROC) curves for protected and unprotected groups without requiring clients to reveal their sensitive attribute values or data distribution. The *PrivFairFL-Post* rectifies the predicted outcomes by determining the optimal classification thresholds for each group using the almost balanced subset of data. Post-construction ROC curves are perturbed using DP for formal privacy assurance. A key characteristic of these strategies is their independence from the model training phase, offering flexibility to be integrated with any federated learning model training technique.

PrivFairFL empirically proved to achieve group fairness in federated learning with formal privacy assurances and also improved the utility of the model even when put on the scale with the baseline approach in [30]. PrivFairFL also mitigated bias and established group fairness using real-world datasets without clients having to provide sensitive features or data distributions. *PrivFairFL*'s computational efficiency and scalability also demonstrate its suitability for real-world applications. However, achieving this group fairness may come at the expense of reduced overall accuracy or performance of the model. Moreso, the use of MPC and DP may incur additional computational overhead. Furthermore, the efficiency of *PrivFairFL* might also be influenced by the unique features of the data and the particular challenge being addressed.

IV. CONCLUSION

This review paper has provided a comprehensive survey of bias mitigation techniques for federated learning systems. We discussed how bias is propagated in federated learning. We categorized existing debiasing strategies based on their implementation during the pre-processing, in-processing, or post-processing phases of model training. Managing non-IID and unbalanced distributed data requires that robust bias mitigation techniques be developed. To enhance model generalization, potential directions include client weighting schemes, multi-task representation learning, and transfer learning. Moreover, regulations, such as the General Data Protection Regulation (GDPR) that restrict data sharing, impose constraints on bias mitigation necessitating solutions compatible with encryption and differential privacy. What needs to be solved is the problem of advancing privacy-aware debiasing techniques tailored for FL. As FL gains increasing traction across domains like healthcare, finance, and smart cities, ensuring fairness and mitigating unintended biases in FL systems will only grow in importance.

V. DIRECTIONS FOR FURTHER RESEARCH

Future works should focus on developing customizable and dynamic debiasing frameworks. These frameworks should allow practitioners to strike a balance between fairness, Integrity, precision, privacy, scalability, and resiliency without sacrificing one or the other. Training data bias refinement is a crucial FL research area. The cross-silo scenario dominates FL bias reduction literature because correcting bias in the cross-device situation, where data variations are very dynamic, is more complicated. Recent issues, including the necessity to account for non-IID (non-independent and identically distributed) data and the complexity of guaranteeing fairness across different and changing devices, require additional investigation, especially in the cross-device scenario.

REFERENCES

[1] S. B. Aher and L. Lobo, "Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data," *Knowledge-Based Systems*, vol. 51, pp. 1–14, 2013.

[2] Z. Shi, L. Zhang, Z. Yao, L. Lyu, C. Chen, L. Wang, J. Wang, and X.-Y. Li, "Fedfaim: A model performance-based fair incentive mechanism for federated learning," *IEEE Transactions on Big Data*, 2022.

[3] M. Magdziarczyk, "Right to be forgotten in light of regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec," in *6th International Multidisciplinary Scientific Conference on Social Sciences and Art Sgem 2019*, 2019, pp. 177–184.

[4] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] O. Odeyomi and G. Zaruba, "Differentially-private federated learning with long-term constraints using online mirror descent," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1308–1313.

[6] O. T. Odeyomi and G. Zaruba, "Privacy-preserving online mirror descent for federated learning with single-sided trust," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 1–7.

[7] H. White, "Consumer data privacy in a networked world: A framework for protecting a privacy and promoting innovation in the globaeconom," *http://www. whitphi) nse pnY/siles/default/files/privac*.

[8] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE journal on selected areas in communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[9] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li, and S. Guo, "A survey of incentive mechanism design for federated learning," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 2, pp. 1035–1044, 2021.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[12] J. Goetz, K. Malik, D. Bui, S. Moon, H. Liu, and A. Kumar, "Active federated learning," *arXiv preprint arXiv:1909.12641*, 2019.

[13] N. Yoshida, T. Nishio, M. Morikura, K. Yamamoto, and R. Yonetani, "Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data," in *ICC 2020-2020 IEEE International Conference On Communications (ICC)*. IEEE, 2020, pp. 1–7.

[14] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with cpu-gpu heterogeneous computing," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 7947–7962, 2021.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[16] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[17] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *Usenix Security*, 2021.

[18] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, 2021.

[19] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, pp. 965–978.

[20] S. Liu, Q. Lin, J. Li, and K. C. Tan, "A survey on learnable evolutionary algorithms for scalable multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, 2023.

[21] Z. Liu, Y. Chen, Y. Zhao, H. Yu, Y. Liu, R. Bao, J. Jiang, Z. Nie, Q. Xu, and Q. Yang, "Contribution-aware federated learning for smart healthcare," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 396–12 404.

[22] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[23] X. Li, Z. Qu, B. Tang, and Z. Lu, "Fedlga: Toward system-heterogeneity of federated learning via local gradient approximation," *IEEE Transactions on Cybernetics*, 2023.

[24] J. Bian, A. Al Arafat, H. Xiong, J. Li, L. Li, H. Chen, J. Wang, D. Dou, and Z. Guo, "Machine learning in real-time internet of things (iot) systems: A survey," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8364–8386, 2022.

[25] M. Duan, D. Liu, X. Chen, Y. Tan, J. Ren, L. Qiao, and L. Liang, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," in *2019 IEEE 37th international conference on computer design (ICCD)*. IEEE, 2019, pp. 246–254.

[26] Y. Guo, F. Liu, T. Zhou, Z. Cai, and N. Xiao, "Privacy vs. efficiency: Achieving both through adaptive hierarchical federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1331–1342, 2023.

[27] J. Chayes, "How machine learning advances will improve the fairness of algorithms," *Huffington Post, August*, vol. 23, 2017.

[28] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[29] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed *et al.*, "Author correction: Do no harm: a roadmap for responsible machine learning for health care (nature medicine,(2019), 25, 9,(1337-1340), 10.1038/s41591-019-0548-6)," *Nature medicine*, vol. 25, no. 10, p. 1627, 2019.

[30] A. Abay, Y. Zhou, N. Baracaldo, S. Rajamoni, E. Chuba, and H. Ludwig, "Mitigating bias in federated learning," *arXiv preprint arXiv:2012.02447*, 2020.

[31] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[32] R. Gao, F. Liu, K. Zhou, G. Niu, B. Han, and J. Cheng, "Local reweighting for adversarial training," *arXiv preprint arXiv:2106.15776*, 2021.

[33] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher, "Diffprivlib: the ibm differential privacy library," *arXiv preprint arXiv:1907.02444*, 2019.

[34] Y. Zhu, J. An, E. Zhou, L. An, J. Gao, H. Li, H. Feng, B. Hou, W. Tang, C. Pan *et al.*, "Mˆ3 fair: Mitigating bias in healthcare data through multi-level and multi-sensitive-attribute reweighting method," *arXiv preprint arXiv:2306.04118*, 2023.

[35] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 4, pp. 1–30, 2023.

[36] J. Gao, Y. Zhu, W. Wang, Y. Wang, W. Tang, and L. Ma, "A comprehensive benchmark for covid-19 predictive modeling using electronic health records in intensive care: Choosing the best model for covid-19 prognosis," *arXiv preprint arXiv:2209.07805*, 2022.

[37] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[38] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 2012, pp. 35–50.

[39] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[40] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 3–1, 2019.

[41] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, 2018.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[43] D. He, V. K. Soo, H. C. Kim, P. Compston, and M. Doolan, "Comparative life cycle energy analysis of carbon fibre pre-processing, processing and post-processing recycling methods," *Resources, Conservation and Recycling*, vol. 158, p. 104794, 2020.

[44] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Information Fusion*, vol. 90, pp. 148–173, 2023.

[45] S. Pentyala, N. Neophytou, A. Nascimento, M. De Cock, and G. Farnadi, "Privfairfl: Privacy-preserving group fairness in federated learning," *arXiv preprint arXiv:2205.11584*, 2022.

[46] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 245–248.

[47] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.