

Fair Client Selection in Federated Learning: Enhancing Fairness in Collaborative AI Systems

Ranim Bouzamoucha¹^a, Farah Barika Ktata²^b and Sami Zhioua³^c

¹Higher Institute of Applied Sciences and Technology, Sousse, Tunisia

²Higher Institute of Applied Sciences and Technology of Sousse, Université de Sousse, Tunisia

³College of Computing and Information Technology (CCIT), University of Doha for Science and Technology (UDST), Doha, Qatar

Keywords: Federated Learning, Fairness, Client Selection, Multi-Armed Bandit, Bias Mitigation, Demographic Fairness.

Abstract: Fairness in machine learning (ML) is essential, especially in sensitive domains like healthcare and recruitment. Federated Learning (FL) preserves data privacy but poses fairness challenges due to non-IID data. This study addresses these issues by proposing a client selection strategy that improves both demographic and participation fairness while maintaining model performance. By analyzing the impact of selecting clients based on local fairness metrics, we developed a lightweight algorithm that balances fairness and accuracy through a Multi-Armed Bandit framework. This approach prioritizes equitable client participation, ensuring the global model is free of biases against any group. Our algorithm is computationally simple, making it suitable for constrained environments, and promotes exploration to include underrepresented clients. Experimental results show reduced biases and slight accuracy improvements, demonstrating the feasibility of fairness-driven FL. This work has practical implications for applications in recruitment, clinical decision-making, and other fields requiring equitable, high-performing ML models.

1 INTRODUCTION

Machine learning (ML) has become a cornerstone of decision-making in critical domains such as healthcare, finance, criminal justice, and education, where prediction-based algorithms are widely adopted by governments and organizations (Dwivedi et al., 2021). While these systems enhance efficiency and accuracy, they often struggle with fairness issues, embedding societal biases that can lead to discriminatory outcomes.


For instance, automated hiring algorithms have been shown to favor male candidates, perpetuating gender biases present in historical data (Dastin, 2022). Similarly, pulse oximeters—devices used to measure oxygen saturation—have been found to be less accurate for individuals with darker skin tones, resulting in higher misdiagnosis rates among minority groups (Bickler et al., 2005). Such cases highlight the urgent need for fairness-aware ML models, particularly


in high-stakes scenarios where biased predictions can have severe consequences.


Incidents of algorithmic discrimination have eroded public trust in ML systems, partly due to their opaque “black-box” nature. This lack of transparency fosters skepticism about the fairness and reliability of these technologies (Toreini et al., 2023).

Ensuring fairness in ML is particularly challenging when protecting sensitive attributes such as race, gender, or socioeconomic status. While fairness is essential for detecting and mitigating bias, it is often constrained by privacy regulations like the GDPR. Users are understandably concerned about data security during auditing processes, creating a demand for solutions that conduct fairness audits while preserving privacy.

Federated Learning (FL) addresses privacy concerns by enabling decentralized model training, where data remains on client devices and only model updates are shared (Shokri and Shmatikov, 2015). However, FL inherently struggles with fairness. Its decentralized nature exacerbates biases, as non-IID (non-independent and identically distributed) client data can lead to the overrepresentation of specific demo-

^a <https://orcid.org/0009-0005-2599-1626>

^b <https://orcid.org/0000-0001-5706-4548>

^c <https://orcid.org/0000-0003-2029-175X>

graphic groups during training (Zhao et al., 2018).

(Li et al., 2023) examined the privacy-fairness trade-off in FL, proposing methods to ensure privacy does not undermine fairness. Their work addresses challenges such as attack resistance, sensitive attribute sharing, algorithmic fairness, and privacy protection.

Addressing fairness in FL is crucial in sensitive fields where biased outcomes can have severe consequences. Research shows that non-representative data distributions in FL skew model predictions, disproportionately affecting marginalized communities (Buolamwini and Gebru, 2018). Agnostic Federated Learning (AFL) (Mohri et al., 2019) promotes fairness by minimizing the worst-case loss across client groups, ensuring "good-intent fairness." However, this focus on worst-case outcomes may overfit minority groups, degrading overall model performance. Additionally, AFL treats all groups equally without explicit client selection, risking imbalances with skewed data distributions.

FedMinMax (Papadaki et al., 2021) improves fairness by optimizing for the worst-performing demographic group. However, it relies on sensitive attributes (e.g., race, gender), which may be unavailable due to privacy policies or legal restrictions. Using such attributes also introduces privacy risks and compliance challenges under GDPR or CCPA, potentially exposing sensitive data through model updates.

In contrast, our approach implements a fair client selection strategy based on local fairness metrics. By prioritizing clients according to fairness criteria, we address bias at the source, ensuring balanced representation. For example, in a federated diagnostic model across hospitals, our method prioritizes clients with underrepresented demographics, guaranteeing their consistent inclusion. This prevents overfitting to majority groups and captures diverse perspectives from the outset.

Moreover, our approach dynamically adapts to shifts in data distributions and client demographics during training. This makes it suitable for real-world applications where fairness and privacy are critical. The algorithm's simplicity ensures applicability in resource-constrained environments, promoting equitable outcomes without compromising performance or privacy.

2 RELATED WORK

Federated Learning (FL), introduced by Google (McMahan et al., 2016), offers a privacy-preserving framework for model training across

distributed data sources while avoiding data centralization. However, the heterogeneity inherent to FL, particularly with non-IID (non-independent and identically distributed) data, leads to significant fairness challenges. This has spurred extensive research into multiple dimensions of fairness to build unbiased and inclusive FL models.

2.1 Client Participation Fairness

Client Participation Fairness aims to provide clients with diverse computational resources and network conditions with equitable participation opportunities, preventing the model from skewing toward data-rich or frequently participating clients. For example, FedCS (Nishio and Yonetani, 2019) enhances efficiency by selecting clients based on deadlines, though it tends to favor resource-rich clients, leaving resource-limited ones underrepresented. Reputation-Based Client Selection (RBCS) (Tiansheng Huang et al., 2020) introduces long-term fairness by modeling client reputations, allowing low-resource clients to participate more consistently over time, but this approach relies on historical data, raising privacy concerns. FairFedCS (Shi et al., 2023) goes further by using Lyapunov optimization to dynamically adjust selection probabilities and balance participation across clients while allowing initially low-performing clients to improve gradually.

2.2 Demographic Fairness

Demographic fairness is critical for ensuring equitable performance across diverse demographic groups, particularly in sensitive areas like healthcare or finance. FedMinMax (Papadaki et al., 2021) seeks to optimize demographic fairness by enhancing performance for the least-performing group, but it relies on sensitive demographic attributes, posing privacy and regulatory challenges. Alternatively, HA-FL (Roy et al., 2024) achieves demographic fairness without demographic data by minimizing the top eigenvalue of the Hessian matrix during training, which preserves privacy but may be computationally intensive.

2.3 Individual Fairness

Individual fairness ensures consistent, equitable treatment of each client, regardless of their data distribution or participation frequency. Methods like q-Fair Federated Learning (q-FFL) (Li et al., 2020) attain individual fairness by reweighting client loss functions and prioritizing clients with higher disparities

in performance, albeit at higher computational cost. The Power-of-Choice selection (Wang and Kantarci, 2020) accelerates convergence by focusing on challenging data from clients with high error rates, yet it risks over-representing outlier data. Dropout techniques (Wen et al., 2022; Bouacida et al., 2020) allow clients to participate by training on subsets of the global model, which improves accessibility for clients with limited resources but could lower model capacity and accuracy for complex tasks.

2.4 Data Heterogeneity Fairness

Data heterogeneity fairness tackles the challenge of balancing model performance when clients possess highly diverse data distributions. Agnostic Federated Learning (AFL) (Mohri et al., 2019) optimizes for the worst-case client, achieving equitable outcomes across clients with skewed data. GIFAIR-FL (Yue et al., 2023) expands on this by using dynamic reweighting to balance both group and individual fairness during communication rounds, though it incurs increased communication costs. FedGCR (Cheng et al., 2024) addresses performance and fairness by implementing group customization and reweighting to effectively reduce disparities in models without excessive computational demands, though it doesn't directly address demographic subgroups within clients.

Advanced solutions for fairness also include post-processing and adversarial methods. Post-FFL (Duan et al., 2024) enhances fairness by applying fairness constraints in post-processing, making it easy to integrate with existing FL workflows, though it cannot address internal biases formed during training. Another approach (Li et al., 2023) treats fairness violations as adversarial attacks and generates fair adversarial samples on each client to ensure consistent treatment across sensitive attributes, but local adversarial training demands significant computational resources and potentially excludes low-resource clients.

In summary, previous research introduced various approaches to tackle fairness challenges in Federated Learning (FL). Our client selection strategy specifically aims to enhance demographic fairness and ensure a balanced global model that supports equitable decision-making without disadvantaging any group. Additionally, our approach focuses on achieving a balanced trade-off between accuracy and fairness—two aspects that have proven challenging to optimize simultaneously in previous studies. By incorporating an exploration parameter, we seek to select clients in a way that promotes fair participation and contributes to an overall fairer client selection process.

Our client selection strategy uniquely incorporates fairness metrics directly into the client selection process, setting it apart from existing approaches. While previous approaches made significant strides in addressing fairness in Federated Learning (FL), they often focused on specific aspects of fairness or introduced trade-offs that limited their scalability and applicability. For instance, FedCS (Nishio and Yonetani, 2018) and FairFedCS (Shi et al., 2023) targeted fairness in client selection but tended to favor clients with higher resources, which could inadvertently exclude under-resourced clients and skew model performance. Demographic fairness approaches, such as FedMinMax (Papadaki et al., 2021) and HA-FL (Roy et al., 2024), achieved group-level fairness but often relied on sensitive demographic data or computationally intensive calculations, which raised privacy concerns and limited their feasibility in large-scale applications. Individual fairness methods, including q-Fair Federated Learning (q-FFL) (Li et al., 2020), aimed to balance individual contributions but could incur high computational costs, impacting convergence times and overall efficiency. Additionally, data heterogeneity fairness techniques like Agnostic Federated Learning (AFL) (Mohri et al., 2019) and GIFAIR-FL (Yue et al., 2023) ensured equitable outcomes across varied client data distributions but were challenged by increased communication or computational demands, particularly in resource-constrained or decentralized settings.

Recent research in federated learning (FL) has predominantly concentrated on refining aggregation methodologies and implementing post-processing techniques to enhance fairness (Duan et al., 2024; McMahan et al., 2016; Yue et al., 2023). For instance, the FairFed algorithm (Yahya H. Ezzeldin and Avestimehr, 2021) introduced a fairness-aware aggregation method that improved group fairness, particularly in scenarios with highly heterogeneous data distributions across clients. Similarly, the FedFB (Yuchen Zeng and Lee, 2021) algorithm modified the FedAvg protocol to better mimic centralized fair learning, thereby boosting the fairness model compared to non-federated approaches. These strategies, while effective, often involved computationally intensive steps that increased resource demands, especially in large-scale federated learning environments where resource constraints were critical.

In contrast, our approach focuses on optimizing the initial client selection phase, addressing fairness concerns at the outset of the federated learning process. By strategically selecting clients, we aim to minimize the need for intensive post-processing or complex aggregation adjustments. This preemptive strat-

egy not only streamlines the workflow but also aligns with the growing demand for efficient and scalable federated learning solutions. By reducing the computational overhead associated with subsequent aggregation and processing steps, our approach offers a more resource-efficient solution that maintains performance without necessitating additional computational power. Our approach actively explores a broad range of clients and encourages participation from clients with diverse data distributions. This not only improves fairness by including clients with less biased data but also enriches the global model's exposure to a wider range of data patterns, leading to a more robust and equitable model. Unlike RBCS, which requires historical data to address client reputation, our approach integrates Group Fairness in Demographics without requiring sensitive data, reducing privacy concerns while enhancing demographic fairness.

3 METHODOLOGY

The client selection process in Federated Learning (FL) involves choosing a subset of participating clients (devices or nodes) in each training round to contribute updates to the global model (Fu et al., 2023). This selection process is crucial as it directly impacts the performance, fairness, and efficiency of the FL system. Traditionally, client selection is often driven by technical criteria such as computational power, network speed, and data volume. Clients with higher computational resources and stable connectivity are typically favored to ensure faster training and reliable communication and contribute to more accurate and efficient updates for the global model ((Yae Jee Cho and Joshi, 2020), (Jaemin Shin and Lee, 2022), (Jiang et al., 2022)). To manage the training load and ensure diversity, many FL systems employed a random or weighted sampling strategy to select a subset of clients candidates ((Zhao and Joshi, 2022), (Li et al., 2020), (Li et al., 2020)).

Our approach integrates fairness criteria into the client selection process. Rather than selecting clients solely based on technical factors, we included fairness metrics, specifically Statistical Parity Difference (SPD), to assess the demographic balance of each client's local model outcomes. By evaluating clients based on their local SPD values, we focused on incorporating models with less biased outcomes at the client level, thus reducing demographic disparities in the aggregated global model. Statistical Parity Difference (SPD) evaluates whether the likelihood of re-

ceiving a positive outcome is the same between different demographic groups, regardless of sensitive attributes such as race. A model achieves statistical parity if the predicted positive outcome rate is equal for both privileged and underprivileged groups. Mathematically, SPD is defined as:

$$\text{SPD} = P(\hat{Y} = 1 | A = 0) - P(\hat{Y} = 1 | A = 1) \quad (1)$$

where \hat{Y} represents the predicted employment status, $A = 0$ denotes the privileged group, and $A = 1$ denotes the underprivileged group. A positive SPD value indicates bias against the underprivileged group, while a value close to 0 suggests no bias, and a negative value indicates bias against the privileged group.

SPD is particularly well-suited to our approach for several reasons: it directly measures fairness in terms of outcome equality, highlighting whether certain groups are disproportionately favored. Moreover, SPD can be computed without centrally accessing sensitive demographic information, thus, preserving client privacy. By constructing the global model from locally fair models, we achieved a more reliable and generalizable global model that can perform equitably across diverse demographic groups.

We implemented three client selection strategies to evaluate their impact on the global model's fairness in federated learning. These strategies rely on the Statistical Parity Difference (SPD) metric, which measures bias between demographic groups. By focusing on local fairness metrics like SPD, we aim to understand how client selection influences global fairness by implementing an algorithm while changing the selection condition every time.

- **Selection of Clients with Highest SPD Values**

This strategy selects clients with the highest SPD values, representing the worst-case scenario for fairness. By focusing on clients with the most biased data, we observe how the global model adapts and whether the bias is amplified or mitigated. This scenario exemplifies a worst-case outcome for the global model, where aggregating biased client data significantly degrades fairness throughout the federated system.

- **Selection Based on Lowest Fairness Metrics**

In this strategy, clients with the lowest SPD values are selected. These clients have data that is less biased, but it's important to note that having a negative SPD indicates a bias against the privileged group. This approach helps to examine the trade-offs between fairness and performance. This scenario highlights how selecting clients with a strong bias against the privileged

group results in a biased global model with a consistent negative SPD value, suggesting a reverse bias in favor of the underprivileged group. While this strategy may reduce bias against marginalized groups, it risks introducing unfairness toward other demographic groups.

- **Selection Based on Optimal Fairness Metrics**

This strategy aims to eliminate clients whose data could introduce significant bias into the global model. Clients with SPD values close to 0, indicating minimal bias, are selected for participation. This selection strategy leads to a significant reduction in the global model's SPD. This result demonstrates that selecting clients with balanced data can effectively minimize bias in the global model. The model maintains fairness across subsequent rounds, confirming that a careful selection of client with near-optimal fairness metrics has a positive impact on the overall system.

These findings emphasize the importance of carefully selecting clients based on fairness metrics. The results suggest that choosing clients with minimal bias produces a fairer global model, while selecting clients with extreme bias, either positive or negative, can skew the model in unintended ways. This study underscores the value of client selection driven by fairness in federated learning, particularly when aiming to balance fairness across different demographic groups in the dataset. These strategies facilitate analysis of the relationship between local fairness (within clients) and global fairness (in the aggregated model). They allow us to evaluate whether selecting biased or unbiased clients can lead to a global model that is both accurate and fair across demographic groups. The experiments present results that further illustrate the effects of these strategies on fairness and performance.

As the number of biased clients included in the training process increased, the fairness of the global model deteriorated. This finding emphasizes the importance of thoughtful client selection in preventing the propagation of local biases into the global model. The direct influence of client selection on global fairness underscores the critical role that client selection strategies play in federated learning. While choosing clients with favorable fairness metrics can help mitigate bias, it introduces several challenges:

- *Sacrificing Fairness in Client Contribution*
Repeatedly selecting the same “best-case” clients can create a new form of bias by excluding other clients, undermining equal contribution opportunities.
- *Reducing the Importance of Training Rounds*

Consistently selecting the same clients limits data diversity and diminishes the iterative nature of federated learning.

- *Excluding Valuable Data*

Focusing solely on fairness metrics may exclude evolving data from certain clients, missing valuable insights that could benefit the global model.

To balance these concerns, a more holistic client selection approach is necessary—one that promotes fairness while maintaining data diversity.

We frame the client selection problem as a stochastic multi-armed bandit (MAB) problem, where each client a is treated as an ‘arm’, with the reward for selection based on improvement in global fairness relative to local fairness. The federated learning system, as the decision-maker, iteratively selects clients to maximize cumulative fairness across multiple rounds. The reward function prioritizes clients whose local fairness contributions significantly enhance the global model's overall fairness.

Reward Calculation. When a client a is selected, the observed reward Reward_a considers both fairness and accuracy improvements, defined as:

$$\text{Reward}_a = \varepsilon \left(\frac{\alpha \cdot \text{Fairness}_{\text{global},t+1}}{\beta \cdot \text{Fairness}_{\text{local},a,t}} \right) + \gamma (\text{Accuracy}_{\text{global},t+1} - \text{Accuracy}_{\text{global},t}) \quad (2)$$

where:

- $\varepsilon \in [0, 1]$ adjusts the exploration-exploitation balance,
- α, β, γ are scaling parameters that weigh fairness and accuracy contributions.

Mean Reward Update. The mean reward for client a at round t is updated using an incremental average to ensure stability over time:

$$\hat{\text{Reward}}[a_t] = \hat{\text{Reward}}[a_t] + \frac{1}{N[a_t]} (\text{Reward}_t) \quad (3)$$

where:

- $\hat{\text{Reward}}[a_t]$ is the estimated mean reward for client a_t ,
- $N[a_t]$ is the count of times client a_t has been selected,
- Reward_t is the observed reward at round t .

This averaging method mitigates noise in reward observations, yielding a consistent and stable selection process. To balance exploration and exploitation, we use an epsilon-greedy strategy with an exploration probability ϵ . With probability ϵ , the algorithm performs exploration by randomly selecting K clients; otherwise, it performs exploitation by choosing the top K clients based on average rewards.

Input: $\epsilon \in [0, 1]$
Output: Updated mean rewards $\text{Reward}[a]$
Initialization: Set $\epsilon \in [0, 1]$

```

for each round  $t$  do
  Generate random number  $r \in [0, 1]$ 
  if  $r < \epsilon$  then
    | Select  $K$  clients randomly
  else
    | Select  $K$  clients with the highest
    |   mean rewards  $\text{Reward}[a]$ 
  end
  Evaluate rewards for each selected client
  Update  $\text{Reward}[a]$  based on Equation 3
end

```

Algorithm 1: Fair Client Selection Approach.

At each round t , the server updates the estimated mean reward $\hat{R}[a]$ for each client a , leveraging cumulative rewards to refine client selection while maintaining a balance between accuracy and fairness. This epsilon-greedy approach with averaging ensures a fair yet robust selection process, where fairness gains do not excessively compromise accuracy.

Exploration, achieved through a non-zero value of ϵ , is essential in federated learning, where clients often have varying data distributions, computational resources, and levels of data quality. By occasionally exploring new or less frequently selected clients, the algorithm can incorporate diverse data sources, leading to a more comprehensive representation of the data in the global model. This is crucial for fairness, as limiting the selection to high-performing clients or those with the most balanced data might skew the model toward these clients' data characteristics, neglecting underrepresented or marginalized client groups.

Without exploration, the algorithm risks falling into a local optimum, where only a subset of clients—those with initially high rewards—are repeatedly selected. This might prevent the model from discovering other clients whose contributions could lead to even greater long-term gains in fairness and accuracy. Exploration ensures that the algorithm does not prematurely settle on a suboptimal client selection

strategy. By periodically exploring different clients, the system mitigates the risk of overfitting to a specific subset of clients with high initial rewards.

This broader exploration helps balance global fairness with model accuracy by ensuring a wider data sampling, preventing the model from being overly influenced by frequent contributors. The choice ϵ is generally determined by the problem context; A higher ϵ encourages exploration, where there is a substantial client performance variation or where fairness across multiple demographics is critical. A lower ϵ value favors exploitation, which is more suitable when the system is confident in the stability and representativeness of the selected clients.

4 EXPERIMENTS SETUP

In this research, we evaluated our approach using US Census data within a distributed learning framework characterized by natural data partitioning. Specifically, we used the ACS Employment dataset introduced by Ding et al. in 2021. The main task was to predict an individual's Employment Status Record (ESR)—whether they are employed or not—based on various features from the Census survey. To facilitate the distribution of data across clients, we modeled 50 clients, each client corresponds to a state. Each client has a distinct data size. For simplicity, we focused on race as a sensitive feature, modifying the dataset to include only two races: White and Black for the sake of simplicity. The distribution of races across states is illustrated in the Figure 1 below

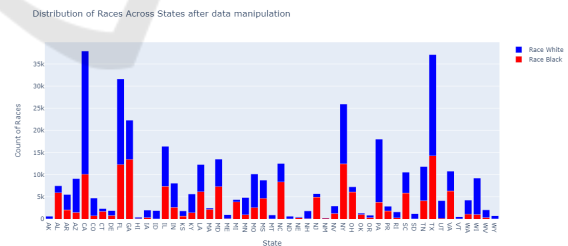


Figure 1: Distribution of races across clients.

Statistical Parity Difference (SPD) was chosen as the fairness metric for evaluating and selecting clients in federated learning because of its clarity, efficiency, and effectiveness in assessing demographic fairness. when SPD values are close to zero, the model treats all client groups fairly across demographic differences, indicating a minimal disparity in outcomes. The distribution of SPD across clients is illustrated in Figure 2

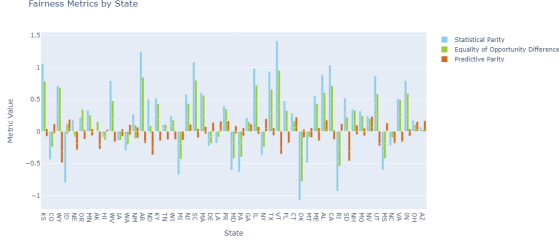


Figure 2: SPD distribution across clients.

This metric’s simplicity also reduces computational overhead, enabling fast fairness assessments—a crucial benefit in federated learning systems, where efficient computation is essential due to the distributed nature of data and model updates. By focusing on aggregated outcome rates across groups, SPD offers a practical and privacy-compliant way to assess fairness, making it an effective tool for ensuring fair client selection and promoting demographic equity in model performance.

5 RESULTS

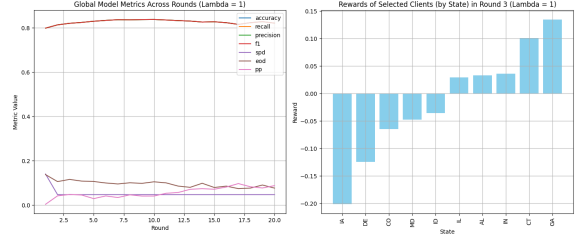
We evaluated five client selection strategies on the US Census dataset using a 50-state federated learning setup. Table 1 summarizes the comparative results.

Table 1: Comparative performance of client selection strategies.

Strategy	SPD	Accuracy
Random Baseline	0.140	0.79
Highest SPD	0.600	0.75
Lowest SPD	-0.350	0.77
Optimal SPD (≈ 0)	0.030	0.80
Reward-Based ($\lambda = 1$)	0.046	0.82

By selecting clients based on their contributions to both fairness and performance, the global model maintained strong predictive capabilities. The reward-based strategy achieved the best overall trade-off, with an SPD of 0.046 and the highest accuracy of 0.82. This corresponds to a 67.1% reduction in bias compared to the random baseline (SPD = 0.140), alongside a 3.8% improvement in accuracy.

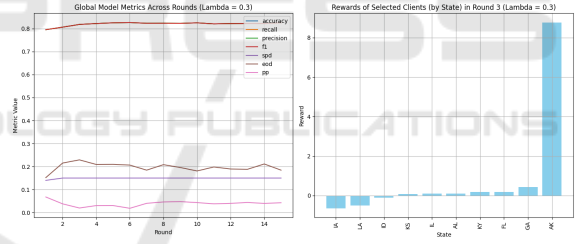
Figure 3 shows that various performance metrics (accuracy, precision, recall, F1) and fairness metrics (SPD, EOD, PP) evolve in a consistent manner over training rounds, suggesting that improvements in performance align with reductions in bias. This convergence supports the effectiveness of the reward-based selection strategy, as multiple independent measures yield similar outcomes. The reward distribution in

Figure 3: Client selection with $\lambda = 1$. Left: Global model metrics across training rounds. Right: Rewards of selected clients (by state) in Round 3.

Round 3 further illustrates that the strategy favors clients contributing positively to both fairness and accuracy, reinforcing the reliability of the approach.

Figure 3 also highlights the impact of pure exploitation ($\lambda = 1$), where only clients with the highest mean rewards are selected. This approach avoids exploration and focuses exclusively on performance-driven selections.

These results suggest that fairness and performance are not necessarily in conflict. The fairness-driven client selection process reduced bias while simultaneously improving accuracy. The decrease in SPD indicates enhanced fairness, while the rise in accuracy shows that performance was not compromised.

Figure 4: Client selection with $\lambda = 0.3$. Partial exploration allows 30% of clients to be chosen randomly.

In Figure 4, we introduce partial exploration by setting $\lambda = 0.3$, allowing 30% of clients to be selected randomly, while the remaining 70% are chosen based on their average reward. In each round, 7 clients are selected based on rewards and 3 are chosen at random. This ensures participation from under-explored clients and promotes data diversity.

Following this adjustment, SPD values fluctuated slightly between 0.14 and 0.15. Although new clients were introduced via exploration, the overall fairness improvement was limited compared to the pure exploitation case. Accuracy increased from 0.79 to 0.80 and stabilized around 0.82. While partial exploration slightly enhanced performance, the gains were less pronounced than with $\lambda = 1$.

With $\lambda = 0.3$, the model incorporated under-represented clients into training. However, the fair-

ness improvement (as indicated by SPD) remained modest. Compared to pure exploitation, which achieved a significantly lower SPD, partial exploration did not substantially enhance fairness. The slight accuracy improvement also suggests that exploration added diversity but did not outperform focused exploitation.

The confidence intervals calculated for accuracy and SPD offer insights into the reliability and consistency of these metrics under the client selection strategy. The accuracy shows a relatively narrow confidence interval of (0.814, 0.822) around a mean of 0.818, suggesting a low variability and a high precision in the model's performance across the selected clients. This narrow interval suggests that the model's accuracy is stable across rounds, with minimal fluctuations, indicating a consistent performance for the global model. In contrast, the confidence interval for SPD is wider, spanning from 0.105 to 0.193 around a mean of 0.149. This broader interval suggests greater variability in the fairness metric, indicating that the model's fairness outcomes vary more significantly across different clients.

These results highlight that fairness and performance are not inherently in conflict. The fairness-driven selection process successfully reduced bias while improving accuracy. The decrease in SPD proves enhanced fairness, and the slight increase in accuracy confirms that prioritizing fairness can complement, rather than hinder, overall model performance.

6 DISCUSSION

Our experiments demonstrate that fairness-driven client selection can significantly enhance the global model's fairness and effectively reduce statistical Parity Difference (SPD) while maintaining or even slightly improving accuracy. This suggests that the trade-off between fairness and performance appears manageable, particularly in federated learning settings where a balanced approach to client selection can ensure equitable client participation without sacrificing the model's quality. Compared to other methods that may prioritize performance over fairness, our strategy achieved a more sustainable balance between these objectives, demonstrating that it is possible to create models that are both accurate and equitable.

The proposed strategy incorporated an exploration mechanism that promoted fairness by encouraging a diverse client pool to contribute to the global model. This mechanism ensured that clients representing underrepresented groups were included, thereby broad-

ening the data distribution and reducing the risk of bias in real-world applications, particularly in sensitive fields like healthcare. In healthcare diagnostics, for example, this approach can prevent majority groups with more consistent data representation from disproportionately influencing the model. By selectively including clients from marginalized groups, the model could be adapted to a more representative distribution, balancing accuracy across demographic groups and reducing the risk of biased medical predictions that could adversely affect specific populations.

Our strategy combined exploration to promote fairness and exploitation to maintain performance. It provided a practical pathway for building suitable models for real-world applications where fairness and performance are equally vital. This adaptive client selection method ensures that the resulting models meet high standards of both fairness and reliability, which is crucial in fields where decisions directly impact individuals' access, health, and outcomes. Traditional federated learning methods often focused on model aggregation without fully considering the representativeness of participating clients, which could inadvertently introduce biases, especially in cases of heterogeneous, non-IID data (Kairouz et al., 2019). Our approach, which integrated a fair client selection mechanism, improved model fairness by prioritizing clients based on fairness rewards that account for demographic representation and balanced participation. This adjustment is impactful across various real-world applications.

In digital healthcare (Zhang et al., 2024), federated learning models trained across diverse hospitals and clinics must ensure fair treatment across different patient demographics. Implementing our client selection strategy enables the selection of clients based on patient demographic fairness rewards, while periodically exploring other clients to prevent demographic and participation biases. This approach fosters equitable healthcare predictions across patient populations, addressing the risk of biased healthcare models that might otherwise favor data-rich institutions.

In smart city management (Wang et al., 2022), where federated learning aids in areas like traffic monitoring and pollution control, data from sensors in affluent or densely populated areas might dominate, potentially skewing model outcomes. Our approach mitigated this by prioritizing sensors in underrepresented or lower-income regions, based on fairness rewards, and periodically exploring new regions to maintain balance. This strategy ensures that city management models provide fair and accurate predictions across diverse neighborhoods, benefiting the en-

tire community.

Similarly, in financial services, especially in credit risk assessment, federated models must avoid biases that favor clients from data-rich, often wealthier, regions. In retail supply chain management, federated learning models need to accurately predict demand across stores in diverse locations, including both urban and rural areas. Larger urban stores typically have more data, which can lead to biases that favor their inventory needs over smaller stores.

However, our approach has certain limitations. It relies heavily on fairness metrics like SPD and Equal Opportunity Difference (EOD), which must be accurately calculated at the client level—a task that can be challenging due to data privacy constraints (Rafi et al., 2024). Furthermore, focusing on clients with balanced data may lead to the underutilization of clients with highly skewed data, potentially affecting the model's generalizability. Addressing these challenges may require advanced privacy-preserving techniques, such as differential privacy ((Saifullah et al., 2024),(Zhou et al., 2024)) and secure multi-party computation (Lindell, 2020), which enable fairness assessments while protecting client privacy.

7 CONCLUSION AND FUTURE WORK

This study presents a novel client selection strategy for federated learning that addresses demographic biases while preserving model accuracy. By incorporating fairness metrics directly into the selection process, the proposed method promotes equitable participation among clients and reduces biases in the aggregated global model. Its simplicity, adaptability, and low computational overhead make it suitable for deployment across diverse real-world applications, including resource-constrained environments.

One of the core strengths of the method lies in its ability to maintain a balance between fairness and performance. Experimental results showed that fairness improvements do not come at the cost of model accuracy. Instead, the approach demonstrated that equitable federated learning is achievable by carefully selecting clients based on fairness indicators. Furthermore, the transparent integration of fairness metrics enhances the interpretability and accountability of the system, allowing stakeholders to better understand and monitor model behavior.

The approach is also highly scalable and robust. It performed consistently across multiple fairness metrics, making it adaptable to different domains

where fairness concerns are context-specific—such as healthcare, finance, and education. Its generalizability allows it to serve as a versatile tool for practitioners aiming to build fair and inclusive machine learning systems.

Despite its strengths, this work also highlights key areas for future research. First, there is a need to develop adaptive fairness mechanisms that dynamically adjust thresholds based on contextual requirements and evolving data characteristics. Such mechanisms would allow federated learning systems to respond to nuanced and domain-specific fairness challenges in real-time.

Second, while the current approach focuses on group fairness—using metrics like Statistical Parity Difference (SPD)—future extensions should consider individual fairness, which ensures that similar individuals are treated similarly regardless of group membership. Combining both fairness notions would offer a more holistic fairness framework in federated learning. Beyond group fairness, future research should incorporate individual fairness, which ensures that similar individuals receive similar model predictions regardless of their group affiliation. This could be done by integrating instance-level fairness constraints within the local client training or reward functions.

Moreover, legal fairness and compliance are essential in sensitive domains like healthcare and finance. Future adaptations of this framework must align with regulations such as the General Data Protection Regulation (GDPR) or California Consumer Privacy Act (CCPA), particularly when fairness evaluations are based on demographic groupings.

Third, privacy concerns remain a significant challenge. Since fairness evaluation often relies on aggregated client data, this could potentially compromise user confidentiality. Incorporating differential privacy techniques can mitigate these risks by enabling fairness-aware computations without exposing individual-level data. Additionally, the adoption of secure multi-party computation (SMPC) can further enhance data security during the exchange of model updates and metrics, ensuring privacy-preserving fairness evaluations ((Dwork and Roth, 2014), (Banse et al., 2024)).

Finally, it is essential to validate the proposed methodology on larger and more diverse datasets across sectors. Such validation will help confirm its scalability, generalizability, and practical utility, providing deeper insights into its real-world impact on fairness, performance, and inclusivity.

In conclusion, this work lays the foundation for fairness-aware federated learning by introducing a

client selection strategy that balances equity and efficiency. With further refinement and rigorous evaluation, it has the potential to become a standard practice in building responsible and trustworthy decentralized machine learning systems.

ACKNOWLEDGEMENTS

The authors acknowledge the use of Copilot (Microsoft, [https://m365.cloud.microsoft/chat]) to summarize the initial notes and to proofread the final draft. The authors have reviewed and validated all AI-generated content for accuracy and coherence.

REFERENCES

- Banase, A., Kreischer, J., and Jürgens, X. O. (2024). Federated learning with differential privacy.
- Bickler, P., Feiner, J., and Severinghaus, J. (2005). Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation. *Anesthesiology*, 102(4):715–719.
- Bouacida, N., Hou, J., Zang, H., and Liu, X. (2020). Adaptive federated dropout: Improving communication efficiency and generalization for federated learning.
- Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.
- Cheng, S.-L., Yeh, C.-Y., Chen, T.-A., Pastor, E., and Chen, M.-S. (2024). Fedgcr: Achieving performance and fairness for federated learning with distinct client types via group customization and reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11498–11506.
- Dastin, J. (2022). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications.
- Duan, Y., Tian, Y., Chawla, N., and Lemmon, M. (2024). Post-fair federated learning: Achieving group and community fairness in federated learning via post-processing. *arXiv preprint arXiv:2405.17782*.
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and Williams, M. D. (2021). Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57:101994.
- Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- Fu, L., Zhang, H., Gao, G., Zhang, M., and Liu, X. (2023). Client selection in federated learning: Principles, challenges, and opportunities.
- Jaemin Shin, Yuanchun Li, Y. L. and Lee, S. (2022). Sample selection with deadline control for efficient federated learning on heterogeneous clients. *CoRR*, abs/2201.01601.
- Jiang, Z., Xu, Y., Xu, H., Wang, Z., and Qian, C. (2022). Adaptive control of client selection and gradient compression for efficient federated learning.
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D’Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2019). Advances and open problems in federated learning. *CoRR*, abs/1912.04977.
- Li, J., Zhu, T., Ren, W., and Raymond, K.-K. (2023). Improve individual fairness in federated learning via adversarial training. *Computers & Security*, 132:103336.
- Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2020). Fair resource allocation in federated learning.
- Lindell, Y. (2020). Secure multiparty computation (MPC). Cryptology ePrint Archive, Paper 2020/300.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2016). Communication-efficient learning of deep networks from decentralized data.
- Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning.
- Nishio, T. and Yonetani, R. (2018). Client selection for federated learning with heterogeneous resources in mobile edge. *CoRR*, abs/1804.08333.
- Nishio, T. and Yonetani, R. (2019). Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7.
- Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2021). Federating for learning group fair models.
- Rafi, T. H., Noor, F. A., Hussain, T., and Chae, D.-K. (2024). Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198.
- Roy, S., Sharma, H., and Salekin, A. (2024). Fairness without demographics in human-centered federated learning.

- Saifullah, S., Mercier, D., Lucieri, A., Dengel, A., and Ahmed, S. (2024). The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods. *Frontiers in Artificial Intelligence*, 7.
- Shi, Y., Liu, Z., Shi, Z., and Yu, H. (2023). Fairness-aware client selection for federated learning.
- Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321.
- Tiansheng Huang, Weiwei Lin, W. W., He, L., Li, K., and Zomaya, A. Y. (2020). An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *CoRR*, abs/2011.01783.
- Toreini, E., Mehrnezhad, M., and Moorsel, A. (2023). Fairness as a service (faas): verifiable and privacy-preserving fairness auditing of machine learning systems. *International Journal of Information Security*, 23:1–17.
- Wang, Y. and Kantarci, B. (2020). A novel reputation-aware client selection scheme for federated learning within mobile environments. In *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–6.
- Wang, Y., Su, Z., Luan, T. H., Li, R., and Zhang, K. (2022). Federated learning with fair incentives and robust aggregation for uav-aided crowdsensing. *IEEE Transactions on Network Science and Engineering*, 9(5):3179–3196.
- Wen, D., Jeon, K.-J., and Huang, K. (2022). Federated dropout – a simple approach for enabling federated learning on resource constrained devices.
- Yae Jee Cho, J. W. and Joshi, G. (2020). Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *CoRR*, abs/2010.01243.
- Yahya H. Ezzeldin, Shen Yan, C. H. E. F. and Avestimehr, S. (2021). Fairfed: Enabling group fairness in federated learning. *CoRR*, abs/2110.00857.
- Yuchen Zeng, H. C. and Lee, K. (2021). Improving fairness via federated learning. *CoRR*, abs/2110.15545.
- Yue, X., Nouiehed, M., and Al Kontar, R. (2023). Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science*, 2(1):10–23.
- Zhang, F., Shuai, Z., Kuang, K., Wu, F., Zhuang, Y., and Xiao, J. (2024). Unified fair federated learning for digital healthcare. *Patterns*, 5(1):100907.
- Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. *CoRR*, abs/1806.00582.
- Zhao, Z. and Joshi, G. (2022). A dynamic reweighting strategy for fair federated learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8772–8776.
- Zhou, R., Dong, A., Yu, J., and Ding, Q. (2024). Fedl-rdp: Federated learning framework with local random differential privacy. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.