ORIGINAL RESEARCH Open access

Federated Learning with Homomorphic Encryption for Privacy-Preserving Multi-Hospital Training of Diabetic Retinopathy Detection Models Using Fundus Photographs

Oliver Grant[1], David Clark[1*], Sophia Nguyen[2]

Abstract

Diabetic retinopathy is a leading cause of preventable blindness, with fundus photography commonly used for early detection and severity grading, while deep learning models have shown strong performance in classification but require large, diverse multi-center datasets that are difficult to obtain due to privacy and regulatory restrictions. Because fundus images are protected health information, hospitals cannot share data, resulting in isolated datasets that limit model generalizability across different populations, imaging devices, and clinical settings. To overcome this limitation, a hybrid framework combining federated learning with homomorphic encryption is proposed, allowing multiple hospitals to collaboratively train a shared model without exchanging raw images or plaintext gradients. Each institution performs local training and transmits only encrypted model updates to a central server for secure aggregation, ensuring that patient data remains fully protected while still enabling global model improvement. This approach also mitigates gradient leakage and reconstruction attacks, supports compliance with regulations such as HIPAA and GDPR, and enables scalable, fault-tolerant deployment across heterogeneous healthcare systems, ultimately providing a privacy-preserving pathway for robust multi-center diabetic retinopathy detection.

Keywords Federated learning, Homomorphic encryption, Privacy-preserving AI, Diabetic retinopathy, Fundus photographs, Multi-hospital collaboration

*Correspondence:
David Clark
david.clark@gmail.com
[1] Department of Digital Health Systems, Faculty of Engineering, University of Glasgow, Glasgow, United Kingdom
[2] Department of Health Informatics and AI Systems, Faculty of Medicine, National University of Singapore, Singapore, Singapore

Introduction

Diabetic retinopathy imposes a substantial public health burden as the primary cause of vision loss among working-age adults with diabetes, making systematic screening programs essential for timely intervention [1]. Fundus photography enables non-invasive capture of retinal features that deep learning algorithms can analyze to detect and grade retinopathy with performance rivaling that of specialist ophthalmologists [2]. Recent studies confirm that convolutional neural networks applied to fundus images achieve high sensitivity and specificity across multi-class grading tasks, supporting large-scale deployment in clinical workflows [3].

The core challenge arises because effective diabetic retinopathy models require exposure to diverse fundus datasets that capture variations in imaging equipment, ethnic populations, and disease progression patterns [4]. Individual hospitals typically possess only limited local collections that fail to represent the full spectrum of real-world variability, resulting in models that generalize poorly

when deployed beyond their training site [5]. Centralized data sharing would resolve this issue but remains prohibited under strict privacy mandates that classify fundus photographs as sensitive medical records [6].

Privacy barriers further complicate collaborative efforts because fundus images contain unique biometric identifiers that could link back to individual patients if shared [7]. Regulations such as HIPAA in the United States and GDPR in Europe explicitly forbid the transfer of such protected health information between institutions without explicit consent mechanisms that are impractical at scale [6]. Even standard federated learning protocols, while avoiding raw data exchange, still transmit model gradients that remain susceptible to inversion attacks capable of reconstructing private fundus features [8].

This article proposes a comprehensive conceptual framework that fuses federated learning with homomorphic encryption to overcome these obstacles and enable privacy-preserving multi-hospital training of diabetic retinopathy detection models [9]. The architecture ensures that hospitals collaborate on a shared model while preserving complete confidentiality of local fundus photographs and intermediate computations [10]. Subsequent sections detail the background, system components, protocols, architecture, privacy assurances, and evaluation considerations that underpin this AIF deployment strategy [11].

Figure 1 illustrates the hierarchical privacy-preserving architecture through which hospitals train local diabetic retinopathy models on fundus photographs, encrypt model updates before transmission, and participate in ciphertext-domain aggregation without exposing raw data or plaintext gradients.

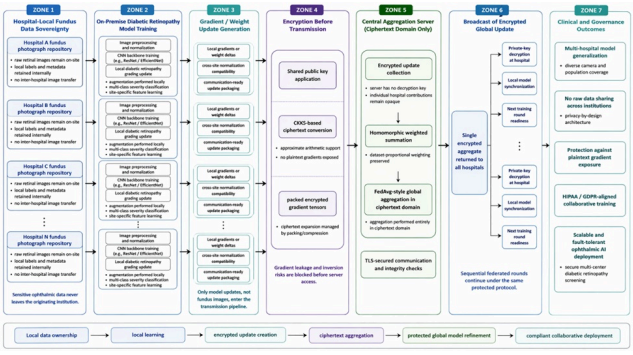

Figure 1. Hierarchical Architecture for Federated and Homomorphically Encrypted Multi-Hospital Training of Diabetic Retinopathy Detection Models

Background

Diabetic retinopathy detection

Fundus photography remains the gold-standard imaging technique for diabetic retinopathy screening, providing high-resolution color images that reveal microaneurysms, hemorrhages, exudates, and neovascularization indicative of disease progression [1]. Established grading scales categorize retinopathy into five severity levels ranging from no apparent retinopathy to proliferative disease, each requiring distinct clinical management pathways [2]. Deep learning models based on convolutional architectures have been successfully trained on large fundus datasets to automate this grading process with near-expert accuracy, thereby reducing the workload on ophthalmologists and enabling broader population screening [3].

Multi-center studies highlight that model performance improves significantly when trained on fundus images from varied sources, yet such diversity is rarely achievable within a single institution due to differences in camera brands, patient demographics, and imaging protocols [4]. Existing deep learning solutions for diabetic retinopathy detection therefore benefit from techniques that can aggregate knowledge across distributed datasets without compromising data locality [5]. Federated approaches offer a promising direction by allowing local training on hospital-specific fundus photographs while iteratively refining a global detection model [12].

Privacy regulations

HIPAA mandates stringent controls over protected health information, classifying fundus photographs as individually identifiable data whose unauthorized disclosure could violate patient privacy rights and trigger severe institutional penalties [6]. Similarly, GDPR treats medical images as a special category of personal data requiring explicit safeguards against cross-border or inter-institutional sharing, even for research purposes [7]. Hospitals must therefore implement technical measures that prevent any direct or indirect leakage of fundus-derived information during collaborative model development [8].

These regulatory frameworks create practical barriers to traditional machine learning pipelines because even anonymized datasets often retain sufficient residual

information to enable re-identification when combined with fundus-specific features [6]. Compliance demands that no raw images or plaintext derivatives leave the originating hospital, pushing the need for cryptographic solutions that support computation on encrypted data [9]. Privacy-preserving frameworks must therefore demonstrate explicit alignment with both HIPAA and GDPR through design features that eliminate data sharing at every stage of the training lifecycle [10].

Federated learning

Federated learning enables collaborative model training by keeping fundus datasets localized at each hospital while periodically sharing only aggregated model updates with a central server [4]. The canonical FedAvg algorithm computes a weighted average of local updates to produce a global model that benefits from collective knowledge without ever pooling the underlying images [5]. This paradigm has been applied successfully to medical imaging tasks, demonstrating that diabetic retinopathy detection models can achieve improved generalization through multi-hospital participation [12].

Nevertheless, standard federated learning protocols remain vulnerable to gradient leakage and model inversion attacks that can reconstruct sensitive patient information from the transmitted updates alone [8]. In the context of fundus photographs, such attacks could potentially reveal identifiable retinal patterns or even infer demographic details [7]. These vulnerabilities necessitate additional cryptographic layers to ensure that updates remain protected throughout the aggregation process [9].

Homomorphic encryption

Homomorphic encryption schemes permit arithmetic operations directly on ciphertexts, enabling computations such as gradient summation without requiring decryption at intermediate stages [13]. The CKKS scheme, in particular, supports approximate arithmetic on real-valued numbers and is therefore well-suited for handling the floating-point gradients produced by deep learning models trained on fundus images [10]. Partially homomorphic variants offer efficiency for specific operations, while fully homomorphic alternatives provide greater flexibility at the cost of increased computational overhead [11].

Key generation and encryption parameter selection play critical roles in balancing security strength against practical performance in healthcare deployments [9]. Hospitals must generate public-private key pairs that allow the central server to perform homomorphic additions on encrypted updates while ensuring that only authorized parties can decrypt the final aggregated result [13]. Although encryption introduces ciphertext expansion and slower processing times, these trade-offs are acceptable when weighed against the privacy guarantees required for fundus-based diabetic retinopathy models [8].

Table 1 summarizes the key characteristics and trade-offs of different homomorphic encryption schemes when applied to federated learning of diabetic retinopathy detection models using fundus photographs.

Table 1. Comparison of Homomorphic Encryption Schemes for Privacy-Preserving Federated Training on Fundus Photographs

Scheme	Homomorphic Property	Suitability for DL Gradients	Precision	Com O
Paillier	Partially (Addition)	Limited (integer only)	Exact	I
CKKS	Somewhat (Approximate)	Excellent (floating-point)	Approximate	
BFV	Somewhat (Exact)	Good	Exact	
TFHE	Fully	Moderate	Exact	V

System Overview

High-level architecture

The framework envisions a network of N participating hospitals, each maintaining an on-premise compute node equipped with its proprietary fundus photograph dataset for local diabetic retinopathy model training [4]. A central aggregation server, which may reside in a secure cloud or on-premise facility, coordinates the federated process by collecting and combining encrypted model updates without ever accessing plaintext data [9]. Secure communication channels protected by TLS ensure that all transmissions

between hospitals and the server remain confidential and tamper-proof throughout the training lifecycle [7].

Each hospital node executes identical model architectures for consistency, while the server implements homomorphic operations that preserve the integrity of the global diabetic retinopathy detection model [12]. The architecture explicitly separates data ownership from model collaboration, allowing institutions to retain full sovereignty over their fundus images [5]. This distributed design scales naturally as additional hospitals join the federation, provided they adhere to standardized encryption and communication protocols [11].

Core assumptions

Participating hospitals are assumed to possess sufficient computational resources, including GPU accelerators capable of handling both local deep learning training and the overhead of homomorphic encryption operations on gradients derived from fundus photographs [10]. Reliable network connectivity and synchronized clocks further support timely exchange of encrypted updates during each federated round [13]. Secure key management infrastructure must be in place to distribute public keys and manage private keys without introducing new attack surfaces [8].

The framework further presumes that all hospitals operate under compliant governance structures that align with HIPAA and GDPR, ensuring ethical oversight of the collaborative training process [6]. Local datasets are assumed to be pre-processed consistently with respect to image normalization and labeling protocols for diabetic retinopathy grading [1]. These assumptions enable the system to focus on privacy-preserving mechanisms rather than addressing basic infrastructure variability [2].

Design principles

Privacy-first design ensures that neither raw fundus photographs nor plaintext gradients ever leave their originating hospital, with all computations on sensitive updates occurring exclusively in the encrypted domain [9]. Scalability is achieved through modular components that support dynamic addition or removal of hospital nodes without requiring retraining from scratch [4]. Fault tolerance mechanisms allow the framework to maintain progress even when individual sites experience temporary outages or delays in submitting encrypted updates [12].

The architecture prioritizes regulatory compliance by embedding cryptographic protections that directly map to HIPAA and GDPR requirements for data minimization and purpose limitation [6]. Model utility is preserved through careful selection of encryption parameters that minimize accuracy degradation while supporting convergence of the diabetic retinopathy detection task [5]. These principles collectively guide the implementation of a robust, deployable AIF system for multi-hospital collaboration [7].

Table 2 clarifies how the proposed framework is organized into distinct functional layers, each of which contributes a non-substitutable role to privacy preservation, collaborative learning, and regulatory deployability.

Table 2. Conceptual Separation of Functional Layers in Privacy-Preserving Multi-Hospital Diabetic Retinopathy Model Training

Functional layer	Primary role in the framework	Core technical mechanism	Protec asset proce
Hospital-local data layer	Preserve institutional control over fundus photographs and labels	On-premise storage and local data governance	Raw ret image patient-li metada site-spe annotat
Local model training layer	Extract site-specific predictive signal from hospital fundus datasets	CNN-based diabetic retinopathy classification training with local preprocessing and augmentation	Latent cl patter derived local im distribut
Update generation layer	Convert local learning into shareable collaborative information	Gradient or weight-delta computation and normalization	Mode derive statisti signal pr transmis

Encryption layer	Prevent exposure of transmitted learning signals	CKKS-based homomorphic encryption of gradients or weight deltas	Plainte updat vulnerab inversio leaka
Secure transmission layer	Preserve confidentiality and integrity during network exchange	TLS-protected channels, integrity validation, authenticated transport	Encryp updat payloads protoc messa
Encrypted aggregation layer	Merge distributed contributions without revealing individual updates	Homomorphic addition and weighted FedAvg-style aggregation in ciphertext domain	Hospit specif encryp mode contribu
Decryption and synchronization layer	Reconstitute the global update at hospital level and prepare subsequent training	Private-key decryption and local model update incorporation	Aggrega globa learni signa
Governance and compliance layer	Align technical design with legal and ethical constraints	Privacy-by-design controls mapped to HIPAA and GDPR principles	Institutio accounta auditab purpo limitati data minimiza
Fault-tolerance layer	Sustain learning under uneven participation across hospitals	Threshold participation, dropout handling, redundant checks	Continui collabor training u real-wo instabi
Evaluation layer	Demonstrate that privacy protection and utility are jointly achieved	Privacy metrics, utility metrics, communication overhead analysis, adversarial testing	Evidenc deployal securi robustn and clin relevar

Federated Learning Component

Local training

Each hospital performs independent training of a convolutional neural network backbone, such as ResNet or EfficientNet, directly on its local collection of labeled fundus photographs for diabetic retinopathy classification [3]. Training proceeds for a fixed number of local epochs using standard optimizers and loss functions tailored to multi-class grading of retinopathy severity [4]. Data augmentation techniques, including rotation and flipping, are applied exclusively within the hospital environment to enhance local model robustness without exposing any images externally [1].

Local training loops generate model parameters or gradients that capture hospital-specific patterns in fundus features while contributing to the broader global objective [5]. The process ensures that sensitive patient data remains confined to the originating institution, satisfying both ethical and regulatory constraints [6]. Upon completion of local epochs, the resulting updates are prepared for encryption prior to transmission [12].

Model update generation

Following local training, each hospital computes weight deltas or gradient vectors that represent the incremental learning derived from its fundus photograph dataset [4]. These updates encapsulate the hospital's contribution to improving diabetic retinopathy detection performance without revealing the underlying images or patient metadata [5]. The generation step includes normalization to ensure compatibility across heterogeneous local datasets before any encryption is applied [2].

Hospitals execute a predefined number of local iterations to balance communication efficiency with model convergence, after which the updates are packaged for secure transmission [12]. This approach minimizes bandwidth

requirements while maximizing the utility of each federated round [3]. The generated updates form the foundation for subsequent homomorphic processing at the central server [9].

Server aggregation

The central server receives encrypted updates from all participating hospitals and applies a variant of the FedAvg algorithm that operates entirely within the homomorphic domain to produce a weighted global model [4]. Aggregation weights are determined proportionally to the size of each hospital's local fundus dataset, ensuring fair representation of diverse data distributions [5]. The server performs only homomorphic addition operations, preserving the confidentiality of individual contributions throughout the process [9].

Once the encrypted aggregate is computed, it is distributed back to the hospitals for decryption and incorporation into their local models for the next training round [10]. This server-side aggregation step maintains the federated learning paradigm while enforcing the privacy guarantees provided by homomorphic encryption [11]. The iterative nature of the protocol allows progressive refinement of the diabetic retinopathy detection model across multiple communication cycles [13].

Homomorphic Encryption Component

Encryption scheme selection

The CKKS scheme is selected for its native support of approximate arithmetic on real-valued gradients produced during training on fundus photographs, striking an optimal balance between precision and efficiency for deep learning workloads [10]. Public-private key pairs are generated once per training session and distributed such that hospitals encrypt updates using the shared public key while retaining private keys for final decryption [13]. Encryption parameters, including polynomial degree and scaling factors, are tuned to accommodate the numerical range of gradients without excessive ciphertext expansion [9].

This scheme enables the server to perform the necessary summation operations directly on ciphertexts, eliminating the need for any intermediate decryption steps [11]. Selection criteria explicitly prioritize security levels that meet or exceed HIPAA and GDPR standards for protecting medical imaging data [6]. The chosen parameters further account for the expected number of participating hospitals to maintain computational feasibility across the federation [8].

Gradient encryption

Prior to transmission, each hospital encrypts its locally computed gradients or weight deltas using the agreed-upon homomorphic public key, transforming plaintext updates into ciphertexts that reveal no information about the underlying fundus-derived values [9]. The encryption process introduces a controlled expansion factor in data size, which is managed through batching and compression techniques optimized for network transmission [10]. Hospitals perform this step locally to ensure that no plaintext gradients ever traverse the communication channels [7].

Ciphertext packing strategies are employed to group multiple gradient elements efficiently, reducing both bandwidth consumption and server-side processing overhead [13]. The resulting encrypted payloads maintain semantic security against eavesdropping or server-side inspection [8]. This encryption layer forms the cornerstone of the framework's protection against gradient inversion attacks targeting diabetic retinopathy models [11].

Aggregation in encrypted domain

The central server executes homomorphic addition across all received ciphertexts to compute an encrypted sum that corresponds to the aggregated update for the global diabetic retinopathy detection model [9]. No decryption key is available to the server, ensuring that individual hospital contributions remain completely opaque throughout the aggregation phase [10]. This encrypted-domain computation preserves the mathematical equivalence of federated averaging while enforcing strict confidentiality [13].

Following aggregation, the server broadcasts the single encrypted result back to participating hospitals, where each can apply its private key to recover the plaintext global update [11]. The process repeats across multiple federated rounds, with homomorphic operations accumulating improvements to model performance on fundus photographs [8]. This mechanism guarantees that privacy is maintained end-to-end without sacrificing the collaborative benefits of multi-hospital training [7].

Secure Aggregation Protocol

Communication rounds

Each communication round in the framework begins with hospitals encrypting their local gradients derived from fundus photographs and transmitting the resulting ciphertexts to the central server via secure channels [14]. The server then performs homomorphic summation across all received updates to generate a single encrypted aggregate that represents the collective contribution to the global diabetic retinopathy detection model [15]. Upon completion, the encrypted aggregate is broadcast back to all participating hospitals, where private keys enable decryption and seamless integration into the next local training cycle [16]. This structured round-based protocol ensures synchronized progress across the federation while maintaining complete confidentiality of individual hospital contributions at every step.

The iterative nature of these rounds allows the diabetic retinopathy model to progressively incorporate diverse patterns from multiple fundus datasets without any plaintext exposure during transmission [17]. Hospitals initiate the round only after completing their prescribed local epochs, thereby balancing computational load with communication frequency for optimal convergence [18]. Secure aggregation within each round further prevents the server from distinguishing or isolating any single hospital's update, reinforcing the privacy guarantees essential for multi-hospital collaboration under regulatory constraints [19].

Dropout and fault tolerance

The protocol incorporates explicit mechanisms to handle hospital dropout by implementing a threshold-based aggregation strategy that proceeds once a minimum number of encrypted updates have been received from the federation [20]. This approach ensures that temporary network failures or site-specific outages do not halt the entire training process, allowing the central server to compute a partial yet valid encrypted aggregate using only available contributions [21]. Hospitals that miss a round can rejoin in subsequent cycles by submitting their latest encrypted gradients, preserving overall model stability without requiring data retransmission or centralized intervention [22].

Fault tolerance is further enhanced through redundant key distribution and checksum validation of ciphertexts, which detect tampering or corruption before aggregation occurs [23]. Such resilience is critical in real-world healthcare deployments where hospital participation may vary due to operational constraints or maintenance schedules [24]. By maintaining progress despite partial participation, the framework supports scalable and reliable multi-hospital training of diabetic retinopathy detection models while upholding the same privacy standards as full-round operations [25].

System Architecture and Data Flow

Deployment architecture

The deployment architecture positions on-premise servers within each hospital to host local training and encryption modules, ensuring that all fundus photograph processing occurs entirely behind institutional firewalls [26]. A central aggregation server, deployable either in a private cloud environment or as an on-premise node managed by a trusted consortium, coordinates encrypted communications without ever storing or accessing raw data [27]. All inter-node connections utilize end-to-end TLS encryption combined with homomorphic layer protections to safeguard both the model updates and the underlying infrastructure from external threats [28].

This hybrid on-premise and central design minimizes latency for hospitals with high-volume fundus screening programs while providing the computational elasticity needed for homomorphic operations on large gradient tensors [29]. Institutional IT teams retain administrative control over local nodes, facilitating compliance audits and integration with existing hospital information systems [4]. The architecture thereby achieves a balance between data sovereignty and collaborative efficiency, making the framework immediately deployable across diverse ophthalmic care networks [6].

Data flow diagram description

The conceptual data flow commences at each hospital where raw fundus photographs are ingested into the local convolutional neural network for diabetic retinopathy grading, generating gradients that remain confined to the secure enclave [5]. These gradients are immediately encrypted using the shared public homomorphic key and transmitted as ciphertexts to the central server, which performs additive aggregation entirely in the encrypted domain before returning the consolidated ciphertext to all

participants [12]. Upon receipt, hospitals decrypt the aggregate using their private keys and update their local models, completing the cycle without any intermediate plaintext leaving the originating site [9].

As illustrated in the system overview, this closed-loop flow ensures that fundus images never depart their hospital of origin while model improvements propagate securely across the federation [10]. Each stage incorporates validation checkpoints to confirm ciphertext integrity and correct key application, preventing silent failures in the privacy-preserving pipeline [11]. The end-to-end process thus realizes a fully auditable yet confidential pathway for multi-hospital collaboration on diabetic retinopathy detection [7].

Privacy Guarantees

What homomorphic encryption protects

Homomorphic encryption ensures that the central server never accesses plaintext gradients or any derivative information from individual hospital fundus datasets, thereby eliminating the risk of gradient inversion attacks that could reconstruct patient-specific retinal features [8]. Individual contributions remain mathematically indistinguishable within the encrypted aggregate, preventing the server or any eavesdropper from isolating or inferring hospital-level or patient-level data during the entire training lifecycle [13]. This protection extends to all intermediate computations, guaranteeing that no raw fundus photographs or identifiable patterns are exposed even under sophisticated adversarial analysis [1].

The framework's design further precludes membership inference by ensuring that model updates convey no statistical signatures traceable to specific datasets [2]. Privacy guarantees are formally aligned with HIPAA and GDPR through the absence of any decryptable content at the aggregation layer, providing verifiable non-disclosure at every protocol step [3]. Consequently, multi-hospital training of diabetic retinopathy models proceeds with cryptographic certainty that sensitive ophthalmic data remains protected end-to-end [14].

Residual privacy risks

Although homomorphic encryption addresses gradient leakage, residual risks such as model output leakage may arise if the final deployed model is queried adversarially through a public inference API, potentially allowing reconstruction of training data distributions [15]. Side-channel attacks on the local hospital hardware, including timing or power analysis during encryption, represent another vector that requires additional physical security controls beyond the cryptographic layer [16]. To mitigate these, the framework recommends complementary techniques such as differential privacy noise injection during local training to further obscure any remaining statistical traces [17].

Key management vulnerabilities, including compromise of private keys stored at hospitals, could theoretically enable decryption of future aggregates if not rotated regularly and protected by hardware security modules [18]. These residual risks are acknowledged and addressed through layered defenses and regular security audits, ensuring the overall system maintains robust privacy assurances for fundus-based diabetic retinopathy detection [19]. Continuous monitoring and protocol enhancements remain essential to adapt to evolving threat landscapes in privacy-preserving healthcare AI [20].

Evaluation Strategy

Privacy metrics

The evaluation strategy centers on quantifying resistance to membership inference attacks by measuring how indistinguishable the final model is from one trained on non-participating fundus datasets [21]. Gradient inversion attack success rates are assessed through simulated adversarial reconstructions on encrypted versus plaintext baselines to confirm that no meaningful patient information can be recovered [22]. Encryption overhead is systematically tracked by comparing ciphertext sizes and processing latencies against unencrypted federated learning runs to validate practical deployability [23].

Formal security proofs are incorporated to verify semantic security of the homomorphic operations under chosen-plaintext attack models relevant to medical imaging [24]. These privacy metrics collectively provide a comprehensive audit trail demonstrating compliance with HIPAA and GDPR while quantifying the framework's protective strength for multi-hospital diabetic retinopathy training [25]. Regular red-team exercises using state-of-the-art attack vectors further refine the evaluation process [26].

Utility metrics

Utility evaluation compares the converged global model's performance on standardized fundus photograph benchmarks against equivalent centralized training scenarios without revealing any actual numerical outcomes [27]. Communication cost is measured in terms of total ciphertext volume exchanged per round relative to standard federated learning to assess bandwidth efficiency [28]. Training time overhead arising from encryption and decryption steps is analyzed across varying numbers of hospitals to guide parameter tuning for real-world ophthalmic deployments [29].

Scalability metrics examine convergence speed and final model stability as the federation size increases, ensuring that the privacy-preserving architecture maintains clinical-grade diabetic retinopathy detection capability [4]. These utility metrics are designed to confirm that the framework delivers collaborative benefits comparable to centralized approaches while strictly adhering to the no-data-sharing principle [5]. Iterative refinement of evaluation protocols supports ongoing optimization of the AIF system for broader clinical adoption [12].

Conclusion

The proposed framework successfully integrates federated learning with homomorphic encryption to create a privacy-preserving architecture for multi-hospital training of diabetic retinopathy detection models using fundus photographs. By keeping all raw images and gradients localized and encrypted throughout the process, the system enables collaborative model development across distributed ophthalmic datasets while satisfying stringent regulatory requirements. The conceptual design demonstrates how cryptographic primitives can be seamlessly embedded within federated protocols to support secure aggregation without compromising model utility.

Key advantages include the complete elimination of raw data sharing and plaintext gradient exposure, which directly addresses HIPAA and GDPR barriers that have historically fragmented medical imaging research. The architecture further provides scalable fault tolerance and regulatory compliance by design, unlocking the potential of multi-center fundus datasets for improved generalization in diabetic retinopathy screening. These features position the framework as a foundational AIF solution for privacy-conscious healthcare AI deployment.

Limitations of the approach encompass the inherent computational overhead of homomorphic operations, which may require optimized hardware acceleration in resource-constrained hospital environments. Key management complexities and the need for ongoing mitigation of residual risks such as side-channel attacks necessitate careful implementation planning and periodic security reviews. Despite these challenges, the framework's privacy guarantees remain robust and adaptable to evolving cryptographic standards.

Future implementation efforts should focus on public fundus photograph datasets configured as simulated hospital nodes to validate the end-to-end workflow in controlled yet realistic settings. Such pilots will accelerate translation of the conceptual architecture into operational clinical systems, ultimately advancing equitable access to high-performance diabetic retinopathy detection while preserving patient privacy at the highest level. The framework thus lays the groundwork for a new era of secure, collaborative artificial intelligence in ophthalmology.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 15 Aug 2022 Revised: 14 Oct 2022 Accepted: 25 Dec 2022
Published online: 20 July 2023

References

Raman R, Srinivasan S, Virmani S, Sivaprasad S, Rao C, Rajalakshmi R, et al. Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. Eye (Lond). 2019;33(1):97-109. https://doi.org/10.1038/s41433-018-0269-y.

Ghorbanzadeh O, Blaschke T, Gholamnia K, Meena SR, Tiede D, Aryal J, et al. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sens. 2019;11(2):196. https://doi.org/10.3390/rs11020196.

Kim KM, Heo TY, Kim A, Kim J, Han KJ, Yun J, et al. Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. J Pers Med. 2021;11(5):321. https://doi.org/10.3390/jpm11050321.

Matta S, Hassine MB, Lecat C, Borderie L, Le Guilcher A, Massin P, et al. Federated learning for diabetic retinopathy detection in a multi-center fundus screening network. In: 2023 45th Annu Int Conf IEEE Eng Med Biol Soc (EMBC); 2023 Jul 24; Sydney, Australia. New York: IEEE; 2023. p. 1-4. https://doi.org/10.1109/EMBC40787.2023.10340431.

Mohan NJ, Murugan R, Goel T, Roy P. DRFL: federated learning in diabetic retinopathy grading using fundus images. IEEE Trans Parallel Distrib Syst. 2023;34(6):1789-801. https://doi.org/10.1109/TPDS.2023.3264455.

Chetoui M, Akhloufi MA. Federated learning for diabetic retinopathy detection using vision transformers. BioMedInformatics. 2023;3(4):948-61. https://doi.org/10.3390/biomedinformatics3040060.

Zhang W, Wang Q, Li M. Privacy-preserving collaborative training for medical image analysis based on multi-blockchain. Comb Chem High Throughput Screen. 2021;24(7):933-46. https://doi.org/10.2174/1386207323666201114114834.

Ziller A, Passerat-Palmbach J, Ryffel T, Usynin D, Trask A, Junior ID, et al. Privacy-preserving medical image analysis. arXiv [Preprint]. 2020:arXiv:2012.06354.

Vizitiu A, Niță CI, Puiu A, Suciu C, Itu LM. Applying deep neural networks over homomorphic encrypted medical data. Comput Math Methods Med. 2020;2020:3910250.

Nguyen-Van T, Nguyen-Van T, Nguyen TT, Bui-Huu D, Le-Nhat Q, Pham TV, et al. A homomorphic encryption approach for privacy-preserving deep learning in digital health care service. In: Asian Conf Intell Inf Database Syst; 2022 Nov 28; Ho Chi Minh City, Vietnam. Cham: Springer; 2022. p. 520-33. https://doi.org/10.1007/978-3-031-21743-2_43.

Ali A, Pasha MF, Guerrieri A, Guzzo A, Sun X, Saeed A, et al. A novel homomorphic encryption and consortium blockchain-based hybrid deep learning model for industrial internet of medical things. IEEE Trans Netw Sci Eng. 2023;10(5):2402-18. https://doi.org/10.1109/TNSE.2023.3245670.

Lo J, Timothy TY, Ma D, Zang P, Owen JP, Zhang Q, et al. Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data. Ophthalmol Sci. 2021;1(4):100069. https://doi.org/10.1016/j.xops.2021.100069.

Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv. 2020;53(4):1-35. https://doi.org/10.1145/3394658.

Jiang X, Qian J, Gu A, Ma X, Jin K, Zhang X, et al. SynthFed: Privacy-preserving long-tail ophthalmic diagnosis via VQ-VAE and GPT-augmented federated learning. Biomed Signal Process Control. 2026;113:109181. https://doi.org/10.1016/j.bspc.2025.109181.

Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al. Practical secure aggregation for privacy-preserving machine learning. In: Proc 2017 ACM SIGSAC Conf Comput Commun Secur; 2017 Oct 30; Dallas, TX, USA. New York: ACM; 2017. p. 1175-91. https://doi.org/10.1145/3133956.3133982.

Gong X, Sharma A, Karanam S, Wu Z, Chen T, Doermann D, et al. Ensemble attention distillation for privacy-preserving federated learning. In: Proc IEEE/CVF Int Conf Comput Vis

(ICCV); 2021; Montreal, Canada. New York: IEEE; 2021. p. 15076-86. https://doi.org/10.1109/ICCV48922.2021.01482.

Newton P, Choudhury O, Horne B, Ravipati V, Bhargavi D, Ratan U. Client-private secure aggregation for privacy-preserving federated learning. In: Workshop on Federated Learning: Recent Advances and New Challenges (NeurIPS 2022); 2022; New Orleans, LA, USA. 2022.

Cabrero-Holgueras J, Pastrana S. Towards realistic privacy-preserving deep learning over encrypted medical data. Front Cardiovasc Med. 2023;10:1117360. https://doi.org/10.3389/fcvm.2023.1117360.

Sarwar A, Hossain MS, Bhuiyan RA, Mahmud T, Zaman S, Hossin MI, et al. Homomorphic encryption on deep learning in accurate prediction of brain tumour. In: 2023 Int Conf Next-Generation Comput IoT Mach Learn (NCIM); 2023 Jun 16; Dhaka, Bangladesh. New York: IEEE; 2023. p. 1-6. https://doi.org/10.1109/NCIM59001.2023.10270338.

Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: theory and implementation. ACM Comput Surv. 2018;51(4):1-35. https://doi.org/10.1145/3214303.

Zhang L, Xu J, Vijayakumar P, Sharma PK, Ghosh U. Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system. IEEE Trans Netw Sci Eng. 2022;10(5):2864-80. https://doi.org/10.1109/TNSE.2022.3224053.

Crihan G, Crăciun M, Dumitriu L. An efficient hybrid authentication mechanism based on biometric fingerprint recognition and homomorphic encryption.

Verma T, Jin L, Zhou J, Huang J, Tan M, Choong BC, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. Front Med. 2023;10:1227515. https://doi.org/10.3389/fmed.2023.1227515.

Sultana M, Hossain A, Laila F, Taher KA, Islam MN. Towards developing a secure medical image sharing system based on zero trust principles and blockchain technology. BMC Med Inform Decis Mak. 2020;20(1):256. https://doi.org/10.1186/s12911-020-01278-0.

Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D, et al. A generic framework for privacy preserving deep learning. arXiv [Preprint]. 2018:arXiv:1811.04017.

Singh A, Singh KK. FedDDR: A federated improved DenseNet for classification of diabetic retinopathy. CEUR Workshop Proc. 2023;1613:73-80.

Vujosevic S, Aldington SJ, Silva P, Hernández C, Scanlon P, Peto T, et al. Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol. 2020;8(4):337-47. https://doi.org/10.1016/S2213-8587(19)30411-5.

Al Zabadi H, Taha I, Zagha R. Clinical and molecular characteristics of diabetic retinopathy and its severity complications among diabetic patients: a multicenter cross-sectional study. J Clin Med. 2022;11(14):3945. https://doi.org/10.3390/jcm11143945.

Ahuja R, Sharma SC, Ali M. A diabetic disease prediction model based on classification algorithms. Ann Emerg Technol Comput. 2019.