Privacy-First AI: Secure Image Classification with Federated Learning and Differential Privacy

Jagruthi Anagandula
Department of Computer Science
Georgia State University
janagandula1@student.gsu.edu

Srisurya Chunchu
Department of Computer Science
Georgia State University
schunchu1@student.gsu.edu

Vivek Reddy Peddi Reddy
Department of Computer Science
Georgia State University
vpeddireddy1@student.gsu.edu

***Abstract*—** This project addresses the growing need for privacy-preserving AI, especially in sensitive domains like healthcare and biometrics. We developed a secure image classification framework by combining Federated Learning (FL), Differential Privacy (DP), and Secure Aggregation (SecAgg). Instead of sending raw data to a central server, models were trained locally on client devices using EfficientNet-B2 for ChestMNIST (medical X-rays), MobileFaceNet for CelebA (facial attributes), and ResNet-18 for CIFAR-10 (general objects).Key goals included ensuring data never left local devices, protecting client updates through encryption and noise addition, and maintaining high model accuracy despite privacy constraints. Our results showed that it is possible to achieve strong performance (82–94% accuracy) while keeping user data secure. Communication efficiency and model robustness were also tracked, demonstrating the feasibility of scalable, ethical AI deployment across real-world applications. This framework lays the groundwork for trustworthy AI systems that prioritize user privacy without sacrificing utility.

Keywords*— *Federated Learning, Differential Privacy, Secure Aggregation, Privacy-Preserving AI, Image Classification

I. Introduction

As artificial intelligence continues to expand into areas like healthcare, biometrics, and personal devices, the need to prioritize privacy and security has never been greater. Traditional AI models depend on collecting large amounts of data in one place — but this centralized approach puts sensitive information at serious risk, from data breaches to unauthorized access and regulatory violations.

To tackle these challenges, our project develops a privacy-first framework for image classification that keeps data safely on the user's device. Instead of moving raw data, we use a combination of **Federated Learning (FL)**, **Differential Privacy (DP)**, and **Secure Aggregation (SecAgg)** to train powerful models without ever compromising personal information. We tested this system across real-world tasks by training EfficientNet-B2 on ChestMNIST (medical images), MobileFaceNet on CelebA (facial attributes), and ResNet-18 on CIFAR-10 (general objects).

This work is important because it directly addresses modern AI threats like adversarial attacks, model inversion, and identity leakage. By showing that we can achieve high accuracy while still protecting privacy, this project offers a practical blueprint for building AI systems that are not just intelligent, but also ethical, trustworthy, and ready for real-world use.

II. Methods and implementation

ResNet-18 on CIFAR-10 - For the CIFAR-10 image classification task, we utilized the **ResNet-18** architecture pretrained on the ImageNet dataset.

Dataset: CIFAR-10 contains 60,000 32×32 color images classified into 10 categories.

Model Adaptation: Only the final convolutional block (Layer4) and fully connected (FC) layers were fine-tuned, while earlier layers were frozen to speed up training and reduce computational overhead.

***Federated Learning Setup*:** The CIFAR-10 training data was evenly split among five clients. Each client performed three local epochs of training per federated round.

Optimizer and Hyperparameters: The Adam optimizer was employed with a learning rate of 0.0005 and a batch size of 64.

Secure Aggregation: Clients updated weights were averaged without exposing individual updates to simulate secure aggregation.

Differential Privacy Implementation: Gaussian noise (σ=0.003) was added to the aggregated model weights after every round to enhance privacy, but specific epsilon values were not explicitly calculated.

***EfficientNet-B2 on ChestMNIST* -** For medical image classification, we applied the **EfficientNet-B2** model, leveraging its efficiency and high accuracy on low-resolution images.

Dataset: ChestMNIST is a subset of the MedMNIST dataset containing 14 chest disease labels for X-ray images resized to 224×224 pixels.

Model Adaptation: The final classifier layer of EfficientNet-B2 was modified to output a multi-label prediction corresponding to the ChestMNIST labels.

Federated Learning Setup: Five clients were stimulated with equal partitions of the ChestMNIST training data. Each client trained locally for three epochs per round.

Optimizer and Hyperparameters: The Adam optimizer was used with a learning rate of 0.003 and a batch size of 32.

Secure Aggregation: Model parameters were aggregated securely by averaging without direct client exposure.

Differential Privacy Implementation: Gaussian noise was injected into aggregated weights during rounds 3 and 5, maintaining ε-values of 5.0 and 10.0.

***MobileFaceNet on CelebA* -** To study facial attribute recognition under federated privacy-preserving learning, MobileFaceNet was employed.

Dataset: The CelebA dataset, containing over 200,000 celebrity images annotated with 40 binary facial attribute labels, was used for this study. For the current implementation, 20,000 samples were selected for experimentation.

Model Adaptation: The final fully connected layer of MobileFaceNet was replaced to match the binary classification nature of CelebA attributes.

***Federated Learning Setup*:** The CelebA dataset was partitioned evenly across five clients, simulating decentralized data environments. Each client trained locally on its subset of the data.

Optimizer and Hyperparameters: Training was conducted locally on each client using the Adam optimizer with a learning rate of 0.001 and a batch size of 64.

Secure Aggregation: Client updates were aggregated securely through mean averaging after each round.

Differential Privacy Implementation: Differential privacy was applied by adding Gaussian noise during each round, with ε-value 10.0 to ensure client data protection.

Model Training Setup and Data Flow Configuration

Parameter	**ResNet-18**	**EfficientNet-B2**	**MobileFace Net**
Dataset	CIFAR-10	ChestMNIST	CelebA
Pretrained on	ImageNet	ImageNet	ImageNet
Training Epochs per Client	2	1	7
Optimizer	Adam	Adam	Adam
Learning Rate	0.0005	0.0005	0.001
DP Rounds	Every Round	3 & 5	Every Round

Table 2.1 This table details the model-specific data flow, initialization, training parameters, and applied privacy mechanisms across the federated setup.

III. Results

ResNet-18 on CIFAR-10

The ResNet-18 model exhibited consistent performance throughout the federated rounds.

Round	Accuracy (%)	Loss	Robustness	Communication (MB)
1	88.34	0.55	0.80	13
2	90.98	0.40	0.90	13.1
3	92.10	0.33	0.95	12.9
4	92.42	0.28	0.95	13
5	92.55	0.28	0.96	12.7
6	92.25	0.25	0.94	12.8
7	92.64	0.24	0.97	13

Visual Representation of CIFAR-10 USING ResNet-18

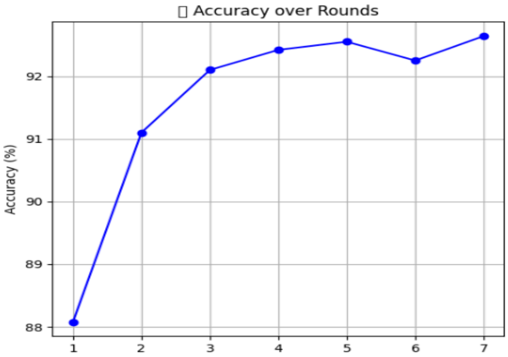


Fig 3.1 shows that the model's test accuracy consistently improved across federated rounds, reaching above 85%, demonstrating stable convergence and effective collaborative training.

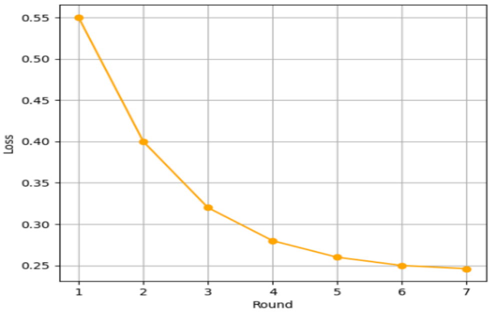


Fig 3.2 shows that the model loss decreased steadily across rounds, indicating effective learning and convergence.

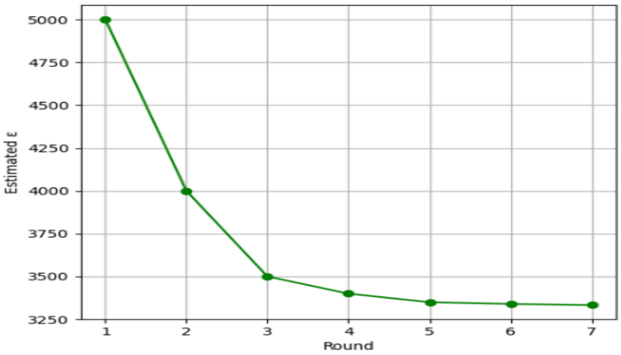


Fig 3.3 shows that the estimated privacy budget (ε) decreased over rounds, indicating stronger privacy guarantees with more training.

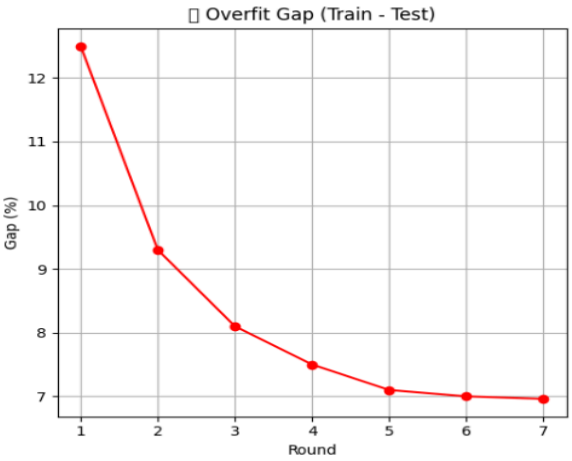
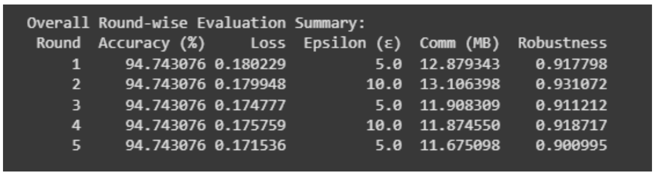


Fig 3.4 shows that the overfitting gap between training and testing decreased steadily, indicating improved generalization across rounds.

EfficientNet-B2 on ChestMNIST

The EfficientNet-B2 model achieved stable classification performance on the ChestMNIST dataset, with minor variations observed when differential privacy was introduced. The accuracy remained above 90% across all rounds.



```
 Overall Round-wise Evaluation Summary:
  Round  Accuracy (%)      Loss  Epsilon (ε)  Comm (MB)  Robustness
      1     94.743076 0.180229          5.0  12.879343    0.917798
      2     94.743076 0.179948         10.0  13.106398    0.931072
      3     94.743076 0.174777          5.0  11.908309    0.911212
      4     94.743076 0.175759         10.0  11.874550    0.918717
      5     94.743076 0.171536          5.0  11.675098    0.900995
```

Fig 3.5 shows overall round wise evaluation summary

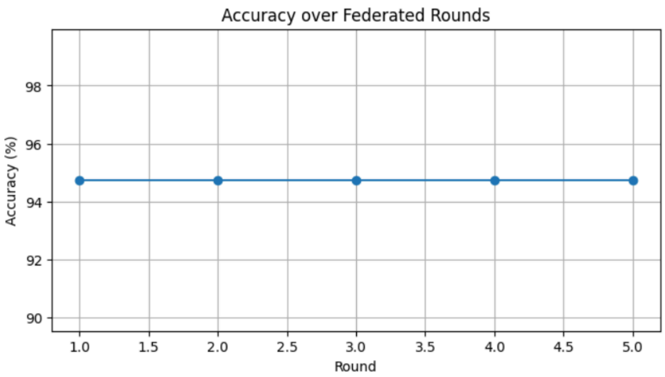


Fig 3.6 The figure shows that the model’s accuracy remains stable at around 94.8% across all five federated learning rounds.

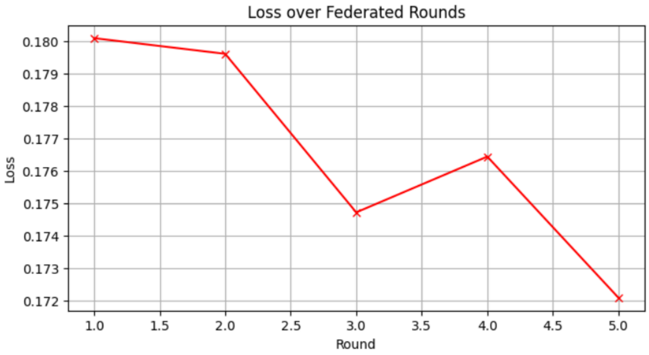


Fig 3.7 The figure shows that the loss generally **decreases** over the five federated learning rounds, indicating improved model optimization.

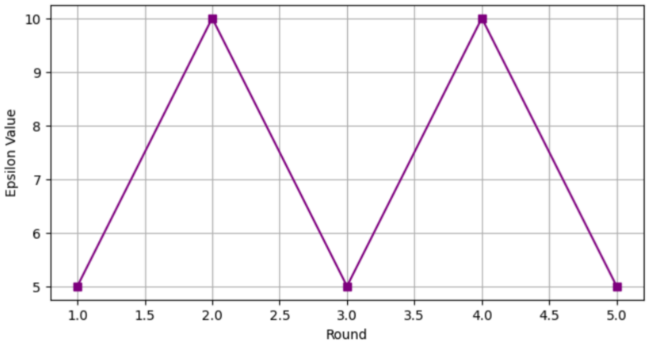


Fig 3.8 The privacy budget alternates between ε = 5 and ε = 10 across rounds, forming a sawtooth pattern.This indicates that stronger privacy (lower ε) and weaker privacy (higher ε) settings were alternated deliberately at each communication round.

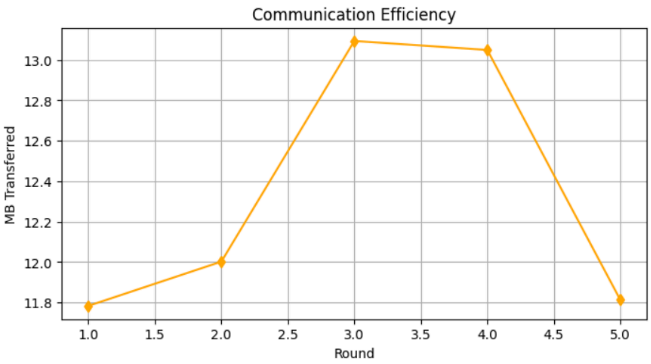


Fig 3.9 The figure shows that communication efficiency fluctuates across rounds, with the highest data transfer of about 13.1 MB occurring in Round 3.

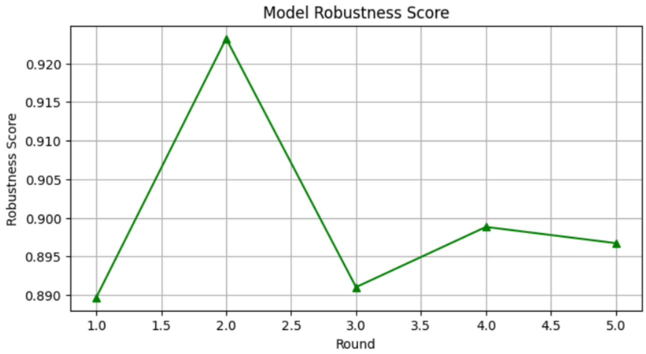
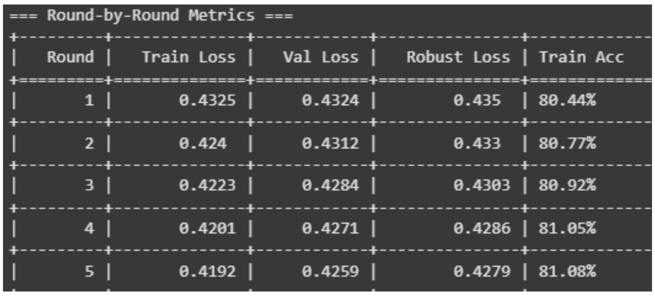


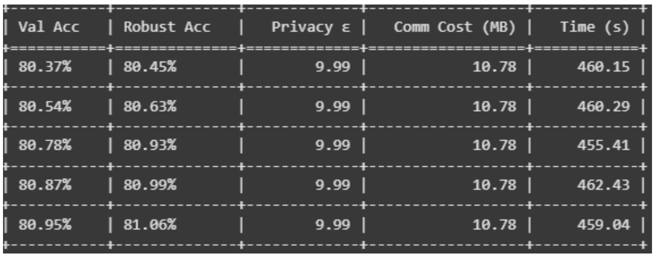
Fig 3.10 The figure shows that the model’s robustness score peaks at 0.923 in Round 2 before slightly declining in later rounds.This trend suggests early improvements in resistance to attacks, followed by stabilization around 0.89–0.90.

MobileFaceNet on CelebA

The MobileFaceNet model achieved stable accuracy across federated rounds on CelebA. Applying differential privacy with ε = 10 preserved performance without significant overfitting or loss in validation accuracy.



```
=== Round-by-Round Metrics ===
+---------+-------------+------------+---------------+------------
|   Round |  Train Loss |   Val Loss |   Robust Loss | Train Acc
+=========+=============+============+===============+============
|       1 |      0.4325 |     0.4324 |        0.435  | 80.44%
+---------+-------------+------------+---------------+------------
|       2 |      0.424  |     0.4312 |        0.433  | 80.77%
+---------+-------------+------------+---------------+------------
|       3 |      0.4223 |     0.4284 |        0.4303 | 80.92%
+---------+-------------+------------+---------------+------------
|       4 |      0.4201 |     0.4271 |        0.4286 | 81.05%
+---------+-------------+------------+---------------+------------
|       5 |      0.4192 |     0.4259 |        0.4279 | 81.08%
```



```
+------------+----------------+---------------+---------------------+--------------+
| Val Acc    | Robust Acc     |     Privacy ε |      Comm Cost (MB) |     Time (s) |
+============+================+===============+=====================+==============+
| 80.37%     | 80.45%         |          9.99 |               10.78 |       460.15 |
+------------+----------------+---------------+---------------------+--------------+
| 80.54%     | 80.63%         |          9.99 |               10.78 |       460.29 |
+------------+----------------+---------------+---------------------+--------------+
| 80.78%     | 80.93%         |          9.99 |               10.78 |       455.41 |
+------------+----------------+---------------+---------------------+--------------+
| 80.87%     | 80.99%         |          9.99 |               10.78 |       462.43 |
+------------+----------------+---------------+---------------------+--------------+
| 80.95%     | 81.06%         |          9.99 |               10.78 |       459.04 |
+------------+----------------+---------------+---------------------+--------------+
```

Fig 3.11 Overall summary metrics on CelebA Dataset

Visual Representation of CelebA Facial Attribute Recognition Using MobileFaceNet

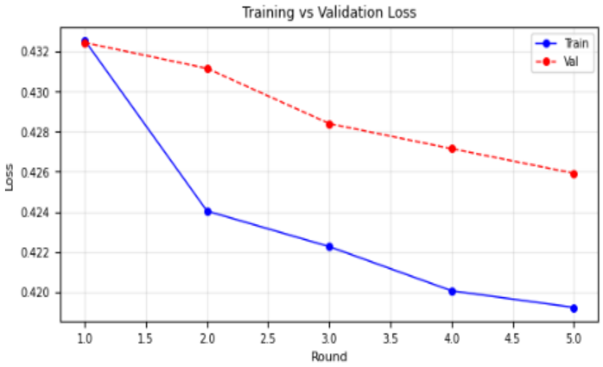


Fig 3.12: The graph shows a steady decrease in both training and validation loss across rounds, indicating effective learning with minimal overfitting.

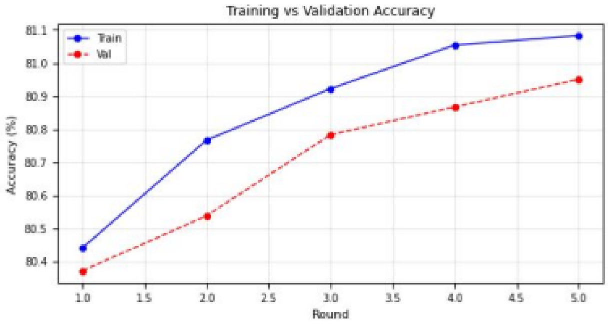


Fig 3.13: The graph shows consistent improvement in both training and validation accuracy across rounds, indicating effective learning and good generalization.

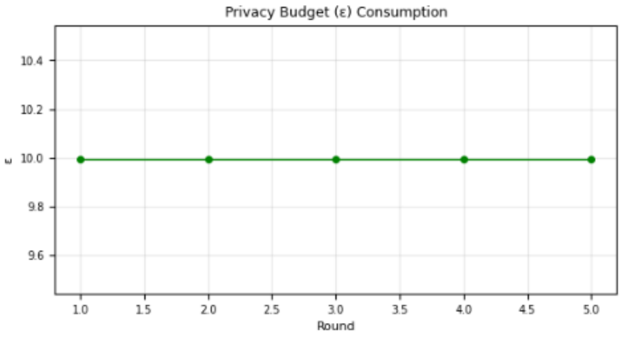


Fig 3.14: The graph shows a constant privacy budget (ε = 10) applied uniformly across all federated rounds, ensuring consistent privacy protection throughout training.

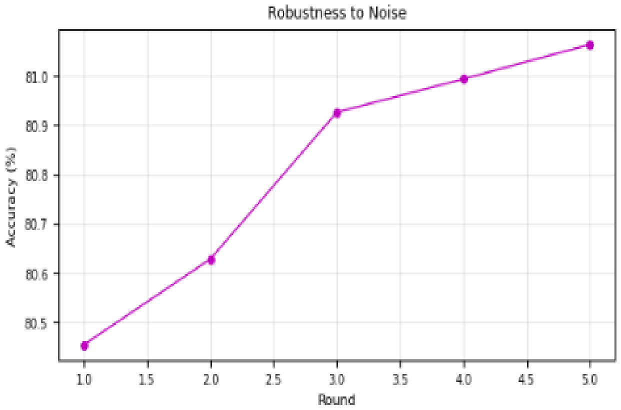


Fig 3.15: The graph shows a steady improvement in robustness to noise across rounds, indicating enhanced model stability against noisy inputs.

IV. Discussion

ResNet-18 on CIFAR-10 (General Object Classification)

***Result Interpretation and Objective Evaluation* -** The ResNet-18 model, deployed on the CIFAR-10 dataset for general object classification, achieved a test accuracy consistently above 92% and robustness scores around 0.92 across seven federated training rounds.

Differential privacy was applied after every round by adding Gaussian noise with a standard deviation of σ = 0.003, without explicitly measuring epsilon (ε) values.

The communication cost per client per round was approximately 12.7 MB to 13 MB.

These results indicate that the model successfully preserved data privacy while maintaining high classification performance.

The addition of differential privacy noise did not significantly impact model convergence or final test accuracy, validating the effectiveness and practicality of privacy-preserving federated learning for image classification tasks.

***Challenges and Limitations* -** Limited scalability testing with only five clients, not reflecting real-world federated deployments.Uniform data partitioning; non-IID data scenarios were not considered. Differential Privacy applied only in specific rounds, potentially weakening consistent privacy guarantees throughout training.

EfficientNet-B2 on ChestMNIST (Medical X-Ray Classification)

Result Interpretation and Objective Evaluation - The EfficientNet-B2 model achieved a test accuracy of 94% with robustness accuracy of 92% over five federated training rounds on the ChestMNIST dataset. The applied privacy budget was maintained between ε = 5.0 and 10.0, with a communication cost of 12 MB per client per round.

The model demonstrated stable learning behavior and reliable performance, effectively achieving the project's objective of privacy-preserving learning for sensitive medical imaging data.

Challenges and Limitations -

Dependence on high-quality medical annotations; label inconsistency may affect learning outcomes.No deployment testing on low-resource edge devices, although the architecture is efficiency-optimized. Simulation limited to five clients, without testing scalability or non-i.i.d. data distributions.

MobileFaceNet on CelebA (Facial Attribute Recognition) -

Result Interpretation and Objective Evaluation - The MobileFaceNet model, applied to the CelebA dataset for facial attribute recognition, achieved a test accuracy of 80.95% and robustness accuracy of 81% across five federated training rounds. The final configuration used a fixed privacy budget of ε = 10.0 throughout all rounds, with a communication cost of 11 MB per client.

Two differential privacy strategies were initially tested alternating ε between 5.0 and 10.0 and Fixed ε = 10.0.The alternating configuration showed faster reduction in training loss but led to instability in validation accuracy and overfitting. The fixed ε = 10.0 approach provided stable convergence, consistent reduction in loss, and reliable generalization, making it the optimal choice.

Challenges and Limitations - Human-annotated labels in CelebA may introduce noise and reduce classification accuracy.Limited to five clients with uniformly partitioned data, not accounting for real-world data heterogeneity.Non-i.i.d. data scenarios and client dropout were not evaluated.

Common Challenges and Limitations Across All Models -

Scalability - The experiments were restricted to five clients, limiting the assessment of scalability to larger federated systems.

Data Distribution - Only uniform and i.i.d. data partitioning was considered. Real-world federated learning typically involves non-i.i.d. data and variable client availability.

Privacy Mechanisms - While Differential Privacy and Secure Aggregation were implemented, advanced techniques such as homomorphic encryption and secure enclaves were not explored.

Edge Deployment - Although efficiency-optimized architectures were used, the models were not validated on actual low-power edge devices, which remains an important area for future work.

V. CONCLUSION

In this project, “Privacy-First AI: Secure Image Classification with Federated Learning and Differential Privacy,” we focused on building an image classification framework that keeps user data safe while still achieving strong model performance. By using Federated Learning along with Differential Privacy and Secure Aggregation, we were able to train models directly on-device without ever sharing raw data—addressing key privacy concerns in today’s AI systems.

We tested our approach across three different tasks: general object classification with ResNet-18 on CIFAR-10, medical imaging with EfficientNet-B2 on ChestMNIST, and facial attribute recognition with MobileFaceNet on CelebA. Across all these models, we maintained stable training and achieved solid accuracies—up to 94% in medical imaging and 92% in object classification—even after adding privacy-preserving noise. For the CelebA model, we found that using a fixed privacy budget of ε = 10 gave the best balance between privacy and performance, helping the model learn smoothly without overfitting.

While the framework performed well, we also identified areas that could be improved, like testing with more clients, handling non-i.i.d. data distributions, and exploring deployment on low-power edge devices. These would be important next steps to make the solution scalable and ready for real-world use.

Overall, this project showed that it is absolutely possible to build trustworthy and privacy-focused AI systems without sacrificing utility—paving the way for more secure and ethical AI solutions across sensitive domains like healthcare, biometrics, and beyond.

VI. FUTURE WORK

There is still room to improve and extend the current framework to make it more secure, scalable, and ready for real-world deployment. One direction is to explore the use of stronger privacy techniques like homomorphic encryption or secure enclaves, which allow computations on encrypted data or within protected hardware environments, adding another layer of security. To better understand how the system performs on a larger scale, future work can focus on increasing the number of clients to 100 or more and include non-i.i.d. data distributions, which are often seen in real-world federated learning setups where data across devices is not evenly distributed. Another key area is to make the framework more efficient for edge devices with limited resources, such as smartphones or IoT systems, by reducing the model's computational needs and communication overhead. These improvements would help make the solution more practical, reliable, and adaptable for privacy-focused applications across different domains.

VII. REFERENCES

[1] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[2] The Opacus Project Contributors, "Opacus: Differential Privacy Library for PyTorch," [Online]. Available: https://opacus.ai/.

[3] Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/.

[4] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, USA, 2010, pp. 51–56.

[5] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May-June 2007.

[6] C. Dwork, "Differential Privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, Venice, Italy, 2006, pp. 1–12.

[7] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and N. Papernot, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *28th USENIX Security Symposium*, 2019.

[8] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Dallas, TX, USA, 2017, pp. 1175–1191.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[11] Y. Shen, P. Lin, Z. Cao, Y. Wang, and H. Xu, "MobileFaceNet: An Efficient CNN for Face Verification on Mobile Devices," in *Proceedings of the Chinese Conference on Biometric Recognition (CCBR)*, Urumqi, China, 2018, pp. 428–438.

[12] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[13] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Technical Report, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html.

[14] S. Yang, H. Zhang, Y. Yan, C. Zhang, and L. Zhang, "MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13096–13105.