Resilient Federated Vision Transformers for Alzheimer's Disease Prediction with Brain Imaging Data Jittering

Mahyar Mohammadi*, Mohammad Hossein Badiei*, Mohammad Mashreghi*, Keyvan Kazemi, and Hamed Kebriaei (*Senior Member, IEEE*)

***Abstract*— Neural networks in Alzheimer's MRI prediction are vulnerable to subtle adversarial perturbations and limited data diversity. In this paper, we present a robustness strategy for federated learning (FL) with vision transformers (ViTs), termed FL-ViTs, aimed at improving classification performance under domain uncertainties. We employ worst-case robust optimization techniques in FL-ViT models to effectively handle uncertainty regions and address data heterogeneity in FL. Specifically, we implement a min-max optimization framework, where the inner maximization step simulates worst-case adversarial perturbations for each local model, and the outer minimization step adjusts the model parameters to minimize loss under these conditions. Our experiments use benchmark models (e.g., ResNet-18, DenseNet121) on standard brain MRI data, applying recent robust optimization methods like jittering, TIFGSM, and DIFGSM to boost resilience and predictive accuracy. Experimentally validated, our approach achieved a 38% increase in robustness, with accuracy improving from 25.72% in the baseline scenario to 63.29% under adversarial attacks. In addition, our approach shows an efficient trade-off between robustness and accuracy, with a 0.28% improvement in accuracy, from 99.63% in the baseline to 99.90% in robust, attack-free conditions. Overall, the Jitter-robust method outperforms TIFGSM-robust and DIFGSM-robust in both attack-free and attacked scenarios, achieving the best balance between accuracy and adversarial resilience with an average accuracy of 81.34%.**

***Index Terms*— Alzheimer's disease prediction, Federated learning, Vision transformers, Adversarial machine learning, Robust optimization.**

I. Introduction

In the realm of distributed machine learning, federated learning (FL) enables collaborative model training across decentralized datasets without sharing raw data, offering inherent scalability, computational efficiency, and privacy preservation, particularly in healthcare applications [1], [2]. Although privacy-preserving techniques like homomorphic encryption (HE) and secure multi-party computation (SMPC) ensure data confidentiality, FL offers distinct advantages, especially in Alzheimer's research, where multi-institutional collaboration is essential to overcome data scarcity while maintaining strict privacy compliance (e.g., HIPAA, GDPR) [3], [4]. Moreover, FL aligns naturally with the decentralized structure of medical data repositories, enabling institutions to retain full control over their sensitive imaging datasets. By integrating FL with Vision Transformers (ViTs), such approaches not only preserve privacy but also harness ViTs' global modeling capabilities to capture cross-site patterns in Alzheimer's disease (AD) progression—something that centralized or cryptographically intensive methods often struggle to achieve [5].

While ViT structures within FL have proven effective in analyzing imaging data, recent advancements in deep learning techniques have further enhanced the ability to identify and categorize complex patterns. However, the effectiveness of these models relies on access to diverse patient data in a domain characterized by sparse data, significant heterogeneity, and potential adversarial threats. Despite some progress in collecting brain imaging data, its availability remains limited, highlighting the need for generalization and robustness to overcome these challenges, even within FL-ViTs [6], [7].

To address these challenges, we propose Jitter-robust FL-ViT, TIFGSM-robust FL-ViT, and DIFGSM-robust FL-ViT algorithms that utilize adversarial robust optimization with gradient-based optimization within the domain-specific data synthesis. Fig. 1 shows our FL-ViT enhanced with data jitter for robustness. Our proposed framework enhances the robustness and generalizability of FL-ViT models for AD prediction by leveraging the network layer characteristics of Vision Transformers with worst-case adversarial robust optimization techniques, ensuring resilience against adversarial attacks while improving the model's ability to generalize across decentralized medical datasets. This offers a promising solution to bolster AI models in FL for Alzheimer's diagnosis.

To further strengthen our approach, we extend it to work with FedBN, a FL framework designed to handle non-independent and identically distributed (non-iid) data. By utilizing FedBN's adaptive normalization with ViT's attention mechanisms, our method addresses the dual challenges of heterogeneity and adversarial robustness. This advancement goes beyond existing FL-ViT frameworks, such as FedMHA [8] and FedViT [9], offering a scalable and privacy-preserving solution for AD prediction. Finally, the key contributions are:

- **Development of Resilient FL-ViT Algorithms:** We introduce three novel algorithms, Jitter-robust, TIFGSM-robust, and DIFGSM-robust FL-ViTs, designed to enhance the robustness and accuracy of ViTs within FL. These algorithms effectively address the challenges of data diversity and vulnerability to adversarial attacks.

M. Mohammady, MH. Badiei, M. Mashreghi, K. Kazemi and H. Kebriaei are with the Department of Electrical and Computer Engineering, University of Tehran, Tehran, 1417614411, Iran, (email: mahyar.mohammady@ut.ac.ir; mh.badiei@ut.ac.ir; m.mashreghi@ut.ac.ir; kebriaei@ut.ac.ir).

*** These authors contributed equally to this work.**

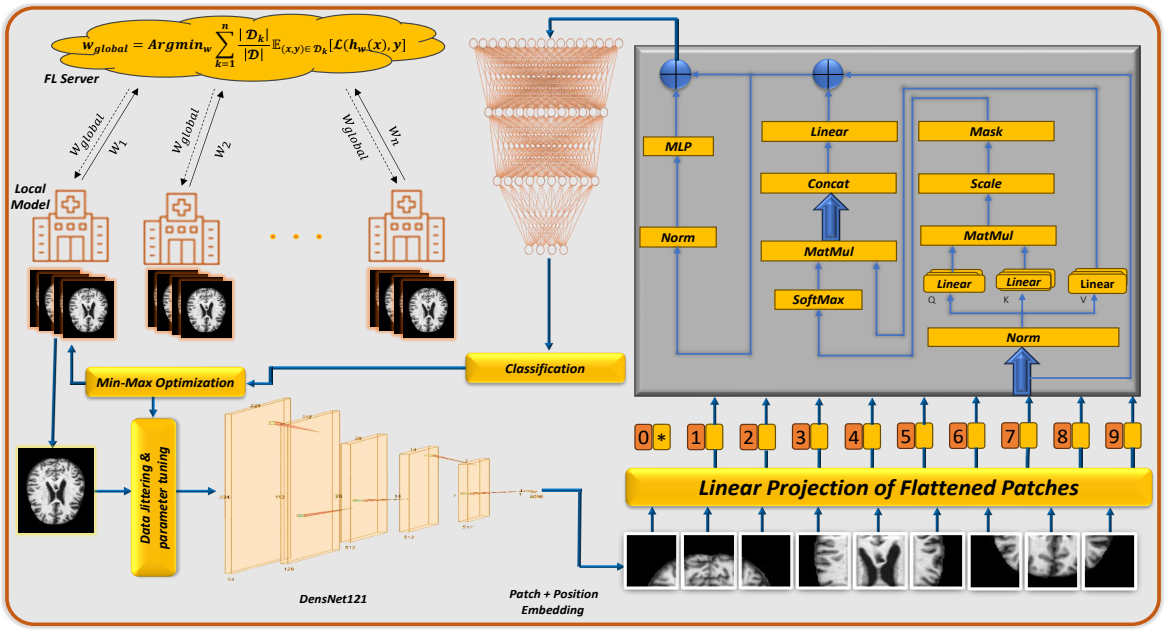


Fig. 1: FL with robust optimization for Deep Networks (e.g., DenseNet121, as shown in the figure) Based on Vision Transformer.

- **Enhanced Generalization:** Our approach improves generalization by exploring uncertainty regions and strengthening data fidelity. By examining and incorporating uncertain scenarios, we enrich the dataset with synthetic examples that reflect potential variations, allowing the model to learn from a broader range of conditions. Additionally, we focus on enhancing data fidelity to ensure high-quality, representative training data. This combination enables our algorithms to perform reliably in complex environments, such as AD prediction, by maintaining robustness against adversarial perturbations and generalizing effectively to unseen data.
- **Resilience Against Adversarial Attacks:** Resilient FL-ViT algorithms demonstrate a notable 38% improvement in accuracy when combating adversarial attack scenarios. Our method increases average accuracy from approximately 25.72% to about 63.29% under attack conditions.
- **Balancing Accuracy and Robustness:** The approach resulted in a 0.28% increase in average accuracy (from 99.63% to 99.90%) in non-attack scenarios, resulting in a well-balanced trade-off between robustness and accuracy, essential for reliable predictions in clinical applications.
- **Effectiveness of the Jitter-Based Robustness Approach:** Our findings demonstrate that the Jitter-robust approach strikes the optimal balance between accuracy and robustness, achieving an average accuracy of 81.34% across both attack-free and attacked scenarios combined. It outperforms DIFGSM-robust (77.86%), TIFGSM-robust (39.81%), and the baseline (31.88%).

Our experiment code is publicly available at GitHub.

The rest of the article is structured as follows. Section II reviews related work. Section III outlines the methodology. Section IV presents simulation results and discussions. Section V illustrates the main challenges and possible future works. Finally, Section VI concludes with key takeaways.

II. Related Work

AD is a growing health concern, especially among older individuals, leading to cognitive decline and memory impairment. It is a leading cause of dementia, affecting a significant portion of the population over 60, with prevalence expected to rise in the coming years [10]. Diagnosing dementia typically involves neuroimaging techniques such as MRI scans to monitor disease progression. However, challenges such as data bias, computational efficiency, and privacy preservation remain in applying FL for reliable and efficient prediction [11], [12].

In the forefront, ViTs have shown great promise in computer vision, with recent efforts to adapt them to medical imaging. One key difference in the ViT architecture, compared to CNNs, is their self-attention mechanism which can capture long-range spatial dependencies, providing a more global perspective [13]. This property is intuitive since anatomical context and spatial patterns are often crucial in analyzing medical images. ViTs typically require large datasets for training, but effective pretraining/fine-tuning techniques have overcome this issue in other contexts. Although ViTs have shown initial success for various neuroimaging tasks, they typically train the models on thousands of MRI scans [14]–[17].

Despite FL's widespread use in computationally and security efficient image analysis, particularly in healthcare systems, it remains vulnerable to source inference attacks (SIAs) and adversarial attacks. To address these challenges, two comprehensive frameworks are proposed: FRESH, which utilizes a ring signature defense to protect physiological data during joint training of FL models [18], and a blockchain-based FL framework with SMPC model verification to securely aggregate local models while detecting and defending against malicious updates [19]. Additionally, to protect data privacy in deep learning-based image synthesis for healthcare research, a Federated Differentially Private Generative Adversarial Net-

work (FedDPGAN) is introduced, combining differential privacy and FL to privately generate diverse patient data without sharing the client's empirical data [20]. Furthermore, a defense strategy called FedDetect is proposed to mitigate backdoor attacks in federated generative adversarial networks (FedGANs) used for medical image synthesis by identifying and blocking malicious clients based on their loss patterns [21].

In addition, the FedBrain framework tackles challenges in brain disease diagnosis by utilizing FL to handle the curse of dimensionality and data distribution differences across sites. It incorporates data augmentation, domain alignment methods, and a personalized predictor based on a mixture of experts, leading to enhanced performance and decreased communication burden compared to other FL-based approaches [22].

Although ViTs are beneficial in this area, FL, such as Federated Domain Adaptation via Transformer (FedDAvT) and Federated Deep Convolutional Neural Network Alzheimer Detection Schemes (FDCNN-AS), has emerged as a pivotal technique for AD classification while addressing data privacy [23], [24]. Recent studies highlight the transformative potential of FL-ViTs in addressing domain-specific challenges. For instance, Rei et al. [25] demonstrated that encrypting ViT embeddings (e.g., patch and positional encodings) preserves model accuracy while preventing inversion attacks like APRIL, a critical consideration for medical data privacy. Similarly, Queyrut et al. [26] proposed shielding early ViT layers via Trusted Execution Environments (TEEs) to mitigate gradient-based evasion attacks, achieving 98.8% robust accuracy on CIFAR-10. However, FL-ViTs still face challenges in domains characterized by data scarcity, such as AD prediction. This limitation stems from insufficient training data, which can be mitigated only when models are designed to operate on distributions beyond the empirical dataset, thereby enhancing their generalization capabilities. Another work by Darzi et al. [8] addresses this by aligning multi-head attention mechanisms across clients to harmonize heterogeneous medical data, improving fairness for underrepresented datasets.

Another FL paradigm, which has addressed data variability in multi-centric studies, utilizes a hierarchical Bayesian latent variable model to estimate client-specific parameters from a global distribution, effectively accounting for data bias and variability. Experimental results on multi-modal medical imaging data and clinical scores from distributed clinical datasets of AD patients demonstrate the method's robustness in handling both iid and non-iid data distributions, providing accurate data reconstruction and outperforming autoencoding models [27].

To address challenges in traditional centralized FL architectures, a decentralized and privacy-preserving training scheme called Robust and Privacy-Preserving Decentralized Deep Federated Learning (RPDFL) has been introduced. RPDFL utilizes a ring FL structure and a Ring-Allreduce-based data sharing scheme, resulting in superior model accuracy and convergence compared to standard FL methods. These advancements make RPDFL well-suited for digital healthcare applications where privacy and robustness are crucial [28].

However, the integration of FL with advanced deep learning models, such as the ViT, has shown promising results in various applications. ViT, a powerful model in computer vision, outperforms traditional models, mainly based on convolutional neural networks by capturing token relationships and generating encoded features. Ongoing research explores the potential of ViT in multi-task learning and distributed learning methods, including medical imaging tasks [29]–[33].

In recent advancements of AD classification, FL-Vit has emerged as a pivotal technique. Traditional domain adaptation methods or CNN-based FL compromise data privacy and enhanced performance [24]. FL-ViT circumvents this by enabling model training across multiple sites without exchanging sensitive data via advanced deep neural networks. Notably, the Federated Domain Adaptation via Transformer framework leverages a Transformer network to mitigate data heterogeneity and maintain data privacy. This approach aligns self-attention maps in source and target domains using mean squared error for subdomain adaptation, leading to impressive accuracy rates in various classification tasks. These frameworks represent cutting-edge methods in FL for AD classification, showcasing significant potential in achieving high accuracy while safeguarding data privacy [23], [34].

III. Methodology

A. Background

The proposed framework integrates a ViT architecture with complementary convolutional neural networks, specifically ResNet18 and DenseNet121, to leverage their synergistic strengths in hierarchical feature extraction. To address data decentralization challenges, a FL paradigm is employed, enabling collaborative model training across distributed nodes while preserving data privacy. Additionally, adversarial training mechanisms are incorporated to enhance robustness against perturbations and potential adversarial attacks. Each component of this hybrid structure—vision integration, distributed learning protocols, and resilience optimization strategies—is systematically designed to address specific technical and operational constraints, as detailed in the subsequent sections.

Vision Transformers. The ViT architecture forms the backbone of our proposed model. ViTs process input images by dividing them into fixed-size patches, which are then linearly embedded into a high-dimensional space. The input sequence to the transformer encoder consists of these patch embeddings, along with a special classification token (T_{cls}) and position embeddings (P_{pos}). The initial input sequence (X_0) is formulated as:

$$X_0 = [T_{cls}; P_{img}] + P_{pos}, \tag{1}$$

where P_{img} represents the embedded image patches, and (P_{pos}) encodes the positional information of each patch.

The transformer encoder comprises multiple identical layers, each consisting of a multi-head self-attention (MSA) mechanism and a multi-layer perceptron (MLP). Layer normalization ($\mathcal{L}_{norm}$) is applied before each operation, and residual connections are used to facilitate gradient flow and stabilize training. The transformations within each encoder layer are described as follows:

$$H_l^* = MSA(\mathcal{L}_{norm}(H_{l-1})) + H_{l-1}, \tag{2}$$

$$H_l = MLP(\mathcal{L}_{norm}(H_l^*)) + H_l^*, \tag{3}$$

here, H_l represents the hidden states of block l, and H_l^* is the intermediate output after the self-attention mechanism. The final output of the classification token (y_{cls}) is obtained by applying layer normalization to the hidden state of the classification token after the last encoder block:

$$y_{cls} = \mathcal{L}_{norm}(H_l^0). \tag{4}$$

In the context of brain imaging, the patch-wise embedding layer is tailored to capture the unique characteristics of neuroimaging data. This adaptation, combined with the global modeling capabilities of ViTs, enables the model to effectively analyze complex spatial relationships in brain scans, making it particularly suitable for AD diagnosis.

Adversarial robustness. Adversarial training of FL-ViTs employs robust optimization principles against evasion attacks. For a ViT model, the adversarial training objective is formulated as Eq.(5):

$$\min_{\Theta} \max_{\delta \in \Delta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\Theta(x_i + \delta), y_i), \tag{5}$$

where $\Delta = \{\delta \in \mathbb{R}^d : ||\delta||_p \leq \epsilon\}$ represents permitted perturbations and the client's local model is represented as f_Θ with parameters Θ and dataset $D = \{(x_i, y_i)\}_{i=1}^N$. ϵ represents the maximum allowable p-norm of the perturbation, setting an upper bound on the magnitude of input variations to control model robustness against adversarial influences. In the basic FedAvg approach, each client performs a local update as follows:

$$\Theta_k^{t+1} = \arg\min_{\Theta}(\frac{1}{N_k} \sum_{i \in D_k} \mathcal{L}(f_\Theta(x_i + \delta^*), y_i)), \tag{6}$$

where δ^* represents an optimal perturbation added to input x_i in the dataset D_k, where the perturbation is specifically chosen to adapt the model training for improved generalization or robustness. FedProx extends this by adding a proximal term to prevent excessive local model drift as Eq.(7) which the term μ is a regularization parameter used to control the degree of model update at each client:

$$\Theta_k^{t+1} = \arg\min_{\Theta}(\tfrac{1}{N_k} \textstyle\sum_{i \in D_k} \mathcal{L}(f_\Theta(x_i + \delta^*), y_i) + \tfrac{\mu}{2}||\Theta - \Theta^t||^2). \tag{7}$$

The FedBN approach maintains client-specific batch normalization statistics with local updates given by:

$$(\Theta_k^{t+1}, \gamma_k^{t+1}, \beta_k^{t+1}) = \arg\min_{\Theta, \gamma_k^t, \beta_k^t} \tfrac{1}{N_k} \textstyle\sum_{i \in D_k} \mathcal{L}(f_{\Theta, \gamma_k^t, \beta_k^t}(x_i + \delta^*), y_i), \tag{8}$$

where γ and β are learnable parameters in the batch normalization layers. Specifically, γ controls the scaling of the normalized data to adjust the feature's standard deviation, while β shifts the normalized data to alter the feature's mean. These parameters help the model adapt to optimal data distribution, enhancing stability and improving convergence during training. In FedBN, batch normalization statistics remain client-specific during aggregation, enabling better handling of data heterogeneity across clients.

B. Resilient FL-ViTs in Healthcare

The critical challenges of adversarial threats, limited data diversity, and poor generalization to unseen data in healthcare AI necessitate robust methodologies. To mitigate these issues, we propose three enhanced FL-ViT approaches: FL-ViT Jitter-robust, FL-ViT TIFGSM-robust, and FL-ViT DIFGSM-robust. These methods enhance adversarial resilience by leveraging robust training frameworks (TIFGSM/DIFGSM) to counter evasion attacks, improve generalization through jitter-based augmentation that simulates domain shifts, and stabilize optimization in federated settings with non-IID data. By aligning with the unique challenges of decentralized medical imaging, our approaches ensure both security and adaptability, reinforcing model reliability in real-world clinical applications.

FL-ViT Jitter-robust. Jitter-robust improves the robustness of untargeted adversarial attacks by fostering diversity in misclassification targets and minimizing perturbations. Traditional untargeted attacks tend to yield limited categories for misclassification, which constrains robustness. To address this limitation, a Euclidean distance-based loss is employed in place of the conventional cross-entropy (CE) loss, maximizing output logits for all classes except the correct one and thereby enhancing the variety of misclassification outcomes. Gaussian noise is introduced into the logits to encourage varied gradient directions, further strengthening robustness. The method updates adversarial examples only when the resulting perturbation is both effective and smaller than previous attempts, thus maintaining minimal perturbations.

The adversarial example $x_{\text{adv}} = x + \delta$ is generated by adding a perturbation δ to the original input x. This adversarial input is then passed through the model f_Θ, producing the logits vector $z = f_\Theta(x_{\text{adv}})$, where z represents the unnormalized class scores. These logits are subsequently rescaled to produce the final class probabilities:

$$\hat{z} = \text{softmax}\left(\alpha \cdot \frac{z}{\|z\|_\infty}\right), \tag{9}$$

where α is a scaling factor that adjusts the range of the logits. The rescaled output $\hat{z}$ provides the class probabilities after the softmax transformation, which is influenced by the adversarial example x_{adv} and the perturbation δ. Thus, the Jitter Loss for client k is defined as follows:

$$\mathcal{L}_{\text{Jitter}}^{(k)} = \begin{cases} \frac{\|\hat{z} - y + N(0,\sigma)\|_2}{\|\delta\|_p} & \text{if } x_{\text{adv}} \text{ is misclassified,} \\ \|\hat{z} - y + N(0,\sigma)\|_2 & \text{if } x_{\text{adv}} \text{ isn't misclassified,} \end{cases} \tag{10}$$

where $N(0,\sigma)$ is Gaussian noise with standard deviation σ, $||\delta||_p$ denotes the norm of the perturbation. This loss function balances the goal of effective misclassification with minimal perturbation norms, thereby enhancing robustness through diversity in misclassifications and reduced perturbation magnitudes.

In fact, Jitter robustness enhances ViT models by introducing controlled positional perturbations during training, improving resilience to spatial variations through random jittering of input patches. This method is particularly valuable in FL-ViTs, where data distributions and quality differ across

clients, as Jitter synthesis applies random spatial offsets while preserving patch order. By fostering invariance to small positional shifts and addressing statistical heterogeneity among federated clients, this approach significantly strengthens model adaptability and robustness to input variations, maintaining the hierarchical structure of visual information [35].

FL-ViT TIFGSM-robust. TIFGSM enhances adversarial attacks by incorporating translation invariance into the gradient-based perturbation generation process [36]. This approach begins with defining a set of T translated versions of the original input x, denoted as $\{x^{(j)}\}_{j=1}^{|T|}$. These translations help capture the model's response to variations in the input, which is crucial for developing more robust adversarial examples.

For each translated image $x^{(j)}$, we compute the gradient of the loss function $\mathcal{L}(f_\Theta(x^{(j)}), y)$ with respect to the input. This gradient, represented as $\nabla_{x^{(j)}}\mathcal{L}(f_\Theta(x^{(j)}), y)$, indicates how changes in the input affect the model's prediction. To obtain a more stable perturbation direction, we average these gradients over all T translations, resulting in:

$$g = \frac{1}{|T|}\sum_{j=1}^{|T|}\nabla_{x^{(j)}}\mathcal{L}(f_\Theta(x^{(j)}), y). \tag{11}$$

Next, the perturbation is generated by applying the sign function to the averaged gradient, scaled by a factor ζ. This yields the adversarial example as:

$$x_{\text{adv}} = x + \zeta \cdot \text{sign}(g). \tag{12}$$

Finally, instead of explicitly defining the adversarial example x_{adv}, we can express the overall loss formulation associated with TIFGSM as:

$$\mathcal{L}_{\text{TIFGSM}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\mathcal{L}\left(f_\Theta\left(x + \zeta \cdot \text{sign}(g)\right), y\right)\right]. \tag{13}$$

FL-ViT DIFGSM-robust. DIFGSM enhances adversarial training by generating diverse adversarial examples through image transformations $T(\cdot)$. By applying these transformations with a specified probability p at each iteration, DIFGSM mitigates overfitting and improves model robustness.

Key transformations include random resizing and random padding, which introduce diversity in input data. The probability p controls the trade-off between success rates on white-box and black-box models: when $p = 0$, DIFGSM reduces to FGSM, while $p = 1$ utilizes only transformed inputs, enhancing black-box success rates but potentially decreasing white-box effectiveness.

The updating rule for DIFGSM is given by:

$$x_{\text{adv}}^{t+1} = \text{Clip}_x^{\epsilon}\left(x_{\text{adv}}^{t} + \epsilon \cdot \text{sign}\left(\nabla_x\mathcal{L}(f_\Theta(T(x_{\text{adv}}^{t}; p)), y)\right)\right). \tag{14}$$

The stochastic transformation function is defined as:

$$T(x_{\text{adv}}^{t}; p) = \begin{cases} T(x_{\text{adv}}^{t}) & \text{with probability } p \\ x_{\text{adv}}^{t} & \text{with probability } 1-p. \end{cases} \tag{15}$$

After T iterations, the final adversarial example is defined as $x' = x^{(T)}$ [37]. Ultimately, the objective of DIFGSM for each client is to minimize the expected adversarial loss, which can be expressed as:

$$\mathcal{L}_{\text{DIFGSM}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\mathcal{L}(f_\Theta(x'), y)\right]. \tag{16}$$

Finally, for all robust methods, global aggregation in the round $t+1$ can be written as follows:

$$\Theta^{t+1} = \frac{1}{N}\sum_{k=1}^{K}(\arg\min_{\Theta}\sum_{i\in D_k}\mathcal{L}_{robust}^{(k)}(x_i, y_i; \Theta)). \tag{17}$$

The FedProx global aggregation can be formulated as:

$$\Theta^{t+1} = \frac{1}{N}\sum_{k=1}^{K}(\arg\min_{\Theta}\sum_{i\in D_k}\mathcal{L}_{\text{robust}}^{(k)}(x_i, y_i; \Theta) + \frac{\mu}{2}\|\Theta - \Theta^t\|^2). \tag{18}$$

For FedBN, the global aggregation can be expressed as:

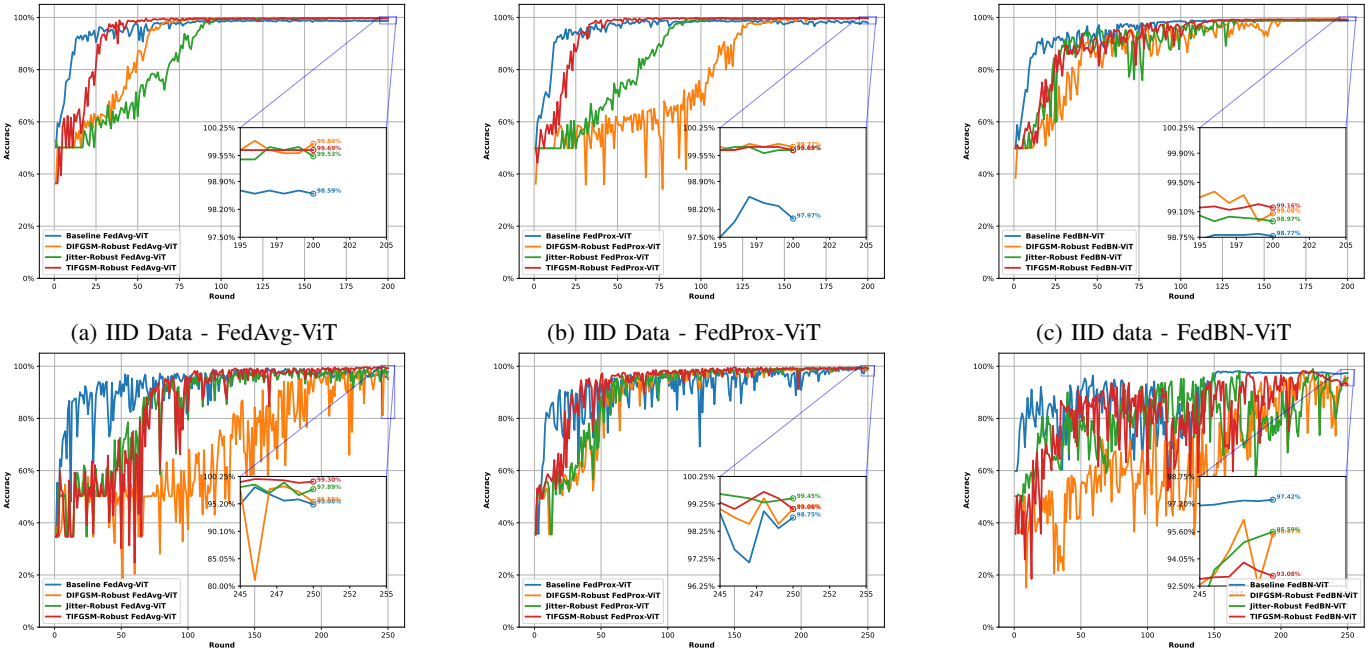
$$\Theta^{t+1} = \frac{1}{N}\sum_{k=1}^{K}(\arg\min_{\Theta}\sum_{i\in D_k}\mathcal{L}_{robust}^{(k)}(x_i, y_i; \gamma^{(k)}, \beta^{(k)}, \Theta)), \tag{19}$$

where $\gamma^{(k)}$ and $\beta^{(k)}$ are the optimal parameters for the batch normalization layers in the k-th client.

IV. Simulation Results and Discussion

We assess the proposed methods using MRI data from the ADNI database. The dataset comprises 6,400 MRI images, divided into four groups: Mild Cognitive Impairment (MCI), Moderate-Stage AD, Very Mild Cognitive Impairment (VMCI), and Control subjects, with a balanced distribution across categories [38]. We utilized ResNet18, which employs residual connections to address the vanishing gradient problem, making it relatively lightweight and fast, and DenseNet121, which utilizes dense connections that improve feature reuse and generally yield higher accuracy, though at the cost of increased computational complexity. In the hybrid ResNet18-ViT model, feature maps are converted into patch embeddings and reshaped to match the input shape required for the Transformer. The positional embeddings and class token are then added to the sequence, allowing the Transformer encoder stack to process the sequence and learn relationships across patches for enhanced classification accuracy. In addition, to better reflect real-world connectivity challenges in distributed learning systems, we employ partial client selection, where 20% of clients are randomly chosen in each round to participate in global model training.

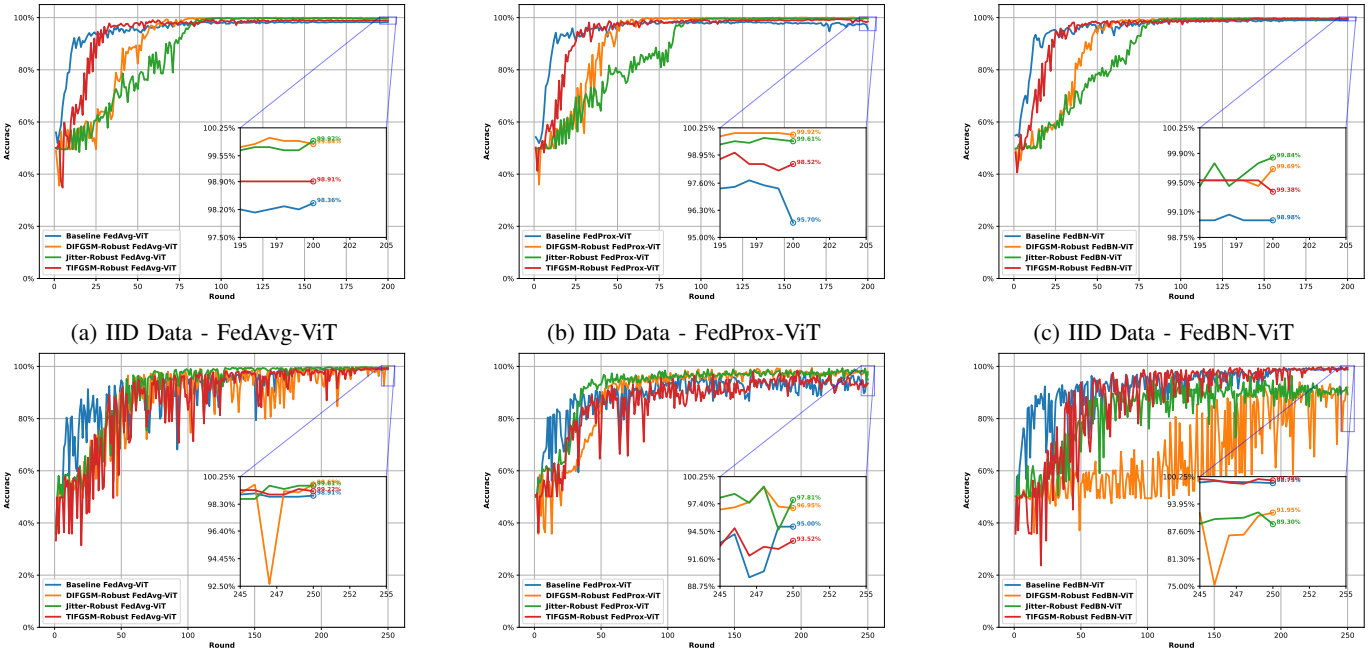
Our selection of adversarial attacks—FGSM, PGD, CW, TIFGSM, etc—was driven by their relevance to practical medical imaging applications, particularly in FL for healthcare disease diagnosis. FGSM and PGD represent gradient-based attacks that simulate fast and iterative adversarial perturbations, reflecting potential security threats in AI-driven clinical systems [39]. For example DeepFool, known for its optimization-based approach, targets healthcare predictive model vulnerabilities with high attack success rates, making it crucial for assessing worst-case robustness [40]. TIFGSM, designed for transferable attacks, mirrors cross-institutional domain shifts, which are common in decentralized MRI data. Additionally, we include natural perturbations like Gaussian noise to model sensor noise and real-world acquisition



(a) IID Data - FedAvg-ViT (b) IID Data - FedProx-ViT (c) IID data - FedBN-ViT

(d) non-IID Data - FedAvg-ViT (e) non-IID Data - FedProx-ViT (f) non-IID Data - FedBN-ViT

Fig. 2: Performance comparison of ResNet18-ViT models using FedAvg, FedProx, and FedBN, showing the accuracy of IID data (a, b, c) and non-IID data (d, e, f) for Jitter-robust, TIFGSM-robust, and DIFGSM-robust across diverse FL-ViT scenarios.



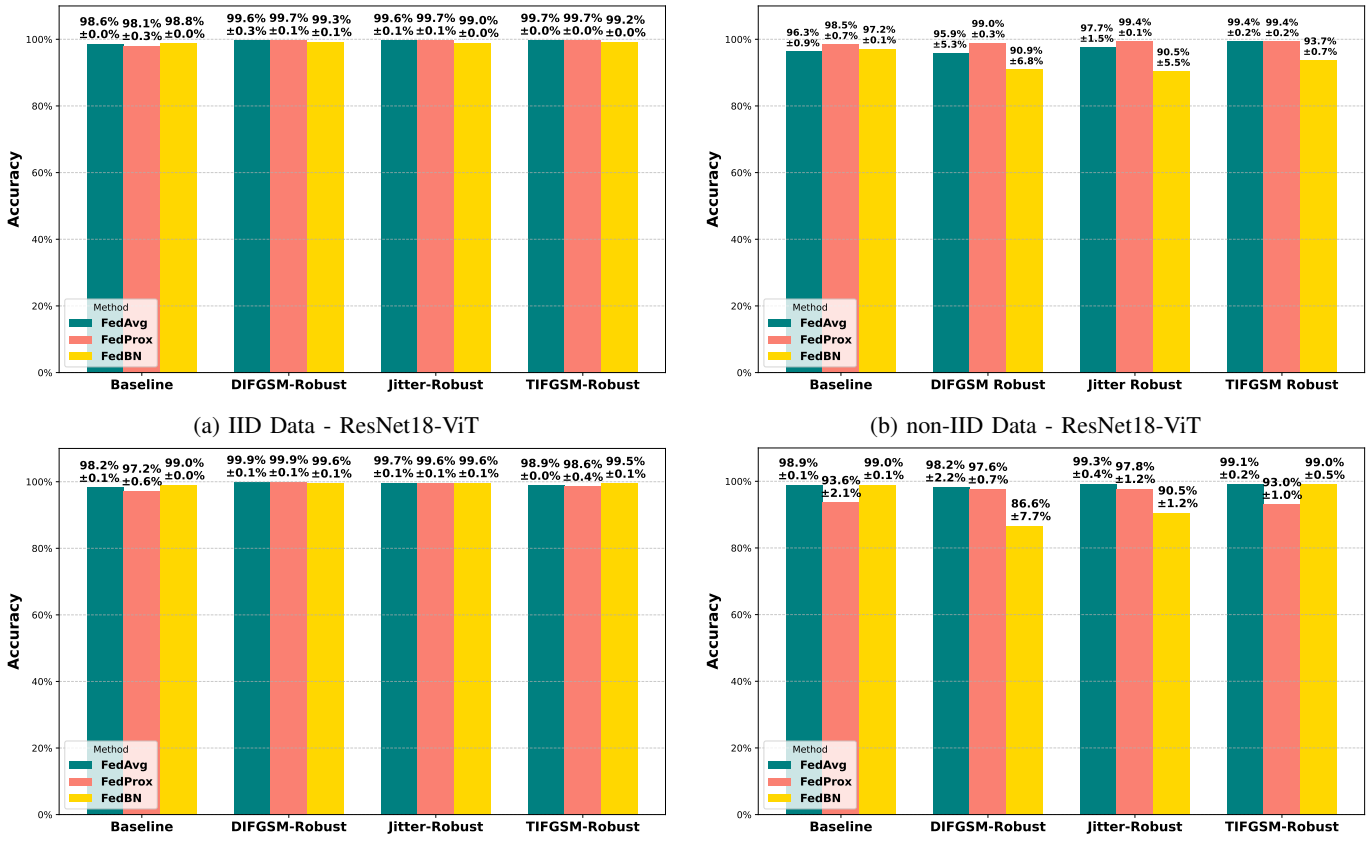
(a) IID Data - FedAvg-ViT (b) IID Data - FedProx-ViT (c) IID Data - FedBN-ViT

(d) non-IID Data - FedAvg-ViT (e) non-IID Data - FedProx-ViT (f) non-IID Data - FedBN-ViT

Fig. 3: Performance comparison of DenseNet-ViT models using FedAvg, FedProx, and FedBN, showing the accuracy of IID data (a, b, c) and non-IID data (d, e, f) for Jitter-robust, TIFGSM-robust, and DIFGSM-robust across diverse FL-ViT scenarios.

variabilities. This comprehensive evaluation ensures that our framework is resilient to both adversarial threats and naturally occurring uncertainties in clinical deployment.

Regarding the performance of ResNet18-ViT across our FL algorithms, as shown in Fig. 2, although baseline methods like FedAvg, FedProx, and FedBN converge faster in the early stages, TIFGSM-robust, DIFGSM-robust, and Jitter-robust methods leverage high-fidelity data diversity, leading to superior final accuracy and generalization. The robust methods perform better, especially when dealing with both in-domain and out-of-domain empirical data distributions, across both heterogeneous and homogeneous environments. In our



(a) IID Data - ResNet18-ViT

(b) non-IID Data - ResNet18-ViT

(c) IID Data - DenseNet121-ViT

(d) non-IID Data - DenseNet121-ViT

Fig. 4: Test Accuracy Comparison of Non-Robust, DIFGSM-Robust, Jitter-Robust, and TIFGSM-Robust Methods under Diverse FL Frameworks (FedAvg, FedProx, and FedBN) for ResNet18-ViT (a: IID, b: non-IID) and DenseNet-ViT (c: IID, d: non-IID).

simulations, we used the Dirichlet distribution (Dir(α)) with $\alpha = 0.5$ to model heterogeneity in the clients' data. Experimentally validated, they more effectively define an optimal virtual radius around the empirical data distribution compared to baseline methods, ensuring each model is trained within a optimal region, thereby enhancing coverage of high-fidelity data and mitigating the impact of out-of-distribution variations.

When perturbations are applied to images using robust algorithms, significant fluctuations are observed during the training process. This is also evident in the FedBN setting, where the personalized batch normalization layers cannot maintain fixed batch parameters, resulting in relatively lower robustness in scenarios such as Non-IID settings. As shown in Subfigure 2f, this leads to improved performance for FedBN in Non-IID settings, highlighting its particular strengths in such scenarios. While this effect is also observed in FedProx compared to FedAvg, the inclusion of the proximal term reduces fluctuations due to the regularization term, which in turn enhances performance with a more robust nature. However, the advantage of the robust optimization methods over the regularization term is evident in these results.

In accordance with Fig. 3, the superiority of robust methods over baseline approaches is also evident for DenseNet121-ViT, with the issue of fluctuations being applicable here as well. It is also observed in Subfigures 3a to 3e that Jitter-robust and DIFGSM-robust outperform other methods, even surpassing TIFGSM-robust in both the IID and Non-IID settings across the aggregation scenarios of FedAvg and FedProx. Meanwhile, in the FedBN setting, FedBN and its variant, TIFGSM-robust, demonstrate superior performance compared to the other methods in the Non-IID setting, as observed for ResNet18-ViT. Furthermore, regarding the training stability, while FedAvg and FedProx exhibit relatively smooth learning curves, the implementations of FedBN and robustness methods introduce notable oscillations, as evidenced in this and previous figures. We present Fig. 4 to visualize the final test performance.

To rigorously compare baseline and robust methods, we define multiple attack scenarios and present the results in TABLE I, showing how gradient-based, optimization-based, and random noise attacks affect model robustness while providing a clear view of the trade-off between test accuracy and adversarial resilience. We assess performance using key evaluation metrics, including accuracy; F1-score, which balances precision and recall to provide a more informative measure for imbalanced datasets; and AUC-ROC score, which evaluates the model's discriminative capability across varying classification thresholds. Additional evaluation criteria are made available in the GitHub repository for further analysis.

The results indicate that the performance of the FL methods varies significantly under different adversarial conditions. For both ViT models (ResNet18-ViT and DenseNet121-ViT), the introduction of adversarial attacks to the baseline methods

TABLE I: Accuracy, F1-Score, and AUC-ROC Score of Resilient FL-ViTs Learning Under Various Adversarial Attacks.

		Attack	Fedavg				Fedprox				FedBN			
			Baseline	DIFGSM	Jitter	TIFGSM	Baseline	DIFGSM	Jitter	TIFGSM	Baseline	DIFGSM	Jitter	TIFGSM
Accuracy	ResNet18ViT	None	100	100	99.92	100	99.53	99.84	99.84	99.92	99.78	99.78	99.81	99.80
		GN(0.2)	48.91	96.64	98.75	48.59	48.91	90.08	99.69	15.00	48.61	48.30	61.08	48.89
		GN(0.4)	48.75	49.30	58.98	36.95	48.91	38.98	70.23	35.70	35.86	48.12	48.97	48.91
		FGSM	38.20	99.77	99.45	94.38	34.84	99.45	99.38	42.42	25.05	26.47	53.20	24.91
		PGD	14.06	99.53	99.38	25.78	3.28	99.45	99.30	5.23	17.86	27.39	31.70	27.73
		CW	11.25	100	99.61	97.66	0.16	99.53	99.69	14.53	12.38	53.59	70.14	36.41
		DeepFool	4.92	1.88	0.31	17.89	4.22	3.44	1.95	1.56	2.94	2.28	4.80	5.61
		BIM	15.63	99.53	99.38	27.58	4.22	99.45	99.30	5.00	19.42	25.69	30.38	26.63
		RFGSM	11.09	99.53	99.38	26.41	6.17	99.45	99.30	15.31	14.52	40.47	31.58	39.52
		TPGD	11.72	99.92	99.53	20.16	9.38	99.61	99.69	25.63	20.72	35.80	31.22	36.31
		TIFGSM	21.88	100	99.84	99.92	31.95	99.69	99.77	99.53	17.77	33.08	80.81	47.44
		EOTPGD	9.84	99.53	99.38	23.59	4.38	99.45	99.30	12.66	12.27	30.39	30.20	32.70
	DenseNet121ViT	None	99.68	99.92	100	99.84	99.06	100	99.92	99.68	99.68	100	100	99.92
		GN(0.2)	42.18	91.71	99.06	34.84	49.84	74.60	98.35	19.76	49.73	83.46	98.81	14.37
		GN(0.4)	49.84	38.04	53.51	32.81	49.84	36.95	60.78	14.37	45.92	40.35	54.46	14.37
		FGSM	70.07	99.60	99.84	91.40	59.06	99.53	99.60	40.15	71.87	99.45	99.68	71.01
		PGD	25.62	99.53	99.60	26.87	9.06	98.75	99.53	9.84	36.71	98.73	99.53	26.96
		CW	1.56	99.68	99.84	52.18	0.00	99.68	99.68	13.59	2.89	99.92	100	58.98
		DeepFool	2.18	2.18	0.54	2.57	8.20	5.39	2.50	3.43	3.28	3.35	1.17	2.81
		BIM	29.53	99.53	99.60	27.81	9.29	98.43	99.53	10.62	38.82	98.67	99.53	27.26
		RFGSM	28.98	99.53	99.68	30.46	14.29	99.21	99.53	18.51	32.82	98.85	99.53	29.81
		TPGD	25.85	99.76	99.84	24.37	14.14	99.53	99.60	15.85	26.35	99.14	99.70	23.81
		TIFGSM	47.50	99.84	99.92	99.45	58.28	99.84	99.84	98.90	52.71	100	100	99.67
		EOTPGD	31.09	99.53	99.68	26.01	13.98	99.21	99.53	11.71	35.65	98.82	99.53	25.84
F1-score	ResNet18ViT	None	1.00	1.00	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
		GN(0.2)	0.32	0.96	0.98	0.34	0.32	0.90	0.99	0.05	0.34	0.42	0.56	0.32
		GN(0.4)	0.32	0.42	0.54	0.32	0.32	0.25	0.68	0.18	0.19	0.44	0.32	0.32
		FGSM	0.31	0.99	0.99	0.94	0.32	0.99	0.99	0.41	0.20	0.24	0.49	0.25
		PGD	0.15	0.99	0.99	0.26	0.03	0.99	0.99	0.05	0.19	0.27	0.32	0.29
		CW	0.11	1.00	0.99	0.97	0.00	0.99	0.99	0.15	0.11	0.46	0.68	0.35
		DeepFool	0.04	0.01	0.00	0.16	0.04	0.03	0.02	0.01	0.02	0.02	0.05	0.05
		BIM	0.16	0.99	0.99	0.28	0.04	0.99	0.99	0.05	0.21	0.25	0.31	0.28
		RFGSM	0.11	0.99	0.99	0.26	0.06	0.99	0.99	0.15	0.15	0.41	0.32	0.41
		TPGD	0.12	0.99	0.99	0.20	0.10	0.99	0.99	0.26	0.21	0.36	0.32	0.37
		TIFGSM	0.20	1.00	0.99	0.99	0.26	0.99	0.99	0.99	0.15	0.33	0.79	0.48
		EOTPGD	0.10	0.99	0.99	0.24	0.04	0.99	0.99	0.13	0.12	0.30	0.31	0.35
	DenseNet121ViT	None	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	1.00	1.00	0.99
		GN(0.2)	0.37	0.91	0.99	0.18	0.33	0.74	0.98	0.13	0.33	0.83	0.98	0.03
		GN(0.4)	0.33	0.24	0.49	0.18	0.33	0.22	0.57	0.03	0.37	0.29	0.50	0.03
		FGSM	0.68	0.99	0.99	0.91	0.59	0.99	0.99	0.40	0.71	0.99	0.99	0.69
		PGD	0.29	0.99	0.99	0.27	0.10	0.98	0.99	0.10	0.38	0.98	0.99	0.26
		CW	0.01	0.99	0.99	0.50	0.00	0.99	0.99	0.12	0.02	0.99	1.00	0.58
		DeepFool	0.02	0.01	0.00	0.02	0.08	0.06	0.02	0.03	0.03	0.04	0.00	0.02
		BIM	0.34	0.99	0.99	0.28	0.10	0.98	0.99	0.10	0.41	0.98	0.99	0.27
		RFGSM	0.32	0.99	0.99	0.30	0.15	0.99	0.99	0.19	0.34	0.98	0.99	0.30
		TPGD	0.28	0.99	0.99	0.24	0.15	0.99	0.99	0.16	0.28	0.99	0.99	0.23
		TIFGSM	0.43	0.99	0.99	0.99	0.54	0.99	0.99	0.98	0.52	1.00	1.00	0.99
		EOTPGD	0.35	0.99	0.99	0.26	0.15	0.99	0.99	0.12	0.37	0.98	0.99	0.25
AUC-ROC-score	ResNet18ViT	None	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99
		GN(0.2)	0.50	0.99	0.99	0.50	0.49	0.99	0.99	0.53	0.49	0.75	0.77	0.49
		GN(0.4)	0.46	0.87	0.92	0.45	0.49	0.71	0.88	0.49	0.50	0.59	0.53	0.46
		FGSM	0.56	1.00	0.99	0.97	0.44	0.99	0.99	0.59	0.34	0.36	0.69	0.39
		PGD	0.29	0.99	0.99	0.43	0.13	0.99	0.99	0.23	0.30	0.37	0.47	0.42
		CW	0.27	1.00	1.00	0.99	0.14	1.00	1.00	0.39	0.29	0.84	0.87	0.57
		DeepFool	0.11	0.33	0.40	0.16	0.30	0.27	0.38	0.33	0.13	0.21	0.23	0.18
		BIM	0.31	0.99	0.99	0.43	0.14	0.99	0.99	0.23	0.33	0.36	0.46	0.41
		RFGSM	0.23	0.99	0.99	0.46	0.18	0.99	0.99	0.32	0.25	0.53	0.50	0.56
		TPGD	0.26	1.00	0.99	0.42	0.22	0.99	0.99	0.52	0.32	0.48	0.50	0.51
		TIFGSM	0.32	1.00	0.99	0.99	0.26	0.99	0.99	0.99	0.27	0.41	0.92	0.62
		EOTPGD	0.23	0.99	0.99	0.45	0.15	0.99	0.99	0.32	0.23	0.42	0.48	0.50
	DenseNet121ViT	None	0.99	0.99	1.00	0.99	0.99	1.00	1.00	0.99	0.99	1.00	1.00	0.99
		GN(0.2)	0.55	0.99	0.99	0.51	0.51	0.94	0.99	0.50	0.48	0.99	0.99	0.54
		GN(0.4)	0.47	0.72	0.82	0.52	0.48	0.62	0.79	0.45	0.46	0.72	0.81	0.46
		FGSM	0.91	0.99	0.99	0.94	0.77	0.99	0.99	0.51	0.92	0.99	0.99	0.77
		PGD	0.46	0.99	0.99	0.39	0.24	0.99	0.99	0.21	0.54	0.99	0.99	0.34
		CW	0.17	0.99	1.00	0.72	0.20	0.99	0.99	0.48	0.19	1.00	1.00	0.81
		DeepFool	0.29	0.37	0.34	0.17	0.40	0.59	0.36	0.28	0.24	0.47	0.42	0.22
		BIM	0.52	0.99	0.99	0.40	0.25	0.99	0.99	0.22	0.57	0.99	0.99	0.34
		RFGSM	0.51	0.99	0.99	0.44	0.30	0.99	0.99	0.30	0.49	0.99	0.99	0.40
		TPGD	0.46	0.99	0.99	0.38	0.28	0.99	0.99	0.30	0.41	0.99	0.99	0.37
		TIFGSM	0.67	0.99	0.99	0.99	0.73	0.99	0.99	0.99	0.71	1.00	1.00	0.99
		EOTPGD	0.54	0.99	0.99	0.39	0.31	0.99	0.99	0.24	0.54	0.99	0.99	0.35

leads to substantial drops in accuracy, F1-score, and AUC-ROC scores, with more sophisticated attacks causing greater degradation. These methods suffer a drastic decrease in performance, especially in terms of accuracy, where the values often drop below 20% under certain attack scenarios.

Among the evaluated algorithms, training with DIFGSM-robust, TIFGSM-robust, and Jitter-robust leads to a noticeable improvement in accuracy and F1-score, suggesting that these approaches enhance the model's ability to generalize in both adversarial and attack-free scenarios. Interestingly, the extent of robustness improvement varies depending on the FL aggregation method and architecture. FedAvg and FedProx for ResNet18-ViTs, in particular, benefit the most from adversarial training, showing the highest retention of robust performance across the evaluated metrics. Overall, these adversarially robust optimization approaches emerge as successful methods within the most commonly used FL frameworks, which require trust besides efficiency, such as in clinical disease prediction.

It's worth mentioning that the effectiveness of robust methods in FL depends on the interplay between perturbation strategies, aggregation mechanisms, model architecture, and data distribution. FedAvg struggles with client drift in non-IID settings, while FedProx stabilizes training through regularization, making it well-suited for gradient-based adversarial

methods like DIFGSM-robust. FedBN, with its client-specific batch normalization, pairs effectively with TIFGSM-robust, as it enhances feature consistency.

In addition, model architecture plays a crucial role in robustness. ResNet18-ViT benefits from FedBN's localized normalization, stabilizing adversarial perturbations from TIFGSM-robust, while DenseNet121-ViT's dense connections inherently resist spatial and adversarial distortions, making it more effective with Jitter-robust and DIFGSM-robust. IID settings require minimal adaptation, where Jitter-robust suffices, while non-IID scenarios demand tailored strategies like TIFGSM-robust and DIFGSM-robust, which FedBN and FedProx effectively normalize and aggregate. The alignment of perturbation type, FL framework, and architectural resilience ensures optimal robustness and accuracy across diverse FL settings, as validated by our experiments.

Conclusively, although all robust approaches show significant improvements in adversarial resilience compared to common baseline FL methods, our results demonstrate that the Jitter-robust method achieves the optimal balance between accuracy and robustness. With an average accuracy of 81.34% across both attack-free and adversarial scenarios, the Jitter-robust method outperforms DIFGSM-robust (77.86%), TIFGSM-robust (39.81%), and the baseline (31.88%), demonstrating its superior performance in maintaining high accuracy while ensuring strong resilience against attacks in both attack-free and attacked conditions.

V. Existing Challenges and Future Directions

Although robust FL-ViTs, particularly jitter-robust, effectively balance robustness and accuracy in both homogeneous and heterogeneous scenarios, maintaining computational efficiency remains a challenge. As shown in TABLE II, we optimize the robustness configuration to minimize additional computational overhead, resulting in an approximate 1.22 times increase in runtime compared to the baseline with the same number of optimization iterations. In addition, another remaining challenge is addressing optimization-based adversarial attacks, such as CW and DeepFool. Our robustness strategies rely on gradient-based approaches, which may be insufficient against such sophisticated adversarial strategies.

In addition, practical challenges such as varying network conditions and client availability are expected to impact system performance. To address this, we simulate real-world scenarios in our experiments where clients may intermittently drop out or experience unstable connections, alongside results obtained with stable connections (see the GitHub repository). Our findings show that the system maintains robust performance despite increased time overhead. However, this introduces a trade-off: although the model still converges with the remaining clients, it does so at a slower rate. Client dropouts can also disrupt noise generation mechanisms, jeopardizing privacy compliance if too few clients participate in secure aggregation. To mitigate this, we suggest increasing per-client noise in such cases in future work, though this could introduce excessive noise when more clients are present. These points emphasize the need for efficient noise generation strategies and scalable

TABLE II: Runtime of IID dataset Different Methods (sec)

	Structure	Robustness			
		Baseline	DIFGSM	Jitter	TIFGSM
Fedavg	ResNet18	3953	5060	5088	5066
	DenseNet121	13246	14980	15530	15259
FedProx	ResNet18	4109	5248	5236	5287
	DenseNet121	13485	15552	15744	15872
FedBN	ResNet18	5104	6177	6233	6189
	DenseNet121	16303	21180	20732	18801

protocols to handle intermittent client availability. Moreover, FL's inherent privacy benefits make it particularly suitable for applications like healthcare data sharing, as data remains on local devices, ensuring compliance with regulations like HIPAA and GDPR. Nonetheless, challenges related to scaling federated systems and managing client dropouts remain critical areas for further research, especially in large-scale, real-world deployments.

For future work, optimizing the computational efficiency of robust FL-ViTs can be achieved through techniques such as low-rank approximations in self-attention, quantization-aware training, and adaptive gradient sparsification to reduce redundancy while maintaining robustness. Additionally, integrating adversarial purification methods, such as diffusion models or certified defenses, could strengthen resilience against optimization-based attacks like CW and DeepFool. Exploring meta-learning strategies for adaptive robustness tuning across heterogeneous clients may further enhance generalization in non-IID settings. Finally, developing hybrid architectures that combine convolutional inductive biases with transformer-based representations can provide a more efficient trade-off between robustness, accuracy, and computational complexity in FL.

VI. Conclusions

We developed a resilient approach for Alzheimer's prediction with FL-ViTs, harnessing robust optimization to enhance FL-ViT performance. Our proposed algorithm employed Jitter-robust, TIFGSM-robust, and DIFGSM-robust methods within a FL framework, aimed at strengthening model resilience against adversarial attacks while effectively managing non-IID medical imaging data. This approach optimized the robustness of Vision Transformers, enabling them to maintain high predictive accuracy in the face of diverse challenges and uncertainties commonly encountered in healthcare applications. Our experiments demonstrated that the ResNet18ViT and DenseNet121 architectures, combined with FL strategies, achieved significant predictive performance while maintaining resilience against adversarial attacks, with FedBN proving particularly effective for non-IID medical imaging data. Our systematic evaluation of robustness strategies, notably the Jitter-robust approach and DIFGSM implementation, revealed that DIFGSM exhibited remarkable resilience against multiple attack vectors when implemented with FedAvg and FedProx. However, its performance varied across different architectures and showed some vulnerability to Gaussian noise attacks, while it demonstrated strong performance when paired with DenseNet121ViT in FedBN scenarios. Overall, our results suggest that the Jitter-robust approach effectively balances accuracy and adversarial robustness, performing more reliably than the other robustness methods. These findings highlighted

the potential of resilient FL-ViTs in precise and reliable medical diagnosis systems, establishing a strong foundation for future research in robust FL for healthcare applications where data diversity and model reliability are critical concerns.

VII. REFERENCE

[1] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[2] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[3] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020.

[4] P. Mohassel and Y. Zhang, "Secureml: A system for scalable privacy-preserving machine learning," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 19–38.

[5] V. Mubonanyikuzo, H. Yan, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Detection of alzheimer disease in neuroimages using vision transformers: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, p. e62647, 2025.

[6] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *Acm computing surveys (csur)*, vol. 53, no. 2, pp. 1–33, 2020.

[7] M. Dehghani and Z. Yazdanparast, "From distributed machine to distributed deep learning: a comprehensive survey," *Journal of Big Data*, vol. 10, no. 1, p. 158, 2023.

[8] E. Darzi, Y. Shen, Y. Ou, N. M. Sijtsema, and P. M. van Ooijen, "Tackling heterogeneity in medical federated learning via aligning vision transformers," *Artificial Intelligence in Medicine*, vol. 155, p. 102936, 2024.

[9] X. Zuo, Y. Luopan, R. Han, Q. Zhang, C. H. Liu, G. Wang, and L. Y. Chen, "Fedvit: Federated continual learning of vision transformer at edge," *Future Generation Computer Systems*, vol. 154, pp. 1–15, 2024.

[10] E.-G. Marwa, H. E.-D. Moustafa, F. Khalifa, H. Khater, and E. AbdElhalim, "An mri-based deep learning approach for accurate detection of alzheimer's disease," *Alexandria Engineering Journal*, vol. 63, pp. 211–221, 2023.

[11] M. Joshi, A. Pal, and M. Sankarasubbu, "Federated learning for healthcare domain-pipeline, applications and challenges," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 4, pp. 1–36, 2022.

[12] A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, and A. Karargyris, "A review of medical federated learning: Applications in oncology and cancer research," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 3–24.

[13] Y. Bi, A. Abrol, Z. Fu, and V. Calhoun, "A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data," *bioRxiv*, pp. 2023–07, 2023.

[14] S. Sarraf, A. Sarraf, D. DeSouza, J. Anderson, and M. Kabia, "The alzheimer's disease neuroimaging initiative ovitad: Optimized vision transformer to predict various stages of alzheimer's disease using resting-state fmri and structural mri data," *Brain Sci*, vol. 13, p. 260, 2023.

[15] C. Chen, H. Wang, Y. Chen, Z. Yin, X. Yang, H. Ning, Q. Zhang, W. Li, R. Xiao, and J. Zhao, "Understanding the brain with attention: a survey of transformers in brain sciences," *Brain-X*, vol. 1, no. 3, p. e29, 2023.

[16] A. Rahman, M. S. Hossain, G. Muhammad, D. Kundu, T. Debnath, M. Rahman, M. S. I. Khan, P. Tiwari, and S. S. Band, "Federated learning-based ai approaches in smart healthcare: concepts, taxonomies, challenges and open issues," *Cluster computing*, vol. 26, no. 4, pp. 2271–2311, 2023.

[17] F. Castro, D. Impedovo, and G. Pirlo, "A federated learning system with biometric medical image authentication for alzheimer's diagnosis." in *ICPRAM*, 2024, pp. 951–960.

[18] W. Wang, X. Li, X. Qiu, X. Zhang, J. Zhao, and V. Brusic, "A privacy preserving framework for federated learning in smart healthcare systems," *Information Processing & Management*, vol. 60, no. 1, p. 103167, 2023.

[19] A. P. Kalapaaking, I. Khalil, and X. Yi, "Blockchain-based federated learning with smpc model verification against poisoning attack for healthcare systems," *IEEE Transactions on Emerging Topics in Computing*, 2023.

[20] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu, "Feddpgan: federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia," *Information Systems Frontiers*, vol. 23, no. 6, pp. 1403–1415, 2021.

[21] R. Jin and X. Li, "Backdoor attack and defense in federated generative adversarial network-based medical image synthesis," *Medical Image Analysis*, p. 102965, 2023.

[22] C. Zhang, X. Meng, Q. Liu, S. Wu, L. Wang, and H. Ning, "Fedbrain: A robust multi-site brain network analysis framework based on federated learning for brain disease diagnosis," *Neurocomputing*, p. 126791, 2023.

[23] B. Lei, Y. Zhu, E. Liang, P. Yang, S. Chen, H. Hu, H. Xie, Z. Wei, F. Hao, X. Song *et al.*, "Federated domain adaptation via transformer for multi-site alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, 2023.

[24] A. Lakhan, M. A. Mohammed, M. K. Abd Ghani, K. H. Abdulkareem, H. A. Marhoon, J. Nedoma, R. Martinek, and M. Deveci, "Fdcnn-as: Federated deep convolutional neural network alzheimer detection schemes for different age groups," *Information Sciences*, p. 120833, 2024.

[25] R. Aso, S. Shiota, and H. Kiya, "Enhanced security with encrypted vision transformer in federated learning," *arXiv preprint arXiv:2308.00271*, 2023.

[26] S. Queyrut, Y.-D. Bromberg, and V. Schiavoni, "Pelta: shielding transformers to mitigate evasion attacks in federated learning," in *Proceedings of the 3rd International Workshop on Distributed Machine Learning*, 2022, pp. 37–43.

[27] I. Balelli, S. Silva, M. Lorenzi, and A. D. N. Initiative, "A probabilistic framework for modeling the variability across federated datasets," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 701–714.

[28] Y. Tian, S. Wang, J. Xiong, R. Bi, Z. Zhou, and M. Z. A. Bhuiyan, "Robust and privacy-preserving decentralized deep federated learning training: Focusing on digital healthcare applications," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

[29] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[30] P. Sahoo, S. Saha, S. Mondal, S. Chowdhury, and S. Gowda, "Vision transformer-based federated learning for covid-19 detection using chest x-ray," in *International Conference on Neural Information Processing*. Springer, 2022, pp. 77–88.

[31] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated split vision transformer for covid-19 cxr diagnosis using task-agnostic training," *arXiv preprint arXiv:2111.01338*, 2021.

[32] J. Chen, W. Xu, S. Guo, J. Wang, J. Zhang, and H. Wang, "Fedtune: A deep dive into efficient federated fine-tuning with pre-trained transformers," *arXiv preprint arXiv:2211.08025*, 2022.

[33] J. Tao, Z. Gao, and Z. Guo, "Training vision transformers in federated learning with limited edge-device resources," *Electronics*, vol. 11, no. 17, p. 2638, 2022.

[34] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7838–7847.

[35] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier, "Exploring misclassifications of robust neural networks to enhance adversarial attacks," *Applied Intelligence*, vol. 53, no. 17, pp. 19 843–19 859, 2023.

[36] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4312–4321.

[37] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.

[38] S. E. Sorour, A. A. Abd El-Mageed, K. M. Albarrak, A. K. Alnaim, A. A. Wafa, and E. El-Shafeiy, "Classification of alzheimer's disease using mri data based on deep learning techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 2, p. 101940, 2024.

[39] E. Darzi, F. Dubost, N. M. Sijtsema, and P. M. van Ooijen, "Exploring adversarial attacks in federated learning for medical imaging," *IEEE Transactions on Industrial Informatics*, 2024.

[40] Y. Chang, S. Laridi, Z. Ren, G. Palmer, B. W. Schuller, and M. Fisichella, "Robust federated learning against adversarial attacks for speech emotion recognition," *arXiv preprint arXiv:2203.04696*, 2022.