2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA) | 979-8-3503-8220-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/DICTA60407.2023.00020

Byzantine-Resilient Federated Learning Leveraging Confidence Score to Identify Retinal Disease

M Sakib Osman Eshan[1], Md. Naimul Huda Nafi[1], Nazmus Sakib[1] , Mehedi Hasan Emon[1] , Tanzim Reza[1] ,
Mohammad Zavid Parvez[2] , Prabal Datta Barua[3,4] , and Subrata Chakraborty[5]
Department of Computer Science and Engineering, Brac University, Bangladesh[1]
School of Computing, Mathematics and Engineering, Charles Sturt University, Australia[2]
School of Management & Enterprise, University of Southern Queensl, Australia[3]
Cogninet Australia Pty Ltd, Level 5, 29-35 Bellevue St, Surry Hills NSW 2010, Australia[4]
School of Science and Technology, University of New England, Australia[5]
eshan.bd@outlook.com, {md.naimul.huda.nafi, nazmus.sakib6, mehedi.hasan.emon}@g.bracu.ac.bd,
tanzim.reza@bracu.ac.bd, mparvez@csu.edu.au, prabal.barua@usq.edu.au, subrata.chakraborty@une.edu.au

***Abstract*—Federated learning is a distributed machine learning paradigm that enables multiple actors to collaboratively train a common model without sharing their local data, thus addressing data privacy issues, especially in sensitive domains such as healthcare. However, federated learning is vulnerable to poisoning attacks, where malicious (Byzantine) clients can manipulate their local updates to degrade the performance or compromise the privacy of the global model. To mitigate this problem, this paper proposes a novel method that reduces the influence of malicious clients based on their confidence. We evaluate our method on the Retinal OCT dataset consisting of age-related macular degeneration and diabetic macular edema, using InceptionV3 and VGG19 architecture. The proposed technique significantly improves the global model's precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) for both InceptionV3 and VGG19. For InceptionV3, precision rises from 0.869 to 0.906, recall rises from 0.836 to 0.889, and F1 score rises from 0.852 to 0.898. For VGG19, precision rises from 0.958 to 0.963, recall rises from 0.917 to 0.941, and F1 score rises from 0.937 to 0.952.**

***Index Terms*—Computer Vision, Federated learning, Deep Learning, Medical Image Processing, Data poisoning, Retinal OCT**

I. Introduction

Retinal OCT (Optical Coherence Tomography) is a non-invasive imaging technology used to see the layers of the retina, the light-sensitive tissue in the back of the eye. It aids in the diagnosis and treatment of retinal disorders by providing detailed images of the retina in cross-section [1]. Age-related macular degeneration (AMD), diabetic retinopathy, retinal vein occlusion (RVO), macular edema, retinal detachment, epiretinal membrane, and glaucoma can all be diagnosed and followed up on with retinal OCT [2]. With the aid of high-resolution images made possible by Retinal OCTs, ophthalmologists can more accurately evaluate the severity and progression of these conditions, detect them earlier, make accurate diagnoses, and take initiative for treatment [3]. It is essential for any ophthalmologist who wishes to evaluate and treat their patients' retinal health appropriately.

Diseases like glaucoma, retinal vein occlusion (RVO), diabetic retinopathy, and age-related macular degeneration (AMD) can be detected with retinal optical coherence tomography (OCT). Globally, approximately 8.7% of people aged 45 and older have AMD, and approximately one-third of diabetics-affected people develop diabetic retinopathy [4]. Between the ages of 40 and 80, approximately 3.5% of persons worldwide get affected by glaucoma, while the prevalence of RVO is estimated to be 0.6% [4]. When attempting to interpret these prevalence rates, it is essential to account for contextual factors such as location, population, and healthcare accessibility. Global health organizations and reputable medical literature databases are excellent resources for up-to-date and accurate information.

Age-related macular degeneration (AMD) is a leading cause of vision loss in the elderly. Timely care and improved patient outcomes depend on early detection and correct categorization of AMD. This research [6], introduces a novel method for automatic AMD classification from retinal OCT images utilizing a multi-scale convolutional neural network. In the paper, the authors used an approach that leverages the unique features present at different scales in OCT images to improve classification accuracy. This research used 2 datasets and both consist of OCT images. The first dataset is divided into 3 categories and the second into 4 categories [7]. As already mentioned, this paper proposed a Multi-Scale CNN Architecture which can be deployed with any commercially available CNN framework. This includes VGG, ResNet, DenseNet, and others. The authors used VGG16 [8], as the primary network for training which can be combined with the FPN structure and the authors named it FPN-VGG16. Following data training on this CNN framework, the suggested model's efficacy was evaluated using several measures, including accuracy, sensitivity, specificity, and AUC-ROC. In all the evaluation scores, the model performed exceptionally well since the classification rate was over 90% for all the metrics. This shows the model classified AMD disease quite well based on retinal OCT images.

Federated Learning has gained popularity in the larger machine learning community, but it is noticeably underrepresented in the field of medical imaging, particularly OCT.

979-8-3503-8220-4/23/$31.00 ©2023 IEEE
DOI 10.1109/DICTA60407.2023.00020

A resource that hasn't yet been fully utilized for improving diagnostic precision, patient privacy, and cooperative research efforts is suggested by the paucity of literature examining FL's potential in OCT, a non-invasive imaging method important in ophthalmology. This study evaluates the applicability and benefits of FL in the OCT environment in an effort to close this gap. A less invasive method for identifying retinal disorders is optical coherence tomography (OCT). However, the incorporation of FL into OCT holds the prospect of bringing about a number of untapped benefits. These include improved patient privacy due to the localization of raw imaging data and the capacity to collaborate with various medical institutions to create more robust and generalizable models. Additionally, FL might make it possible for OCT models to receive real-time updates, guaranteeing that they continue to be capable of spotting new diseases. This unusual relationship between FL and OCT calls for further investigation. A less invasive method for identifying retinal disorders is optical coherence tomography (OCT). However, the incorporation of FL into OCT holds the prospect of bringing about a number of untapped benefits. These include improved patient privacy due to the localization of raw imaging data and the capacity to collaborate with various medical institutions to create more robust and generalizable models. Additionally, FL might make it possible for OCT models to receive real-time updates, guaranteeing that they continue to be capable of spotting new diseases. This unusual relationship between FL and OCT calls for further investigation.

Optical coherence tomography (OCT) of the retina has become more important as a tool for the detection and follow-up of retinal disorders. Clinical decisions rely heavily on OCT picture segmentation of retinal layers and pathology. Also, retinal OCT image processing is quite a difficult and sensitive task as various kinds of challenges are associated with segmentation due to image artifacts and variability. Therefore, in paper [9], the authors describe a deep learning-based approach to segmenting retinal OCT images. To automatically detect abnormal signs and segment retinal layers efficiently, the method in the paper used convolutional neural networks (CNNs). The U. of Miami OCT dataset [10], was utilized in this study and is available to the public. This article features fifty optical coherence tomography (OCT) images from ten patients with mild, nonproliferative diabetic retinopathy. Transverse resolution is 11.11 m/pixel, and axial resolution is 3.867 m/pixel in each of the 768x496 images. Each patient was examined in five areas: one of them is the central fovea, two of them are the periphery, and two of them are the parafovea. Using the definition of a "surface" as the boundary between two neighboring retinal layers, two expert graders labeled a total of five surfaces per image. There are two main phases in this approach for calculating retinal surfaces. In the first stage, a classification algorithm is used to determine which retinal layer best fits each pixel. These estimates of per-pixel categorization are input into a regression technique that exploits the fact that retinal surfaces can be modeled as smooth functions that partition layers along the axial dimension. The scientists employed a fully convolutional neural network (FCN) with the DenseNet [11,12] architecture to categorize the retinal layer. The authors conducted an evaluation of the model after data training with K-fold cross-validation (K=10).

Without quick diagnosis and treatment, diabetic macular edema (DME) can cause permanent visual loss. This research paper [13], presents a deep learning-based technique that can predict OCT measurements of macular thickening from more readily available color fundus pictures(CFP) in clinical settings. The proposed method utilizes convolutional neural networks (CNNs) to acquire accurate estimations of OCT-based metrics through the learning of complicated representations of retinal features. In the paper, two different datasets were used. The first collection of CFPs includes 12,374 FC-CFPs from 725 patients with corresponding Central Subfield Thickness (CST) measurements, and the second set includes 17,997 CFPs from 753 patients with corresponding Central Foveal Thickness (CFT) measures. Both of these datasets are collected via phase-3 RIDE [14], and RISE [15], (DME) clinical trials. Additionally, both the binary classification and regression tasks were implemented using an inception-V3-based architecture. For training the Inception-V3 models, a transfer-learning cascade function was used. When applying transfer learning, the softmax layer of the architecture is typically replaced and trained for binary classification, whereas the linear layer of the design is typically replaced and trained for regression. Two metrics, AUC and R2, were employed to assess the efficiency of the model. The DL (Deep Learning) models' performance in binary classification was evaluated using the AUC metric, while the DL model's regression accuracy was evaluated using the R2 value. For binary classification, the highest AUC value was 0.97 for CST and 0.91 for CFT which means the DL model was able to correctly predict with 97% accuracy for CST and 91% accuracy for CFT. However, R2 values of 0.74 for CST and 0.54 for CFT were found in the best DL regression model. Overall, the model performed better in the CST measurement category than in the CFT category.

The leading cause of blindness among diabetic-affected patients is diabetic retinopathy (DR). To avoid permanent injury, early diagnosis, and treatment are essential. Combining multiple deep convolutional neural networks (CNNs) to enhance classification performance is the focus of this paper [16], given that it presents a deep learning ensemble approach to DR detection. The proposed method trains five alternative CNN models—Resnet50 [17], Inceptionv3 [18], Xception [19], Dense121 [20], and Dense169 [20]—on a publicly available Kaggle dataset of retina images and then aggregates their predictions to enhance the accuracy of DR diagnosis. There are 35126 color fundus images in the authors' utilized retina dataset, and each image is 3888 pixels wide by 2951 pixels high. The dataset is categorized into 5 classes with 5 different modalities(Normal, Mild, Moderate, Severe, PDR). Performance criteria like accuracy, sensitivity, specificity, precision, F1 measures, and AUC-ROC were utilized to assess the validity of the suggested model. As a whole, the proposed model outperformed the state-of-the-art by a wide margin. For example, for class 0, the proposed model had an accuracy of 97%, 84%, 40%, and 90% for recall, precision, specificity, and F1-score respectively. In contrast, an older existing model [21] had an accuracy of 95%, 78%, 19%, and 85% for recall, precision, specificity, and F1-score respectively which means the proposed model predicted much more accurately than the older model. This is true for all other classes too. Overall, this research proves the superiority of the ensemble model compared to individual models, demonstrating its potential to improve DR diagnostic precision and reliability.

To enhance the usefulness of deep learning (DL) in the diagnosis of rare ocular disorders from optical coherence tomography (OCT) pictures, the authors of this research [22] employed few-shot learning (FSL) in tandem with a generative adversarial network (GAN). Using regular OCT pictures and CycleGAN models, they fabricated pathological images of uncommon disorders. Using an enhanced training dataset and an external test dataset, the authors trained a DL model based on the Inception v3 architecture. The authors found that the CycleGAN-generated synthetic images helped in extracting the unique features of each uncommon disease, leading to a more precise diagnosis. The accuracy of OCT diagnosis for rare retinal illnesses was greatly enhanced by the suggested DL model compared to traditional DL models like the Siamese network and prototype network. Several measures were utilized to assess the performance of the DL model. These included overall accuracy, Cohen's, RCI, and the Matthews correlation coefficient. To prove its efficacy, the proposed DL model achieved an overall accuracy of 93.9% when identifying uncommon retinal disorders. The model outperformed other FSL methods, including the Siamese network and the prototype network, when it came to multiclass categorization. Additionally, the DL model that utilized GAN-based data augmentation outperformed baseline models in the detection of uncommon diseases. Using the Grad-CAM technique, the authors created saliency maps to visualize the pathological characteristics of the predicted evidence. In addition, the authors investigated the applicability of their method to a segmentation task utilizing the CycleGAN model. They successfully segmented diseased features in central serous chorioretinopathy and macular telangiectasia using manually segmented OCT pictures and normal images to train the model. Overall, the OCT diagnostic accuracy for rare retinal disorders was enhanced by the proposed DL model when combined with GAN-based data augmentation. This strategy can aid clinicians in preventing the neglect of rare diseases, reducing diagnostic delays, and easing patient burden.

A thorough exploration into the use of Federated Learning (FL) in medical image processing is presented in the research paper [29]. The writers thoroughly examined 17 pertinent studies, dissecting the techniques used and the outcomes that followed. They discovered that FL has been successfully used in a number of medical fields, including the study of diabetic retinopathy, the detection of COVID-19, and cancer diagnostics, among others. The majority of the publications under evaluation trained local FL models using convolutional neural networks (CNNs), typically using well-known models like VGG16, Inception, and ResNet18. In several instances, Differential Privacy and secure aggregation were used to solve privacy concerns related to data security. There is still space for improvement in FL's performance in medical image analysis, as the comparison research showed that FL models typically achieved lower accuracy when compared to traditional machine learning models. The study also emphasized issues with model privacy, data heterogeneity, and the actual application of federated systems. Overall, the research highlights the importance of overcoming current restrictions to fully realize FL's potential in this crucial field while offering insightful information about the state of FL in medical image analysis today.

The aim of this paper [23], was to replace a normal convolutional neural network (CNN) with a capsule network to improve the classification accuracy of optical coherence tomography (OCT) pictures. The authors responded to the criticism CNN has received. They reasoned that capsule networks, which are capable of learning positional information from images despite pooling layers' best efforts to the contrary, would be able to work around this restriction. The choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal eye images were among the 83,484 OCT images used by the researchers. They separated the dataset into a training dataset containing 37,205 images and a test dataset containing 1,000 images, with 250 images per category. The authors built and trained a capsule network to create their proposed model. The trained model was then evaluated with the help of the test dataset. The results demonstrated that the proposed method classified OCT images with an impressive 99.6% accuracy. This accuracy was 3 percentage points greater than other methods reported in the literature. In conclusion, the study effectively demonstrated that replacing the CNN with a capsule network improved the accuracy of OCT image classification. Classification results for CNV, DME, drusen, and normal pictures obtained using the suggested method were on par with or better than those obtained using the state-of-the-art methods. This study contributes to the advancement of computer-assisted diagnosis through the use of deep learning techniques and may have major ramifications for improving the precision of illness classification using OCT.

To identify retinal problems through optical coherence tomography (OCT) images, the method we are proposing is intended to improve federated learning. The method focuses primarily on confidence score-based aggregation, which permits a more sophisticated integration of gradient updates from local models into the global model [24]. This method overcomes the limitations of conventional federated learning systems and incorporates novel performance-enhancing features. The use of confidence scores to modulate the gradient updates from local models is a notable feature of our proposed method. The method allocates appropriate weights to the gradient updates during aggregation by measuring the confidence score of each model based on the average SoftMax probabilities. Softmax is an activation function [25], that is frequently used in applications involving multi-class classification for probabilistic outputs. This assures that the updates from more reliable models have a greater impact on the global model. This mechanism for confidence-based scaling contributes to the sophistication of the proposed method. We used label flipping [26], and data poisoning [27], type adversarial attacks to get the accuracy rate. Data poisoning is a technique employed by cyber-criminals to introduce false or skewed data into a training dataset 27]. Data poisoning aims to simulate potential assault scenarios against a federated learning system. This checkpoint enables us to evaluate the safety and stability of our proposed system.

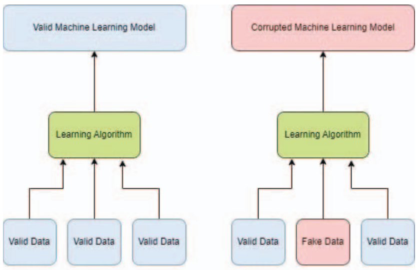


Fig. 1. Data poisoning architecture

The "Retinal OCT Images" dataset available on Kaggle [28] is a collection of Optical Coherence Tomography (OCT) images of the retina, which is the light-sensitive layer at the back of the eye. The dataset contains a total of 84,495 images, each of size 496 x 768 pixels [28].

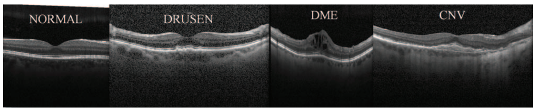
OCT.jpg 

Fig. 2. Retinal OCT dataset samples

Our suggested method addresses several limitations that current federated learning systems are prone to. By incorporating confidence scores and contemplating their impact on the aggregation procedure, the method provides a more refined and precise integration of the contributions of local models. This approach distinguishes it from other methods that rely on straightforward gradient averaging or disregard confidence. Additionally, the authors highlight the prospective advantages of their method for enhancing the precision and performance of the global model.

This paper can be broken down into the following sections: In section 2, we detail the methodologies, experimental design, data sets, and metrics utilized to test the efficacy of the proposed safety measures. Section 3 discusses the analysis and outcomes of the experiments. Finally, section 4 concludes with a review of the study's findings, an analysis of their significance, and the prospect of further study.

A. Problem Statement

This paper focuses on how to mitigate the risk of data contamination in federated learning systems. By inserting erroneous or malevolent data samples into the local datasets of participating devices, data poisoning attacks can jeopardize the integrity and trustworthiness of federated learning's collaborative model [4]. These assaults attempt to influence the model's behavior while it is being trained, which can have catastrophic consequences in mission-critical environments. While there is a growing awareness of data poisoning vulnerabilities in federated learning, the majority of current research focuses on centralized learning settings, leaving a significant void in addressing security issues within the decentralized and privacy-preserving federated learning framework.

In federated learning environments, there is a disparity between what is known and what can be done to prevent data poisoning attacks. While research has examined the susceptibility of federated learning models to data poisoning attacks, robust defenses are still required to ensure the privacy and security of machine learning models. Moreover, the specific requirements of collaborative learning must be met in novel ways that address the unique challenges presented by federated learning's decentralized structure, while maintaining the efficacy of data poisoning attack defense mechanisms [6].

The authors of this article propose a novel technique that employs confidence-based mitigation to enhance the security and privacy of federated learning systems. Our objective is to verify the hypothesis that implementing this mechanism can significantly enhance the security, reliability, and privacy of federated learning [3] by studying the dynamics of malicious attacks, developing a confidence-based mitigation mechanism, conducting extensive experiments, and analyzing the outcomes. By resolving these pressing issues, we can provide dependable collaborative model training using privacy-preserving machine learning technologies and promote their use in privacy-sensitive domains.

B. Research Objectives

The research contributes to the field of federated learning and addresses the issue of data poisoning assaults to improve the robustness and privacy of machine learning models. The primary contributions of this study can be summed up as follows:

- The research investigates the dynamics and vulnerabilities of data poisoning assaults within the context of federated learning [6]. It investigates the attack methodologies, the effects on the training procedure, and the characteristics of tainted data. The research seeks to develop a comprehensive understanding of data poisoning attacks in federated learning environments by gaining an understanding of these aspects.
- The study proposes the development of effective safeguards to mitigate the impact of malevolent local models with high confidence scores. The research seeks to improve the robustness and privacy of the federated learning system by implementing these safeguards [3]. To detect and prevent the inclusion of poisoned data, we will investigate safeguard mechanisms such as anomaly detection, resilient aggregation, and model verification.
- Evaluation of Proposed Safeguards: The research seeks to rigorously evaluate the performance and effectiveness of the proposed safeguards through extensive experimentation and testing [2]. Representative datasets and evaluation metrics commonly used in federated learning research will be employed to assess the efficacy and reliability of the safeguards.

- Contribution to the Literature: This study seeks to address the lack of comprehensive methods to combat data poisoning attacks in the context of federated learning [4] by addressing an existing research gap. This research contributes to the existing literature on adversarial attacks, data poisoning, federated learning, and defense mechanisms by synthesizing and analyzing the existing knowledge.

It is anticipated that these contributions will enhance the understanding of data poisoning attacks in federated learning and pave the way for the development of effective preventive measures, thereby improving the safety, reliability, and privacy of machine learning models in federated learning environments.

II. Proposed Method and Implementation

This section first presents the proposed method in detail, then describes the implementation of our work for its evaluation.

A. Confidence Score Based Aggregation

The proposed method aggregates local models' gradient updates into the global model by scaling them according to their confidence, which enhances the sophistication of the approach.

To measure the model's confidence score C_i, we use the average of the SoftMax probabilities σ for the expected category index i of each input vector z.

$$C_i = \overline{\sigma}(z)_i$$

After calculating the confidence score C_i of each local model and normalizing it, we use the normalized score to scale its respective local model's gradient w_i. Finally, the global model receives the scaled gradient updates from all the local models.

$$w_g = \sum_{i=0}^{N} \left(\frac{C_i}{\sum C_i} * w_i \right)$$

Therefore, the local models with lower confidence score will have less impact on the global model, which leads to a better performing model.

B. Implementation

*1) **Data Organization**:* For the experiment, we used a subset of the "Retinal OCT" dataset that contained 9 local datasets. Each local dataset had 2000 images with balanced categories. We intentionally mislabelled a portion of the images in 3 of the local datasets.

	Correct data	Mislabelled data
Malicious dataset 1	1000	1000
Malicious dataset 2	800	1200
Malicious dataset 3	0	2000

TABLE I
MALICIOUS LOCAL DATASETS

Table 1 shows how much of the local datasets are mislabelled. It can be observed that a substantial proportion of mislabelled data has been used, which provides us with more clear and presentable evaluation results.

Fig. 3. Visualization of Local datasets

Each local dataset was then split into train and validation sets, validation sets containing 10% of the local dataset. Two datasets were created as test sets. One test set containing 400 images is used to measure the confidence score of the local models. The other test set containing 968 images is used to evaluate the global model's performance.

*2) **Data Pre-processing**:* The images are resized into 150x150 pixels and normalized before using them on the model.

*3) **InceptionV3 and VGG19**:* InceptionV3 and VGG19 are used as the base model of the federated learning system. We took a model without the default fully connected layers and added a custom fully connected layer consisting of 4 neurons, one for each category. We applied the SoftMax activation function to the output layer. We only trained the top layer, keeping the rest of the model fixed. We chose Adam as the optimizer for our model.

*4) **Secure Federated Learning System**:* We initialize the global model with the base model and create nine instances of it as local models. Each local model was trained and validated on its own dataset for seven epochs. After that, the confidence score of each model is evaluated. Then, the gradients of the models are aggregated according to their confidence scores before the global model receives them as updates. This concludes one cycle of the Federated Learning System. We repeated this process for 25 cycles.

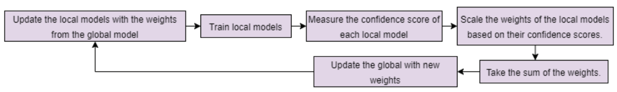


Fig. 4. Secure Federated Learning System

*5) **Insecure Federated Learning System**:* For comparison, we also built another federated learning system that does not use confidence scores. Instead, it uses the traditional FedAvg aggregation method.

*6) **Evaluation**:* After running the FL cycles, we evaluated both global models to demonstrate an in-depth comparison between them. We used several evaluation metrics such as Accuracy, loss, precision, recall, F1 score, and AUC-ROC to highlight the effectiveness of the proposed method.

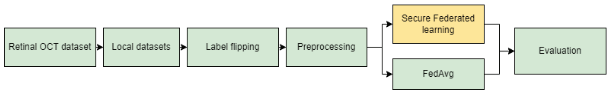


Fig. 5. Workflow

III. Result and Analysis

In this section of the paper, we illustrate various evaluation charts and scores.

A. *Confidence Score*

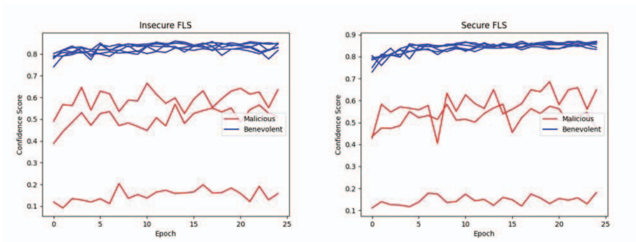


Fig. 6. Confidence score rates for InceptionV3.

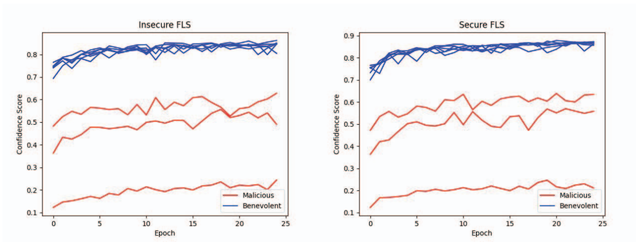


Fig. 7. Confidence score rates for VGG19.

Figure 6 and Figure 7 compare the confidence scores of the local models for secure and insecure federated learning systems (FLS) over 25 cycles. It shows two graphs, one for each type of FLS. The local models are either benevolent or malicious, depending on whether they cooperate or interfere with the global model. The difference in confidence scores between the malicious and benevolent models is quite apparent. The confidence score rates of the benevolent local models are closely clustered together. Moreover, the benevolent local models have more stable confidence scores in the secure FLS than in the insecure FLS. In contrast, the malicious local models have lower and more fluctuating confidence scores in the secure FLS than in the insecure FLS. This is because the malicious local models try to fit the global model with different types of datasets, which are more effective in the insecure FLS and less effective in the secure FLS.

B. *Accuracy and Loss*

Figure 8 and Figure 9 show a line chart comparing the accuracy and loss of the global model for secure and insecure federated learning systems (FLS) over 25 cycles. In the case of InceptionV3, The secure FLS starts with an accuracy of

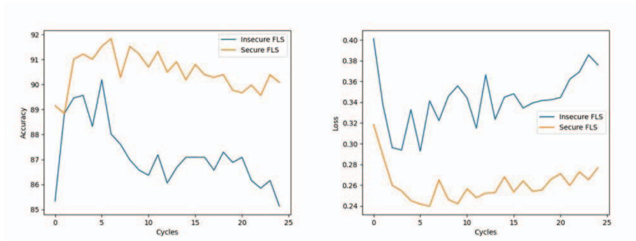


Fig. 8. Accuracy and Loss rates for InceptionV3.

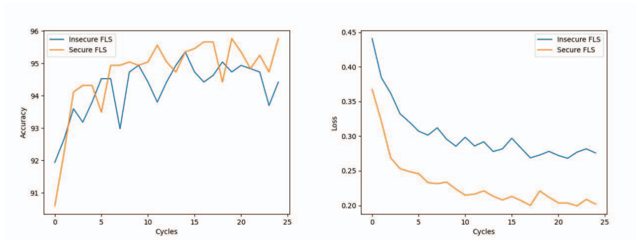


Fig. 9. Accuracy and Loss rates for VGG19.

about 89%, while the insecure FLS starts with about 85%. The insecure FLS reaches a peak accuracy of slightly above 90% before declining, while the secure FLS maintains an accuracy of around 92%. On the other hand, for VGG19, the secured FLS achieves a peak accuracy of 95.8%, in contrast to the insecure FLS, which reaches a lower peak and then declines. The charts clearly demonstrate that the secure FLS has a higher and more stable accuracy than the insecure FLS. Similarly, the charts illustrate that the secure FLS has a lower and more stable loss rate than the insecure FLS for both Inception and VGG19.

C. *Confusion Matrix*

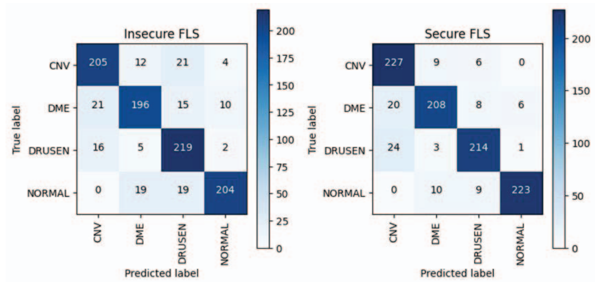


Fig. 10. Confusion matrix for InceptionV3.

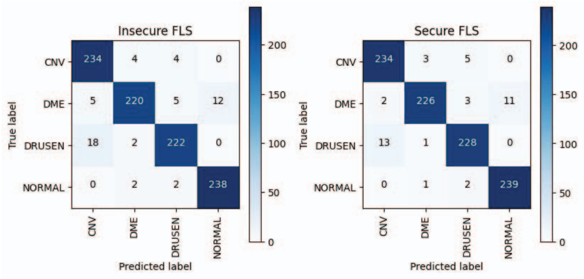


Fig. 11. Confusion matrix for VGG19.

The confusion matrix of the two systems in Figure 10 and Figure 11 provides less clarity to the comparison. However, the secure FLS seems to have better overall results than the insecure FLS. Only the DRUSEN category in the case of InceptionV3 model, has slightly more true positives in the insecure FLS. To evaluate the confusion matrices better, we can use the following metrics.

Table II and Table III illustrate the precision, recall, F1 score, and AUC-ROC of insecure FLS and secure FLS. Secure FLS achieves noticeably better scores in every metric for both InceptionV3 and VGG19.

	InceptionV3	
	Insecure FLS	Secure FLS
Precision	0.8690	0.9063
Recall	0.8357	0.8894
F1	0.8520	0.8978

TABLE II
COMPARISON OF PERFORMANCE METRICS BETWEEN INSECURE FLS AND SECURE FLS FOR INCEPTIONV3

	VGG19	
	Insecure FLS	Secure FLS
Precision	0.9579	0.9630
Recall	0.9174	0.9411
F1	0.9372	0.9519

TABLE III
COMPARISON OF PERFORMANCE METRICS BETWEEN INSECURE FLS AND SECURE FLS FOR VGG19

IV. CONCLUSION AND FUTURE WORK

This research paper addresses the urgent need for a reliable federated learning system in diagnosing retinal disorders. By leveraging confidence scores, we improve the speed and accuracy of diagnosis while ensuring data privacy and security. Our study focuses on the Retinal optical coherence tomography (OCT) dataset, with a particular emphasis on combating data poisoning attacks. We propose a protected system that utilizes SoftMax values to assess the reliability of local updates. Evaluation of our federated learning network with nine local models demonstrates significant improvements in accuracy, recall, F1 score, and AUC-ROC. Notably, precision increased from 0.869 to an excellent 0.906, and the F1 score rose from 0.852 to 0.898 for InceptionV3.

Looking ahead, we plan to expand our model's evaluation with real-world datasets to investigate the effects of detrimental data in multi-client scenarios. Simulated attacks on realistic targets, along with the incorporation of explainable AI techniques, will further enhance transparency and trust in federated learning, fostering collaboration among healthcare professionals and improving patient care outcomes.

REFERENCES

[1] Huang D, Swanson EA, Lin CP, et al. Optical coherence tomography. Science. 1991;254(5035):1178-1181. doi:10.1126/science.1957169
[2] Sadda SR, Staurenghi G, Chakravarthy U, et al. Optical coherence tomography endpoints in neovascular age related macular degeneration clinical trials: a systematic review and consensus recommendations. Eye (Lond). 2015;29(4):441-453. doi:10.1038/eye.2015.34
[3] American Academy of Ophthalmology Retina/Vitreous Panel. Preferred Practice Pattern® Guidelines. Diabetic Retinopathy. San Francisco, CA: American Academy of Ophthalmology; 2019. Available at: https://www.aao.org/preferred-practice-pattern/ diabetic-retinopathy-ppp-2019. Accessed June 13, 2023.
[4] Shields CL, Materin MA, Mehta S, Shields JA. Optical coherence tomography of retinal tumors. JAMA Ophthalmol. 2015;133(5):584-590. doi:10.1001/jamaophthalmol.2015.0103
[5] Shahlaee A, Samara WA, Hsu J, et al. In vivo assessment of macular vascular density in healthy human eyes using optical coherence tomography angiography. Am J Ophthalmol. 2016;165:39-46. doi:10.1016/j.ajo.2016.02.037
[6] Sotoudeh-Paima, S., Jodeiri, A., Hajizadeh, F., Soltanian-Zadeh, H. (2022). Multi-scale convolutional neural network for automated AMD classification using retinal OCT images. Computers in biology and medicine, 144, 105368.
[7] 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.
[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.
[9] Pekala, M., Joshi, N., Liu, T. A., Bressler, N. M., DeBuc, D. C., Burlina, P. (2019). Deep learning based retinal OCT segmentation. Computers in biology and medicine, 114, 103445.
[10] Jing Tian, Boglarka Varga, Erika Tatrai, Palya Fanni, Gabor Mark Somfai, William E Smiddy, and Delia Cabrera DeBuc. Performance evaluation of automated segmentation software on optical coherence tomography volume data. Journal of biophotonics, 9(5):478–489, 2016.
[11] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 1175–1183. IEEE, 2017.
[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.
[13] Arcadu, F., Benmansour, F., Maunz, A., Michon, J., Haskova, Z., McClintock, D., ... Prunotto, M. (2019). Deep learning predicts OCT measures of diabetic macular thickening from color fundus photographs. Investigative ophthalmology visual science, 60(4), 852-857.
[14] RIDE: A Study of Ranibizumab Injection in Subjects With Clinically Significant Macular Edema (ME) With Center Involvement Secondary to Diabetes Mellitus. Registered on ClinicalTrials.gov as NCT00473382. Available at: https:// clinicaltrials.gov/ct2/show/NCT00473382. Accessed May 2, 2017.
[15] RISE: A Study of Ranibizumab Injection in Subjects With Clinically Significant Macular Edema (ME) With Center Involvement Secondary to Diabetes Mellitus. Registered on ClinicalTrials.gov as NCT00473330. Available at: https:// clinicaltrials.gov/ct2/show/NCT00473330. Accessed May 2, 2017.
[16] Qummar, S., Khan, F. G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z. U., ... Jadoon, W. (2019). A deep learning ensemble approach for diabetic retinopathy detection. Ieee Access, 7, 150530-150539.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.
[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2818–2826.

[19] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1251–1258.
[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4700–4708.
[21] Y. Yang, T. Li, W. Li, H. Wu, W. Fan, and W. Zhang, “Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2017, pp. 533–540.
[22] Yoo, T.K., Choi, J.Y. Kim, H.K. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. Med Biol Eng Comput 59, 401–415 (2021). https://doi.org/10.1007/s11517-021-02321-1
[23] Tsuji, T., Hirose, Y., Fujimori, K. et al. Classification of optical coherence tomography images using a capsule network. BMC Ophthalmol 20, 114 (2020). https://doi.org/10.1186/s12886-020-01382-4
[24] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[25] Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.
[26] Kurakin, A., Goodfellow, I., Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[27] Gu, T., Dolan-Gavitt, B., Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.
[28] “Retinal OCT Images (Optical Coherence Tomography).” Retinal OCT Images (Optical Coherence Tomography) — Kaggle, /datasets/paultimothymooney/ker- many2018. Accessed 13 Jan. 2023.
[29] M. F. Sohan and A. Basalamah, ”A Systematic Review on Federated Learning in Medical Image Analysis,” in IEEE Access, vol. 11, pp. 28628-28644, 2023, doi: 10.1109/ACCESS.2023.3260027.