Explainable Privacy-Preserving Federated Learning Using Differential Privacy and FedProx for Non-IID Image Classification

Sujay S[1], Kavyashree H[2], Kavya T N[3]
[1,2,3]*Department of MCA Surana College (Autonomous) Bengaluru, India*

***Abstract*—Federated Learning (FL) is well suited for privacy-sensitive applications because it enables collaborative model training without requiring clients to share raw data. How-ever, real-world federated environments often contain Non-Independent and Identically Distributed (Non-IID) data, which can lead to optimization instability, privacy concerns, and limited model interpretability. This paper presents an explain-able privacy-preserving federated learning framework for image classification by integrating Differential Privacy (DP), FedProx optimization, and GradCAM-based explainability. The proposed framework is evaluated on CIFAR-10 and CIFAR-100 datasets under different federated configurations, including FedAvg, Fed-Prox, DP-FedAvg, and DP-FedProx.**

The experimental results show that FedProx improves op-timization robustness in non-IID settings, while Differential Privacy introduces a measurable trade-off between privacy pro-taction and predictive utility. The explainability analysis further indicates that privacy-preserving optimization produces more spatially distributed activation patterns compared with non-private models. These findings highlight the close relationship between optimization stability, privacy preservation, and inter-pretability, supporting the development of more transparent and trustworthy federated learning systems.

***Index Terms*—Federated Learning, Differential Privacy, Fed-Prox, Explainable Artificial Intelligence, GradCAM, Non-IID Learning, Image Classification, Privacy-Preserving Machine Learning.**

I. INTRODUCTION

Federated Learning (FL) enables multiple clients to train a shared model collaboratively without centrally sharing their raw data. This makes FL a promising approach for privacy-sensitive machine learning applications such as healthcare, mobile intelligence, and distributed image analysis [1]. In practice, however, federated systems often operate in environments where client data is Non-Independent and Identically Distributed (Non-IID). Such heterogeneity can create opti-mization instability, increase client drift, and slow or degrade convergence. To address these issues, FedProx extends fed-erated optimization by adding proximal regularization, which limits excessive divergence between local client objectives and the global model under heterogeneous learning conditions [2]. Although FL reduces the need to expose local datasets directly, it does not eliminate privacy risks completely. Gra-dients and model updates exchanged during training may still reveal sensitive client information. Differential Privacy (DP) offers a mathematically grounded way to reduce this risk through gradient clipping and calibrated noise injection during optimization [3]. At the same time, stronger privacy protection often introduces a privacy–utility trade-off, where model performance decreases as privacy constraints become more stringent.

In addition to robustness and privacy, explainability is an important requirement for trustworthy machine learning systems. Deep neural networks often behave like black-box models, which limits transparency and can reduce user con-fidence, especially in high-stakes applications. GradCAM is a widely used visual explanation method that highlights the image regions most responsible for a model prediction [4]. While privacy-preserving federated optimization has received significant research attention, relatively fewer studies examine optimization robustness, privacy preservation, and explainabil-ity together in heterogeneous federated image classification scenarios.

Motivated by these limitations, this work proposes an explainable privacy-preserving federated learning

framework that integrates Differential Privacy, FedProx optimization, and GradCAM-based explainability for non-IID image classification. Experiments are conducted on CIFAR-10 and CIFAR-100 datasets to evaluate predictive performance, privacy–utility trade-offs, optimization robustness, communication overhead, and quantitative explainability behavior. The main contributions of this work are as follows:

- An explainable privacy-preserving federated learning framework is developed by combining Differential Pri-vacy, FedProx, and GradCAM for heterogeneous image classification.
- Extensive experiments are conducted in non-IID federated settings using CIFAR-10 and CIFAR-100 benchmark datasets.
- Ablation studies are performed on proximal regularization, privacy noise levels, statistical testing, and multi-seed evaluation to strengthen empirical validation.
- A GradCAM entropy-based analysis is presented to ex-amine how privacy-preserving optimization affects ex-planation concentration in federated image classification systems.

II. LITERATURE REVIEW

Federated Learning (FL) has become an important paradigm for privacy-aware collaborative machine learning because it allows distributed clients to train models jointly without di-rectly sharing raw datasets [1]. Although this setting reduces centralized data exposure, real-world deployments frequently encounter statistical heterogeneity, where clients hold Non-Independent and Identically Distributed (Non-IID) data. This heterogeneity can significantly affect optimization stability and often results in client drift, slower convergence, and reduced global model performance.

Recent studies have focused on improving federated optimization under heterogeneous conditions. Reddi et al. pro-pose adaptive federated optimization algorithms, including Feda dam, FedYogi, and FedAdagrad, and demonstrate im-proved convergence behavior in decentralized heterogeneous learning environments [5]. Similarly, Acar et al. introduce FedDyn, which uses dynamic regularization to reduce inconsistency between local client objectives and the global optimization target [6]. Beyond optimization dynamics, Li et al. investigate feature heterogeneity through FedBN, where local batch normalization statistics are preserved during aggregation to improve performance in feature-shift Non-IID settings [7]. Together, these studies show that robust optimization remains a central challenge in practical federated learning.

Among heterogeneous optimization methods, FedProx in-troduces a proximal regularization mechanism that explicitly limits excessive divergence between local updates and the global model during training [2]. By modifying the local optimization objective, FedProx improves learning robustness under statistical heterogeneity and system variability. Since real-world federated systems commonly operate with diverse client data distributions, stabilization techniques such as Fed-Prox remain highly relevant.

Despite these optimization advances, privacy preservation remains a major concern in federated learning. Although FL avoids direct dataset sharing, gradients and model parameters exchanged during training may still expose sensitive client information. Zhu et al. demonstrate through gradient inversion attacks that original training samples can potentially be reconstructed from shared optimization signals, emphasizing the need for stronger privacy guarantees in collaborative learning environments [8]. Differential Privacy (DP) addresses this problem by providing mathematically bounded privacy guarantees through gradient clipping and calibrated noise injection [3].

Building on these foundations, recent work increasingly integrates Differential Privacy into federated learning systems. Chen and Vikalo examine federated learning in non-IID environments using differentially private synthetic data and report improved learning behavior under heterogeneous settings [9]. Privacy-preserving federated learning has also gained attention in image-centric and healthcare applications. Recent work on privacy-aware federated medical image classification explores the interaction between Differential Privacy and decentral-
ized visual learning in sensitive clinical environments [10]. Likewise, MedPerf presents a federated benchmarking frame-work for medical AI evaluation and emphasizes reproducible privacy-aware validation across distributed institutional set-tings [11]. While such studies strengthen privacy protection, they also show that stronger privacy mechanisms often reduce

predictive utility.

Alongside optimization and privacy, explainability has be-come essential for trustworthy deployment of deep learning models. Neural networks often operate as black-box predictors, limiting transparency and reducing confidence in high-stakes applications. GradCAM provides a widely adopted explanation method that generates class-discriminative visual localization maps using gradient information from convolutional layers [4]. Although explainability has received significant attention in centralized deep learning, its role in privacy-preserving federated learning remains relatively un-derexplored.

A review of the existing literature reveals three major research trends. First, considerable effort has been devoted to addressing non-IID data and improving convergence in federated optimization. Second, Differential Privacy has become a widely adopted approach for protecting sensitive information in federated learning systems. Third, advances in explainability techniques have helped improve the transparency and inter-pretability of deep learning models.

Despite this progress, these areas are often studied sepa-rately. Only a limited number of studies examine optimization stability, privacy protection, and explainability within a uni-fied federated image classification framework. In particular, the impact of Differential Privacy on the clarity, focus, and spatial concentration of model explanations in heterogeneous federated environments has not been explored in sufficient depth.

To address these gaps, this study presents an explainable and privacy-preserving federated learning framework that in-tegrates Differential Privacy, FedProx optimization, and Grad-CAM-based explanation analysis for image classification un-der non-IID data conditions.

III. METHODOLOGY

This work proposes an explainable privacy-preserving federated learning framework that combines federated optimization, Differential Privacy (DP), and Explainable Artificial Intelligence (XAI) techniques for non-IID image classification. The framework is designed to address three major challenges at the same time: optimization instability caused by heterogeneous client data, privacy leakage risks during collaborative training, and limited interpretability in deep federated models.

A. Federated Learning Framework

The proposed framework follows a centralized federated learning architecture consisting of a coordinating server and multiple distributed clients. Each client stores its local dataset independently and performs local model optimization without sharing raw training samples.

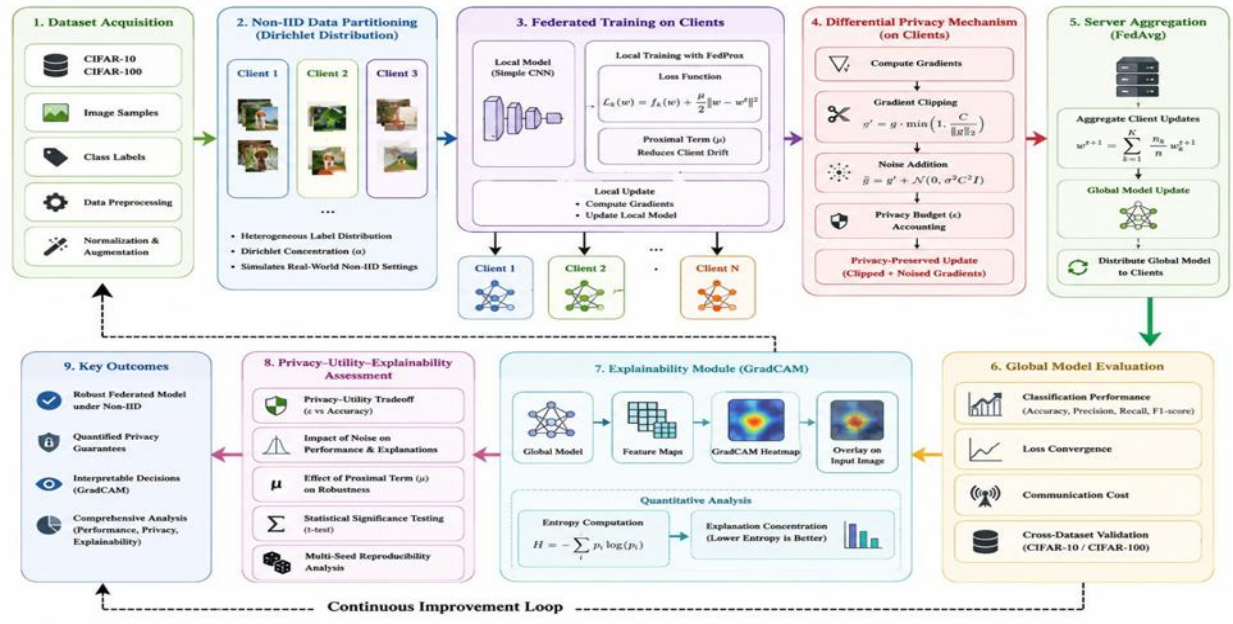

Fig. 1. Proposed explainable privacy-preserving federated learning framework integrating FedProx optimization,

Differential Privacy, and GradCAM-based explainability for non-IID image classification. The framework performs client-side training on heterogeneous data, privacy-preserving model aggregation through FedAvg, quantitative explainability analysis, and comprehensive privacy–utility–interpretability evaluation.

At communication round t, the server sends the current global model parameters w^t to the participating clients. Each client then performs local training using its private dataset and returns the updated model parameters to the server for aggregation.

Global aggregation follows the Federated Averaging (Fe-dAvg) strategy [1]. The global model update is computed as

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{t+1} \quad (1)$$

where K denotes the number of participating clients, n_k represents the number of training samples available at client k, and n denotes the total number of client samples.

To simulate realistic federated environments, heterogeneous client distributions are generated using Dirichlet-based non-IID partitioning. This setup creates statistically imbalanced client datasets and reflects the challenges commonly found in practical federated learning systems.

B. FedProx Optimization for Non-IID Learning

Conventional FedAvg often experiences optimization insta-bility under heterogeneous client distributions because local objectives may drift substantially away from the global optimization target. To reduce this issue, the proposed framework integrates FedProx optimization [2].

FedProx modifies the local training process by introducing a proximal regularization term into each client's optimization objective. For client (k), the local objective is defined as

$$F_k(w) = f_k(w) + \frac{\mu}{2} \|w - w^t\|^2 \quad (2)$$

where f (w) denotes the local empirical loss, wt represents the global parameters received from the server, and μ is the proximal regularization coefficient The additional proximal term discourages the local model from moving too far away from the current global model during client-side training. This helps reduce client drift and improves optimization stability when the data distribution is non-IID. In this study, several values of μ are evaluated experimentally to examine the model's sensitivity to regularization and determine an appropriate configuration.

C. Differential Privacy Integration

Although federated learning prevents direct transmission of client datasets, the optimization signals exchanged during training may still leak sensitive information. To strengthen privacy protection, Differential Privacy is integrated into local optimization.

The framework adopts Differentially Private Stochastic Gra-dient Descent (DP-SGD) [3]. During local training, gradients are first clipped using a predefined norm threshold, after which calibrated Gaussian noise is added.

Given gradient vector g, the clipped gradient is computed as

$$\bar{g} = g \cdot \min\left(1, \frac{C}{\|g\|_2}\right) \quad (3)$$

where C represents the maximum gradient norm.

The noise-perturbed gradient is then generated as

$$\tilde{g} = \bar{g} + \mathcal{N}(0, \sigma^2 C^2 I) \quad (4)$$

where σ denotes the noise multiplier that controls privacy strength.

Privacy guarantees are measured using the privacy budget ϵ, where a smaller ϵ indicates stronger privacy protection. The framework also analyzes privacy–utility behavior through experiments with different noise multiplier settings.

D. GradCAM-Based Explainability Analysis

In addition to predictive performance and privacy preserva-tion, the proposed framework evaluates interpretability using Explainable Artificial Intelligence techniques.

GradCAM is used as the primary explanation method because it produces class-discriminative localization maps from gradient information extracted from convolutional feature layers [4]. For target class c, GradCAM computes feature importance weights as

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

where A^k denotes the k^{th} convolutional feature map and represents the target class prediction score.
The GradCAM explanation map is generated as

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (6)$$

Beyond qualitative visualization, explanation behavior is evaluated quantitatively through entropy analysis. Entropy scores are calculated from normalized heatmaps to measure the concentration of explanations across different federated configurations

E. Experimental Configuration
Experiments are conducted on the CIFAR-10 and CIFAR-100 image classification benchmarks. A lightweight convolutional neural network is used as the base model architecture for all experiments. The federated setting consists of multiple clients with non-IID data distributions generated through Dirichlet partitioning.

Four experimental configurations are evaluated:

- FedAvg
- FedProx
- Differentially Private FedAvg (DP-FedAvg)
- Differentially Private FedProx (DP-FedProx)

Performance is evaluated using classification accuracy, macro F1-score, privacy budget (ϵ), communication cost, statistical significance testing, multi-seed reproducibility analysis, and explainability metrics.

IV. RESULTS AND DISCUSSION

This section evaluates the proposed explainable privacy-preserving federated learning framework under heterogeneous non-IID settings. The experiments examine optimization ro-bustness, the privacy–utility trade-off, explainability behavior, and cross-method performance. The evaluation is carried out using CIFAR-10 and CIFAR-100 image classification bench-marks.

A. Performance Comparison of Federated Methods
The classification performance of the evaluated federated learning approaches is presented in Fig. 2. The results show that FedProx achieves the highest predictive performance among all evaluated methods, reaching approximately 65% classification accuracy compared with 64% for FedAvg. This improvement suggests that proximal regularization helps re-duce client drift and improves optimization stability under heterogeneous client distributions.
where A^k denotes the k^{th} convolutional feature map and y represents the target class prediction score.

The GradCAM explanation map is generated as

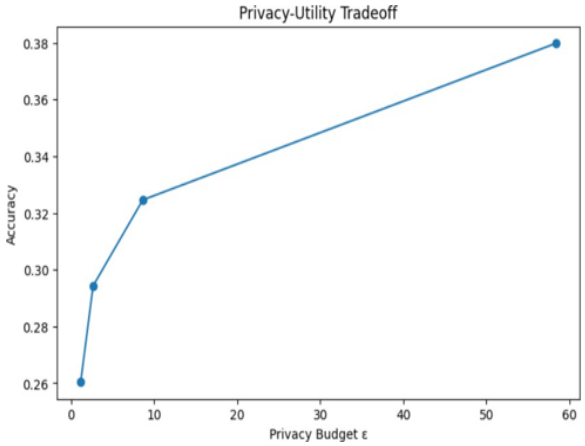

Fig. 2. Accuracy comparison of FedAvg, FedProx, DP-FedAvg, and DP-FedProx on CIFAR-10.

The privacy-preserving variants show substantially lower performance than their non-private counterparts. DP-FedAvg achieves approximately 32% accuracy, while DP-FedProx reaches nearly 34%. Although Differential Privacy reduces performance because of noise injection during optimization, the results indicate that meaningful collaborative learning is still possible under privacy constraints. The slight improvement observed in DP-FedProx also suggests that optimization stabilization remains useful even in privacy-preserving federated environments.
Overall, the findings indicate that FedProx consistently improves learning robustness under non-IID conditions and partially offsets the performance degradation introduced by Differential Privacy mechanisms.

B. Privacy–Utility Tradeoff Analysis
A key objective of this study is to examine the relation-ship between privacy protection and predictive utility. Fig. 3 illustrates the effect of the privacy budget (ϵ) on classification accuracy.

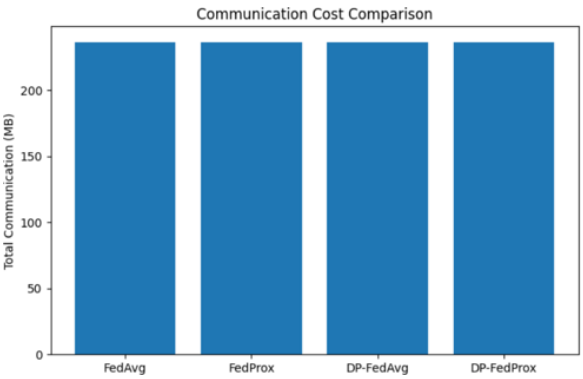

Fig. 3. Privacy–utility tradeoff obtained under different differential privacy configurations.

The results show a clear privacy–utility trade-off. As the pri-vacy budget increases, classification accuracy also improves. Smaller ϵ values indicate stronger privacy guarantees, which require higher noise injection during optimization. This added noise disturbs useful learning signals and leads to lower predictive performance.

When ϵ is close to 1, accuracy remains around 26%. However, when ϵ increases beyond 50, performance improves to approximately 38%. This pattern is consistent with the expected behavior of Differential Privacy, where stronger protection against information leakage typically reduces model utility. Overall, the findings confirm that privacy-preserving design introduces an important trade-off in federated image classification.

C. Explainability Analysis

To study interpretability behavior, GradCAM explanations are generated for all evaluated federated models. A quantitative explainability evaluation is then performed using entropy analysis. Lower entropy values indicate more concentrated ex-planations, while higher entropy values suggest that attention maps are more diffuse and less spatially focused.

Fig. 4 presents the average GradCAM entropy values ob-tained for each federated configuration.

FedAvg and FedProx produce similar entropy values, in-dicating that proximal optimization does not substantially change explanation concentration in the non-private setting. In contrast, the differentially private configurations show higher

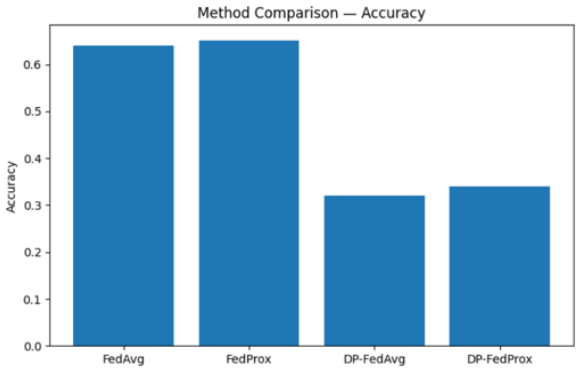

Fig. 4. Comparison of mean GradCAM entropy across federated learning methods.

Entropy Values, with DP-FedProx producing the largest in-crease. This suggests that privacy-aware training leads to more spatially distributed attention maps during inference.

These results indicate that Differential Privacy affects not only predictive performance but also explanation behavior. Noise injection introduces additional uncertainty into learned feature representations, which can lead to broader activation regions in GradCAM outputs. As a result, privacy-preserving models tend to generate visual explanations that are less tightly focused than those produced by non-private federated models.

D. GradCAM Visualization Analysis

Representative GradCAM visualizations are shown in Fig. 5. The original image corresponds to a ship sample from the CIFAR-10 dataset.

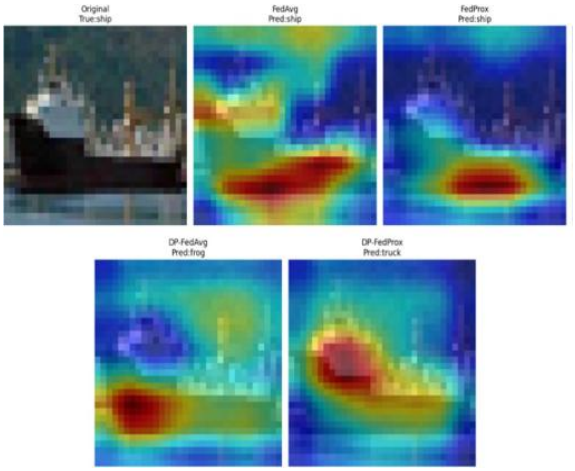

Fig. 5. Comparison of Grad-CAM visual explanations produced by the FedAvg, FedProx, DP-FedAvg, and DP-FedProx models.

As shown in Fig. 5, both FedAvg and FedProx correctly classify the target object as a ship and concentrate their attention on the relevant region. FedProx, however, produces a more localized and clearly defined activation pattern around the ship, which is consistent with its stronger classification performance.

The privacy-preserving models exhibit noticeably different behavior. DP-FedAvg misclassifies the image as a frog, while DP-FedProx predicts it as a truck. Their attention maps are also more dispersed and less closely aligned with the actual object than those produced by the non-private models. These visual patterns support the entropy-based findings and suggest that privacy constraints affect not only predictive accuracy but also the spatial focus of the generated explanations.

Taken together, these results highlight an important trade-off between privacy and interpretability. Although Differential Privacy provides stronger protection against information leak-age, it can alter the spatial structure of the learned feature representations. This leads to explanations that are broader, less focused, and consequently more difficult to interpret.

E. Communication and Statistical Analysis

The communication overhead remains almost the same across all evaluated methods because neither FedProx nor Differential Privacy modifies the underlying model architecture. Each configuration requires approximately 236.65 MB of communication over ten federated rounds. This shows that privacy protection and optimization stability can be added without increasing communication costs.

To examine the reliability of the experimental results, statis-tical significance testing was conducted across multiple trials. Welch's t-test produced a p-value below 0.05, indicating that the differences between the baseline and privacy-preserving configurations are statistically significant. Therefore, the ob-served performance variations are unlikely to have resulted solely from random initialization.

Overall, the experimental results show that FedProx im-proves optimization stability in federated environments with heterogeneous data. Differential Privacy introduces a clear trade-off between privacy and predictive performance. It also affects the behavior of model explanations by reducing their spatial concentration. These results demonstrate the importance of considering optimization, privacy, predictive perfor-mance, and interpretability together when developing reliable federated learning systems.

V. CONCLUSION

This work presents an explainable and privacy-preserving federated learning framework that combines Differential Pri-vacy, FedProx optimization, and Grad-CAM-based interpretation for non-IID image classification. The framework ad-dresses three key challenges in federated learning: unstable optimization caused by heterogeneous client data, the risk of privacy leakage during collaborative training, and the limited interpretability of model predictions.

Experimental results on CIFAR-10 and CIFAR-100 show that FedProx consistently improves optimization robustness in non-IID settings and achieves higher classification performance than standard FedAvg. When Differential Privacy is introduced, privacy protection improves, but a clear privacy–utility trade-off emerges, as stronger privacy guarantees re-duce predictive performance. Cross-dataset evaluation further shows that privacy-related performance degradation becomes more noticeable as task complexity increases. In addition, the GradCAM-based entropy analysis suggests that privacy-preserving optimization changes explanation behavior by pro-ducing more spatially distributed activation patterns.

Overall, the findings indicate that optimization robustness, privacy preservation, and explainability are closely connected in federated learning systems. By studying these dimensions within a unified framework, this work contributes toward the development of more transparent, privacy-aware, and trust-worthy federated image classification models. Future work may explore advanced privacy mechanisms and alternative explanation techniques, particularly in large-scale federated environments where computational and communication con-straints are more demanding.

REFERENCES

[1] B. McMahan *et al*., "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017.

[2] T. Li *et al*., "Federated optimization in heterogeneous networks," in *Proc. Machine Learning and Systems (MLSys)*, 2020.

[3] M. Abadi *et al*., "Deep learning with differential privacy," in *Proc. ACM Conf. Computer and Communications Security (CCS)*, 2016.

[4] R. R. Selvaraju *et al*., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017.

[5] S. J. Reddi *et al*., "Adaptive federated optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[6] D. A. E. Acar *et al*., "Federated learning based on dynamic regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[7] X. Li *et al*., "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[8] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] C. Chen and H. Vikalo, "Federated learning in non-IID settings aided by differentially private synthetic data," *arXiv* preprint arXiv:2206.00686, 2022.

[10] Anonymous (or Authors), "Vision through the veil: Differential privacy in federated learning for medical image classification," *arXiv* preprint, 2023.

[11] Karargyris *et al*., "MedPerf: Open benchmarking platform for medical artificial intelligence evaluation," *Nature Machine Intelligence*, 2023.