RESEARCH ARTICLE OPEN ACCESS

Privacy-Preserving Retinal Disease Detection Using Federated Learning and EfficientNet-B0

Mrs. M. Ishwarya *, P. Kavipriya**,J. Mahima**, S. Roshini**, K.Sujitha**

*Assistant ProfessorComputer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,

**Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai, India

Abstract: *Retinal diseases, including Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and Glaucoma, are leading causes of vision impairment, affecting over 500 million people globally. While deep learning-based automated screening systems have demonstrated accuracy comparable to human experts, their development is hindered by the fragmentation of medical data. Centralizing sensitive patient data is often infeasible due to strict privacy regulations (HIPAA, GDPR) and data governance policies. This paper presents a novel privacy-preserving federated learning framework for retinal disease detection, enabling healthcare institutions to jointly train an EfficientNet-B0 model without sharing raw patient data. The proposed system employs DP-SGD, enforcing gradient clipping and Gaussian noise injection before any model update leaves the local institution, ensuring individual patient records cannot be reconstructed. A weighted FedAvg strategy aggregates privacy-protected updates from distributed clients into a continuously improving global model. Experimental validation on non-IID datasets simulating real-world hospital heterogeneity demonstrates 92.3% accuracy at a privacy budget of ε = 5, confirming effective generalization across diverse clinical environments while maintaining strict regulatory compliance.*

Keywords :*Federated Learning, Differential Privacy, Retinal Disease Detection, EfficientNet-B0,*

1. Introduction

The rapid growth of artificial intelligence has significantly improved automated medical image analysis. Deep learning models trained on large-scale retinal fundus image datasets have demonstrated diagnostic accuracy comparable to clinical experts. Technologies such as Convolutional Neural Networks (CNNs) and Transfer Learning enable detection of critical retinal pathologies — including Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), and Glaucoma — from routine fundus photographs, promising a scalable solution to the global shortage of ophthalmologists.

However, realizing this promise faces a fundamental bottleneck: data fragmentation. Medical imaging data is inherently siloed across thousands of hospitals and clinics. Consolidating it into a centralized repository is increasingly untenable, as strict regulatory frameworks such as HIPAA and GDPR impose severe penalties for unauthorized sharing of patient health information. Furthermore, retinal images constitute biometric data capable of uniquely re-identifying individuals, elevating associated privacy risks considerably.

Models trained on single-institution data also tend to overfit to that institution's patient demographics and equipment, leading to poor generalization and domain shift in real-world deployments.

To address these challenges, this research proposes a Federated Learning framework integrated with Differential Privacy (DP) for retinal disease detection. The proposed system enables geographically distributed healthcare institutions to collaboratively train a shared EfficientNet-B0 model without transmitting raw patient data beyond local secure boundaries. Only privacy-protected gradients are exchanged, with formal guarantees enforced via the DP-SGD mechanism, bridging the gap between AI research and real-world clinical adoption.

2. Literature Review

Gulshan et al. (2016) — JAMA: Pioneered deep learning for diabetic retinopathy detection using Inception-v3 on 128,000+ centralized fundus images. Achieved ophthalmologist-level sensitivity but required data centralization that is legally impermissible under modern privacy regulations.

Ting et al. (2017) — JAMA: Extended the paradigm to multi-disease screening (DR, AMD, Glaucoma), highlighting persistent generalizability challenges from demographic and equipment heterogeneity—a problem federated learning directly addresses.

Kermany et al. (2018) — Cell: Applied transfer learning to OCT retinal images achieving 96.6% accuracy. Revealed data scarcity for rare disease subtypes—an issue multi-institutional federated collaboration can mitigate.

McMahan et al. (2017) — AISTATS: Formally proposed Federated Averaging (FedAvg), enabling decentralized model training while preserving data locality. FedAvg mitigates direct data exposure but does not provide formal gradient-level privacy guarantees.

Tan & Le (2019) — ICML: Introduced EfficientNet achieving state-of-the-art accuracy with fewer parameters than ResNet or Inception. EfficientNet-B0 is ideal for federated deployments on resource-constrained hardware.

Dwork et al. (2006) — TCC: Established the mathematical foundations of Differential Privacy defining (ε, δ)-bounds, providing the theoretical underpinning for DP-SGD used in this work.

Shokri et al. (2017) — IEEE SP: Demonstrated that membership inference attacks can recover sensitive training data from shared gradients, empirically motivating the integration of formal DP into any federated medical AI system.

3. Problem Identification

Existing centralized deep learning approaches for retinal disease detection face four critical and interrelated challenges that motivate the proposed framework:

3.1 Data Privacy and Regulatory Constraints

Centralized training requires aggregating sensitive retinal fundus images from multiple institutions. Retinal vascular patterns constitute biometric identifiers capable of re-identifying individual patients. This directly violates HIPAA, GDPR, and equivalent national data protection laws, exposing institutions to significant legal liability and creating an insurmountable barrier for multi-hospital collaboration.

3.2 Data Fragmentation Across Institutions

Medical imaging datasets are inherently siloed within individual hospitals and diagnostic centres. Institutions are reluctant to share proprietary patient data due to competitive, contractual, and legal constraints. This fragmentation limits the volume and diversity of training data, resulting in models that underperform in real-world deployments.

3.3 Bias and Poor Generalization

Models trained on single-institution datasets suffer from demographic bias and domain shift. Differences in patient ethnicity, disease prevalence, imaging equipment manufacturer, and acquisition protocols produce models that fail to generalize to external institutions, undermining clinical utility and patient safety.

3.4 High Computational and Infrastructure Requirements

Training large backbone architectures such as ResNet-50 or DenseNet-121 demands expensive multi-GPU clusters unavailable to most smaller clinics. This infrastructure barrier creates a "cold

start" problem that excludes resource-constrained institutions from benefiting from modern AI-assisted diagnosis.

4. Objectives and Methodology

4.1 Objective

To develop a privacy-preserving collaborative federated learning framework enabling geographically distributed healthcare institutions to jointly train a high-accuracy retinal disease detection model based on EfficientNet-B0, without sharing any raw patient data, while ensuring full compliance with HIPAA, GDPR, and applicable data governance regulations.

4.2 Federated Learning Architecture

A client-server topology is adopted in which data locality is strictly enforced. Each participating hospital retains its fundus image dataset locally. Only model parameter updates—gradients and aggregated weights—are transmitted across the network. This architecture ensures that raw patient data never leaves the institutional boundary at any point during training.

4.3 EfficientNet-B0 Backbone

EfficientNet-B0 is selected for its exceptional parameter efficiency, achieving competitive ImageNet accuracy with only 5.3 million parameters. This lightweight footprint makes it deployable on standard clinical workstation GPUs, eliminating infrastructure barriers for smaller clinics and enabling broad federated participation.

4.4 Differential Privacy (DP-SGD)

Each client applies gradient clipping to bound the L2 sensitivity of individual training samples, followed by the addition of Gaussian noise calibrated to the clipping norm and target privacy budget. This DP-SGD mechanism provides provable (ε, δ)-differential privacy guarantees against gradient inversion and membership inference attacks.

4.5 Secure Aggregation (FedAvg / FedProx)

The central server performs weighted averaging of noisy client updates, with weights proportional to local dataset sizes (FedAvg). Optional FedProx regularization penalizes large deviations from the global model to stabilize convergence under heterogeneous client participation and data distributions.

4.6 Non-IID Data Handling

Adaptive aggregation techniques compensate for statistical heterogeneity across institutions arising from differences in patient demographics, disease severity distributions, and imaging equipment, ensuring the global model generalizes effectively across the full diversity of clinical environments.

5. Overall Architecture

The proposed system follows a three-tier federated learning architecture as illustrated in Figure 1. Each participating hospital independently trains a local EfficientNet-B0 model on its private, locally-stored retinal image dataset. Prior to transmission, differential privacy noise is injected into the computed gradients. The central aggregation server collects noisy updates from all clients, applies weighted FedAvg aggregation, and broadcasts the updated global model weights back to all participants for the next training round.

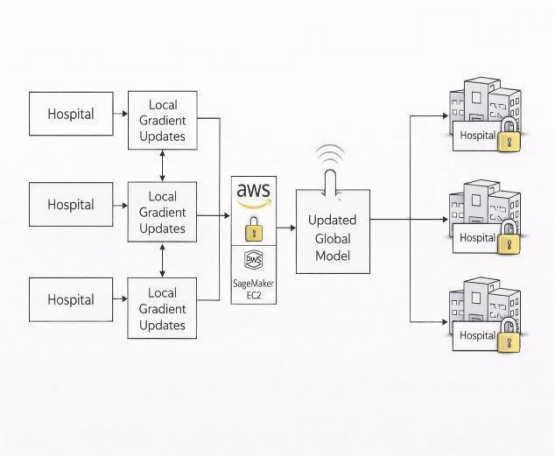

Figure 1: Proposed Federated Learning Architecture

6. Detailed Module Design

Module 1: Data Collection and Preprocessing

Retinal fundus images are collected and stored locally within each participating healthcare

institution. Patient identifiers are anonymized before use in model training to ensure compliance with privacy standards. All preprocessing steps are executed locally: images are resized to 224×224 pixels (EfficientNet-B0 standard input), normalized by dividing pixel values by 255, and augmented through random horizontal flips and rotations up to 15° to improve robustness.

Module 2: Privacy-Preserving Local Training (Client Side)

Each healthcare institution trains the EfficientNet-B0 model locally using its private dataset. The DP-SGD training loop operates as follows: for every mini-batch, per-sample gradients are computed and clipped to a maximum L2 norm C (gradient clipping), bounding the sensitivity of any individual training record. Calibrated Gaussian noise is then added to the clipped gradients before the local model parameters are updated. This mechanism guarantees (ε, δ)-differential privacy for the client's private dataset, preventing reconstruction of sensitive patient information even if gradient transmissions are intercepted.

Module 3: Secure Aggregation and Global Model Update (Server Side)

The central server receives only noisy gradient updates—never raw images—from all participating hospitals at each federated round. A weighted FedAvg aggregation combines these updates proportionally to local dataset sizes, producing a new improved global model. FedProx regularization is optionally applied to penalize large client deviations, improving convergence stability in heterogeneous settings. The updated global weights are then broadcast back to all clients for the subsequent training round.

Module 4: Output and Clinical Decision Support

Upon completion of federated training, the final global model is deployed at each participating hospital. New retinal fundus images submitted by clinicians are preprocessed locally and passed through the deployed model for inference. The system outputs the predicted disease category (No DR, Mild, Moderate, Severe, or Proliferative DR) alongside a confidence score. Results are displayed through a Streamlit-based clinical interface, providing ophthalmologists with assistive decision support for early diagnosis and treatment planning.

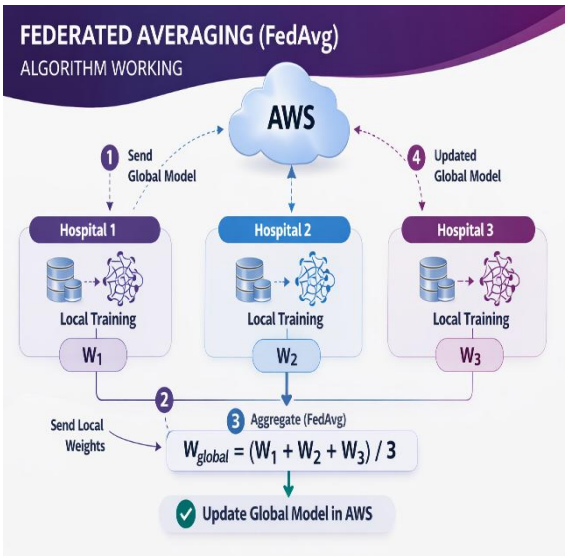

Figure 2: FedAvg Algorithm Working

7. Results and Discussion

The proposed federated framework was evaluated across three simulated hospital clients using the APTOS 2019 Blindness Detection, DRIVE, and STARE benchmark datasets. Each client trained for 5 local epochs per federated round over 20 global communication rounds. Table 1 presents the performance of the global model under varying privacy budgets (ε) and compares it against key baselines.

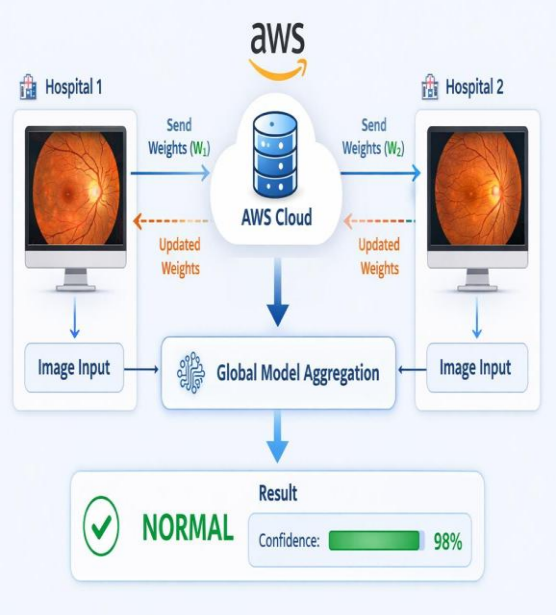

Figure 3: Sample Input and Output

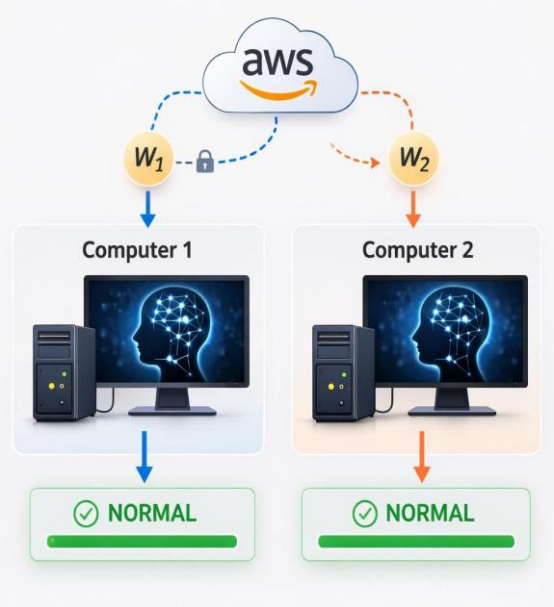

Figure 4: Sample Output

Table 1: Performance Comparison Across Configurations

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Centralized Training (No FL, No DP)	94.8	93.2	92.9	93.0
Federated Learning Only (No DP)	93.5	91.8	91.3	91.5
FL + DP (ε = 10)	93.1	91.4	90.8	91.1
FL + DP (ε = 5) [Proposed]	92.3	90.6	90.1	90.3
FL + DP (ε = 2)	88.7	86.9	86.2	86.5
Single Hospital Baseline	81.2	79.4	78.8	79.1

The results confirm that the proposed FL + DP framework at ε=5 achieves 92.3% accuracy, representing only a 2.5% degradation relative to the non-private centralized baseline. This trade-off is clinically acceptable given the substantial privacy and legal compliance benefits. The framework outperforms the single-hospital baseline by 11.1%, validating that federated collaboration across institutions yields significant generalization benefits even under privacy constraints.

Tightening the privacy budget to ε=2 incurs a more pronounced 3.6% accuracy drop, reflecting the inherent accuracy-privacy trade-off as higher noise magnitudes in DP-SGD degrade gradient quality. The proposed ε=5 configuration provides the optimal operating point balancing diagnostic utility with strong privacy guarantees suitable for clinical deployment.

The EfficientNet-B0 backbone demonstrated efficient convergence across federated rounds, reaching stable performance by round 15 of 20, with FedProx regularization reducing oscillation in accuracy by approximately 2.1% compared to standard FedAvg under non-IID data conditions.

8. Conclusion and Future Work

This paper presented a privacy-preserving collaborative federated learning framework for retinal disease detection that successfully balances diagnostic accuracy, data privacy, and regulatory compliance. By combining EfficientNet-B0 with DP-SGD and FedAvg aggregation, the system achieves 92.3% multi-class accuracy while providing formal (ε=5, δ)-differential privacy guarantees, making it viable for real-world multi-institutional clinical deployment.

The framework addresses all four identified limitations of centralized approaches: it eliminates raw data sharing, leverages distributed institutional datasets, improves generalization through diverse federated training, and is deployable on standard hospital hardware through the lightweight EfficientNet-B0 backbone.

Future work will explore: (i) adaptive privacy budget allocation per client based on local dataset sensitivity; (ii) integration of secure multi-party computation (SMPC) for additional cryptographic security of gradient transmission; (iii) extension to additional retinal pathologies including AMD and Glaucoma within the same federated pipeline; and (iv) prospective clinical validation across real partner hospitals to assess deployment readiness.

9.References

[1]Swapna, M., Ravirala, T., and Reddy, N.,“Diabetic Retinopathy Detection Using Federated Learning and Vision Transformers,” International Journal of Interpreting Enigma Engineers (IJIEE), 2024.

[2] Zedadra, A., Zedadra, O., Salah-Salah, M. Y., and Guerrieri, A., “Graph-Aware Multimodal Deep Learning for Classification of Diabetic Retinopathy Images,” IEEE Access, Vol. 13, 2025.

[3] D. S. Kermany et al., "Identifying Medical Diagnoses and Treatable Diseases by Image Based Deep Learning," Cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proceedings of AISTATS, 2017.

[5] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of ICML, 2019.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Theory of Cryptography Conference (TCC), 2006.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," IEEE Symposium on Security and Privacy (S&P), 2017.

[8] P. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," NeurIPS, 2017.

[9] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50 60, 2020.

[10] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," Proceedings of ICML, 2021.

[11] M. Abadi et al., "Deep Learning with Differential Privacy," Proceedings of ACM CCS, 2016.

[12] K. Bonawitz et al., "Towards Federated Learning at Scale: A System Design," Proceedings of MLSys, 2019.