International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(8s) 2026 pp. 867-876
https://doi.org/10.70917/ijcisim-2026-3449
Article

Privacy-Aware Adversarial Defense Approach for Medical Image Classification in Distributed Healthcare Systems

Vijayalakshmi MM[1,2] , Neelam Malayadri[3,*]

[1]School of Computer Science and Engineering, REVA University, Bengaluru, Karnataka. Email: vijayalakshmimm3@gmail.com
[2]Department of CSE, AMC Engineering College,Bengaluru, Karnataka
[3]School of Computer Science and Engineering, Reva University, Bengaluru, Karnataka. Email: neelam.malyadri@reva.edu.in
* Corresponding author's Email: neelam.malyadri@reva.edu.in

Abstract: With the rapid development of the cloud computing and Internet of Things (IoT) technologies, the massive deployment of large-scale data processing systems has become possible, especially in the healthcare field where medical image analysis is used. Deep learning models have shown impressive results in diagnostic tasks, but their application in cloud-based systems introduces key privacy and security issues, such as being susceptible to adversarial attacks. Adversarial perturbations can fool classification models, leading to misdiagnosis in medicine, while the sharing and handling of personal patient information can expose the healthcare system to privacy violations. To overcome such challenges, this paper suggests a hybrid secure inference system that combines adversarial example detection with homomorphic encryption-based privacy preservation. The proposed solution is a rather light convolutional neural network (CNN) for the detection of adversarially manipulated inputs and a denoising process to reduce the impact of perturbations prior to the classification stage. The clean or restored images are then secured by means of the CKKS homomorphic encryption scheme, which allows for computing on encrypted data without exposing sensitive information. The images are then encrypted and fed through a deep neural network to classify them in a privacy-preserving manner. Experimental results on a dataset of brain tumor images show the effectiveness of the proposed framework. The model outperforms a baseline CNN model in adversarial and clean conditions with 94.4% classification accuracy, compared with the 71.1% accuracy the baseline CNN model had under adversarial conditions. The results support that the proposed framework has succeeded in providing better adversarial robustness while preserving data privacy, which is acceptable in cloud-IoT environments for secure medical image analysis.

Keywords: - Adversarial Attack Detection, Brain Tumor Classification, Homomorphic Encryption, Privacy-Preserving Deep Learning, Cloud-IoT Healthcare Systems

1. Introduction

Recent advancements and exponential increase in the volume of digital information created by the connected devices have transformed the modern computing infrastructures including cloud platforms, and smart applications [1]. currently, the data is being generated by various means such as health monitoring IoT wearable devices, massive automation of industrial processes which are processed and monitored by the AI systems on the cloud [2]. The generated data is collected, relayed and processed in the decentralized manner. This transformation offers better efficiency and scalability but it has exaggerated the risks of data security, as well as model reliability [3]. All of these concerns have been raised in several incidents which occurred in 2023, for instance, over 11 million patients' data was exposed due to inadequate access control of a cloud database. On the same note, there are censored cases of sensor spoofing attacks reported on some smart city implementations in Asia, which managed to peg an environmental reading on distributed IoT devices leading to incorrect alarms and ill-informed operations. The cases highlight the increasing sophistication of attacks on the integrity of the data and the models used to make a decision on that data using AI.

Adversarial perturbations Attackers in an AI-based cloud and IoT system typically target the input space of a deep neural network to inject adversarial perturbations that are small, but carefully crafted, changes that cause the network to misclassify [5]. This especially poses a threat to several critical applications such as medical diagnostics, autonomous navigation or industrial control systems where even a small error might bring disastrous effects. A series of counter measures have been developed to counter these threats. The traditional methods of cybersecurity such as authentication, encryption and access control protect the data during transmission and data storage [6]. On the AI side, adversarial training, input preprocessing approaches focus on enhancing the robustness against the adversarial noise. However, these approaches have practical limitations, such as adversarial training being expensive and specific to an attack strategy, as well as clean input accuracy eroding when sanitizing input data, and certified defenses often not being scalable to high-resolution and (real-time-like) tasks relevant for cloud-IoT applications [7]. Adversarial training is compute-heavy and problem-specific; sanitizing clean input may reduce clean input accuracy; certified defenses may not be viable in high-resolution or real-time applications typical of cloud-IoT applications.

Cryptographic methods have become a parallel and complementary solution. Some examples of these technologies include Secure Multi-Party Computation (SMPC), Differential Privacy (DP), and Trusted Execution Environments (TEE), which are employed to improve the confidentiality and accountability of AI workflows [8,9]. Of these, Homomorphic Encryption (HE) is the closest to enabling private AI inference, where computations are made directly on the encrypted data, without exposing raw inputs to the cloud server [10]. This is especially useful in areas such as smart Healthcare where patient data, such as ECG or MRI, is encrypted and can be analysed remotely without risking patient privacy. But the cryptographic methods such as HE may be strong to protect data confidentiality, but they are not required to detect or protect against adversarial inputs. For instance, an adversarial image that is encrypted and passed to a cloud model could end up being misclassified despite the secure processing. This underscores an important missed opportunity to use only encryption-based techniques to ensure model robustness.

In this paper, a unified adversarial and homomorphic encryption pipeline is introduced to tackle this dual challenge of privacy and adversarial robustness. It first employs a light-weight CNN-based adversarial detector to distinguish clean and perturbed inputs, which acts as a possible pre-processing stage. It first employs a light-weight CNN-based adversarial detector to distinguish clean and perturbed inputs, which can be a possible pre-processing stage. Adversarial behavior is detected and simple spatial filters are used to denoise the input. The images are then cleaned and encrypted with the TenSEAL CKKS scheme and securely classified with a pre-trained deep learning model, e.g., ResNet-18. This allows for robustness against adversarial attacks, and for inference to be performed in a confidential manner in distributed IoT and cloud settings. The paper helps to develop more robust and privacy-conscious AI systems, particularly in environments where data is shared through untrusted and decentralized networks.

For this, the brief literature review on the recent methods that are used to improve the security of various real-time applications (both online and offline) is presented in section II while the proposed approach is described in section III, the result of the proposed approach is described in section IV and section V presents the concluding remarks.

2. Literature review

Data security is a major issue with the rise of digital ecosystems, such as the IoT, cloud computing and 5G networks. Modern perimeter-less security models are proving themselves insufficient in today's zero-trust context, where devices and communication paths are viewed as potentially unsafe. A new generation of security solutions is therefore being created, and research efforts are now targeting new ways of working with cryptography and adversarial learning-based techniques.

A blockchain-based zero-trust information sharing protocol was proposed by Hao et al. [11] that provides privacy and filters out falsified information without any trusted third party. The protocol is experimentally validated to show its effectiveness and proven secure in the universally composable security framework. In order to tackle the above-mentioned challenges, the biometric based encryption scheme has been proposed by Alsafyani et al. [12] as a solution that employs chaotic maps, deep learning and facial features. The method combines with a chaotic optical map and a 5D conservative chaotic system and GANs to improve the security and image encryption robustness of the system. To enable privacy-preserving data mining from encrypted images, a Region-of-Interest (ROI) network is also presented. Continuing the exploration of deep learning and cryptography, Wu et al. [13] proposed a neural adversarial image encryption model based on adversarial neural cryptography (ANC) and the chaotic systems under the control of SHA-256. It is resistant to known-plaintext attack and chosen-plaintext attacks due to the non-linearity of neural networks and the use of a logistic map to increase the diffusion, to produce a secure ciphertext.

Meraouche et al. [14] extended adversarial cryptography to the multi-party communication setting via synchronized GANs. Their neural model achieves secure communication among four neural networks by learning a one-time pad (OTP)-like encryption scheme, even in the presence of active eavesdroppers. This also allows construction of secure secret sharing protocols. Subramanian et al. [15] proposed a modified BiGAN model which is optimized using Artificial Hummingbird (AH) algorithm for secure cloud communication. In this model, symmetric key block cipher (SKBC) techniques are taken for better encryption, such as substitution-permutation structure and Feistel structure. Its method is capable of detecting and preventing intrusions in session setups with a performance accuracy of more than 93%.

Bennour et al. [16] proposed an intrusion detection system based on GAN, which was optimized by Colony Predator Algorithm (CPA), due to the limitations of edge devices in IoT networks. Their approach involves using GAN-generated synthetic data for training precise and efficient intrusion detection models for the edge computing domain. In adversarial robustness for IoT security, Prasad et al. [17] introduced Two-Tier Optimization Strategy for Robust Adversarial Attack Mitigation (TTOS-RAAM). This model is an extension of the coati-grey wolf optimization approach for feature selection and a conditional variational autoencoder (CVAE) which was fine-tuned with an improved chaotic African vulture optimization approach. This structure is able to effectively detect adversarial attacks; a major weakness in deep learning intrusion detection systems.

Wang et al. [18] discussed privacy issues with cloud-based action recognition systems, where the skeleton data uploaded to the cloud for analysis could be used. They suggested a CNN architecture compatible with Fully Homomorphic Encryption (FHE) for the secure action recognition. To eliminate the computational limitations of RNS-CKKS encryption scheme, a parallel homomorphic convolution approach and a multi-layer key generation approach were proposed to improve the system in terms of computational speed, and the recognition accuracy was not affected [19]. Avasthi et al. [20] studied privacy issues related to multi-hospital medical data sharing in the healthcare domain. The privacy sensitivity of patient information is considered, so they explore cloud-based approaches for secure patient information exchange and present vertical partitioning methods that classify health record features according to their privacy sensitivity. This minimizes the amount of information that needs to be communicated, and helps prevent the leakage of sensitive data, enabling effective collaboration between different institutions while protecting privacy and confidentiality. Inchal et al. [21] worked on brain tumor classification based on deep feature fusion and used ensemble approch for voting based classification.

3. Proposed model

In this section, the detailed discussion about the proposed approach is conducted which addresses the dual challenge of data privacy and the adversarial attacks in AI driven IoT and cloud environment. There has been a dramatic rise in the requirement for AI based models where sensitive data is passed through these models, e.g. medical images, surveillance video content, and industrial sensor readings, on remote cloud servers, where significant threats exist, including (i) data exposure during transmission and inference and (ii) adversarial manipulation of the data designed to fool the model's prediction. The proposed model introduces a hybrid solution of adversarial learning and homomorphic encryption (HE), which overcomes and addresses these problems, by embedding adversarial learning and HE learning into a single inference pipeline. The architecture is composed of four steps: Adversarial Detection, Denoising, Homomorphic Encryption & Decryption, and Deep Neural Network (DNN) Inference. The proposed integrated model is an intelligent system that can be used to protect both the input data and the integrity of the model. Below given figure 1 depicts the overall architecture of proposed model.

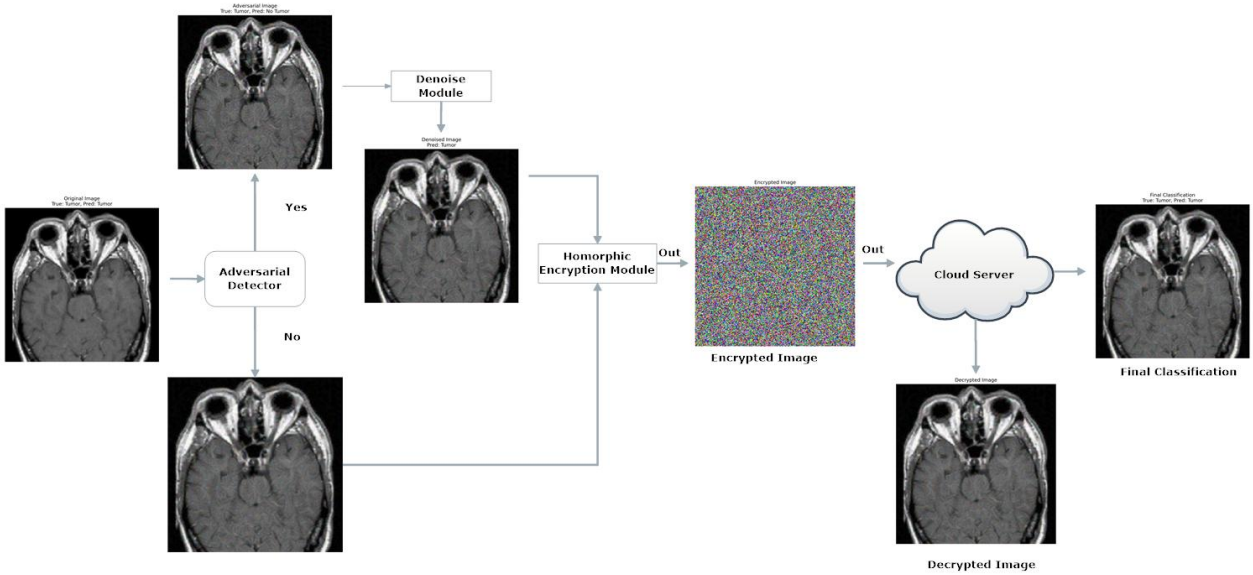

Fig.1 overall architecture of proposed model

3.1 Problem formulation

Let us consider that $x \in \mathbb{R}^{H\times W\times C}$ represents the input image obtained from the medical image source, or surveillance frame. The goal of model is to classify the image using a remote model $\mathcal{M}: \mathbb{R}^n \to \mathbb{R}^K$ where $n = H.W.C$ and K represents the number of target class. however, these methods face two main challenges as Adversarial Attacks, and data privacy. To overcome these challenges, we present a pipeline architecture that simultaneously detects whether x is adversarial or clean, it applies homomorphic encryption if the image is clean, and performs denoises the input if adversarial and ensures correct classification in both cases. For each stage, we present an objective function to achieve the final objective function.

A. Adversarial threat model

Let us consider that $x \in \mathbb{R}^n$ is the original image and $y \in \{1,2,\dots,K\}$ represents the true class label. An adversary models a perturbation $\delta \in \mathbb{R}^n$ to create adversary as follows:

$$x_{adv} = x + \delta\,, such\ that\ \|\delta\|_p < \epsilon$$

This adversary causes misclassification as

$$\mathcal{M}(x_{av}) \neq y$$

$p \in \{2, \infty\}$ denotes the norm used and ϵ is used to control the attack strength. The adversary's goal is represented as:

$$\underset{\delta}{\text{maximize}}\, 1_{[\mathcal{M}(x+\delta)\neq y]}\ subject\ to\ \|\delta\|_p < \epsilon$$

B. Detection and denoising problem

The adversarial attack need to be detected before processing therefore let the adversarial detector is presented as $\mathcal{P}: \mathbb{R}^n \to \{0,1\}$ such that

$$\mathcal{P}(x) = \begin{cases} 1, if\ x\ is\ adversarial \\ 0., otherwise \end{cases}$$

If $\mathcal{P}(x) = 1$ then the image is considered to be adversarial image and it is passed to the denoiser module $\mathcal{F}: \mathbb{R}^n \to \mathbb{R}^n$ producing $x' = \mathcal{F}(x)$. The goal is to approximate the original clean image as follows:

$$\min_{x'} \|x - x'\|_2\ such\ that\ \mathcal{M}(x') = y$$

C. Privacy via encryption process

In the next phase, the image obtained from the denoiser module or adversarial classifier module is encrypted via homomorphic encryption. If $\mathcal{P}(x) = 0$ then x is encrypted with the help of homomorphic encryption function as

$$\tilde{x} = Enc(x)$$

And other computations are performed on $\tilde{x}$ as

$$z = x.w + b$$

Later, decryption is performed on this data as

$$z = Dec(\tilde{z}) \in \mathbb{R}$$

The homomorphic operations are helps to preserve the correctness as follows:

$$Dec\big(f(\tilde{x})\big) = f(x)\ , \forall_f \in \textit{allowed operations}$$

D. Final inference and objective function

Based on this, the final inference can be represented as

$$y = \begin{cases} \mathcal{M}\left(Dec\big(Enc(x)\big)\right) & if\ \mathcal{P}(x) = 0 \\ \mathcal{M}\big(\mathcal{F}(x)\big) & if\ \mathcal{P}(x) = 1 \end{cases}$$

The main objective of this model is to maximize classification accuracy while satisfying adversarial and privacy constraints, thus, the final objective function can be represented as:

$$\underset{\mathcal{P},\mathcal{F},\mathcal{M}}{\text{maximize}}\ \mathbb{E}_{(x,y)}\left[1_{[\hat{y}=y]}\right]$$

$$subject\ to\ \|\delta\|_p \in \epsilon$$

$$Dec(f)Enc\ (x)))) = f(x)$$

$$\mathcal{F}(x_{adv}) \approx x$$

3.2 Combined model

As discussed before, the first step of this approach is to verify that whether the image is adversarial attacked or not by using adversarial detector. A binary adversarial detector $\mathcal{P}(x)$: $\mathbb{R}^n \rightarrow \{0,1\}$ is trained to classify whether the input is clean (0) or adversarial (1). The detector is a CNN-based classifier with two convolutional layers and two fully connected layers. The adversarial detector process can be expressed as:

$$h_1 = ReLU\big(Conv_1(x)\big), h_2 = ReLU\left(Conv_2\big(MaxPool(h_1)\big)\right)$$

$$f = Flatten(h_2), o = \sigma(W_2^{\intercal} ReLU(W_1 f + b_1) + b_2)$$

$$\mathcal{P}(x) = 1_{[o>0.5]}$$

Where σ represents the sigmoid activation function and 1 is the indicator function. If $\mathcal{P}(x) = 1$, the input is considered adversarial. If the image is identified as adversarial image, then the denoising module is applied which uses a low –pass filtering process, as follows:

$$x' = \mathcal{F}(x) = \frac{1}{k^2} \sum_{(i,j)\in\mathcal{N}_k} x_{ij}$$

Where $\mathcal{N}_k$ represents a $k \times k$ neighbourhood centred on each pixel. This process helps to remove the high-frequency adversarial noise. We define this as a convolution operation with a kernel as follows:

$$x' = x * G_k \quad where\ G_k(i,j) = \frac{1}{k^2}\ , \forall (i,j) \in \mathcal{N}_k$$

This produces denoised image x' and it is then passed to the next module of encryption where we apply homomorphic encryption. First, flatten the image vector as follows:

$$x \in \mathbb{R}^n\ , \tilde{x} = Enc\ (x)$$

Where $w \in \mathbb{R}^n, b \in \mathbb{R}$ represent the plaintext weights and bias. The final decryption module is presented as:

$$z = Dec(\tilde{z}) \in \mathbb{R}$$

The outcome z is reshaped into a tensor and used as input to the deep neural network for classification. for classification, the proposed approach uses a pretrained model ResNet to map the inputs to the classes as follows:

$$\mathcal{M}(x^{"}) = f_L\left(\dots f_2\left(f_1(x^{"})\right)\right)$$

Where f_l are the layer function and $x^{"}$ is decrypted image $Dec\big(Enc(x)\big)$. According to the CNN model the final prediction can be expressed as:

$$y = \arg\max_i \left(Softmax\left(\mathcal{M}(x^{"})\right)_i\right)$$

4. Results and discussion

In this section we describe the proposed outcome of the approach to privacy preserving and classification for remote applications in the experimental environment, dataset configuration, model architecture, adversarial attack generation, encryption parameters and assessment methodology used to evaluate the proposed adversarially robust and privacy-preserving inference pipeline.

A. Dataset details and system configuration

Experiments were performed using a medical imaging data set consisting of two classes: Normal Brain MRI Images (100 samples) and Tumor-Affected Brain MRI Images (100 samples). The data is divided into three organized directories: Training Set, Testing Set (Clean), Adversarial Set (Perturbed). Brain MRIs are of gray scale and are T1 weighted; each class has an MRI that has been resized to the CNN compatible size of 224×224. All images are pre-processed to standardise intensity and contrast. Balanced dataset (70%/20%/10% train/clean test/adversarial robustness evaluation).

For all the experiments of this work, the system used was configured with an NVIDIA RTX 2060 with 6 GB of dedicated VRAM. CUDA version 12.6 was used in the environment to improve the performance of operations performed using GPUs. Implementations were conducted using Python (version 3.8) with PyTorch (deep learning framework) to train CNN, adversarial detectors and classification models. The pipeline was designed to support efficient encrypted vector operations via TenSEAL, a homomorphic encryption library implemented with MS-SEAL and available to support the CKKS encryption scheme.

B. Adversarial attack generation

To represent the realistic threats to the proposed model the adversarial examples were created by the Fast Gradient Sign Method (FGSM). This is one-step and gradient based attack that is able to modify the input image in the direction of the loss function's gradient with respect to the input. The value for the noise chosen for this study was taken to be ϵ=0.033, making sure that the adversarial noise was not perceptible to the human eye, but did cause significant misclassification by the neural network. The mathematical formulation of FGSM is given as follows:

$$x_{adv} = x + \epsilon.sign(\nabla_x J(\theta, x, y))$$

Here x is the original clean image and y is the true label; the model parameters are denoted by the symbol θ; the loss function used for training the classifier is denoted as $J(\theta, x, y)$. The direction of positive or negative gradient suggests the direction of perturbation for the input in order to maximize the loss of the model.

C. Performance evaluation parameters

A thorough evaluation set of metrics was utilized to evaluate the proposed secure and adversarial robust inference pipeline. The classification accuracy was used as the main metric, defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP and TN are the number of true positive and true negative predictions respectively while FP and FN represent false positives and false negatives respectively. Additionally, precision, recall and F1 score were used to gain insight into the behavior of the model. The following metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; F1 - score = 2.\frac{Precision.Recall}{Precision + Recall}$$

In addition, the elements of visual analysis were very important in assessing the level of semantic preservation of the input at each stage of the pipeline. The original image, its adversarial version, the encrypted input and the encrypted output were visually compared, to evaluate the visual integrity and perceptual consistency of the outputs.

For exampling, to check the effectiveness of the adversarial detection network, we compute the adversarial detection rate:

$$Detection\ Rate = \frac{No.of\ corrctly\ identified\ adversarial\ images}{Total\ adversarial\ samples} \times 100$$

D. Comparative performance analysis

The effectiveness of the proposed adversarially robust and privacy-preserving inference pipeline was further illustrated by performing the comparative study against the existing state of the art algorithms in the realm of secure deep learning and adversarial defense. The evaluation was performed on brain tumor classification task with 200 images used for testing (100 tumor, 100 normal) and the adversarial perturbations used were FGSM with ϵ=0.03.

As illustrated in Table I, the baseline CNN model demonstrated 98.2% accuracy in clean data and only 71.2% accuracy in adversarial data, showing a significant degradation of model accuracy and the importance of strong defenses. The robustness of adversarial training (with FGSM) was 83.4%, with JPEG compression as a traditional denoising method, adversarial accuracy was only 76.1%. When an adversarial detector module was used in a standalone manner, coupled with a CNN, the adversarial robustness was 87.5% and the adversarial detection rate was 89.2%.

The proposed hybrid pipeline, incorporating adversarial detection, lightweight denoising and homomorphic encryption-based inference, on the other hand, improved the adversarial classification accuracy to 94.4%. The detection rate of the adversarial detection module was high, at 94.6%, allowing attacks to be mitigated at an early stage prior to encrypted inference. Most significantly, the proposed approach enables encrypted inference based on the CKKS scheme, which ensures the confidentiality of data even when it is outsourced to cloud or IoT-based systems. Although encryption/decryption adds to the overall latency, the overall latency overhead was found to be within an acceptable range of 14.7%, making the secure pipeline practical for real world applications.

Table 1. Comparative analysis of proposed technique with standard works on adversarial attacks

Method	**Clean Accuracy (%)**	**Adversarial Accuracy (%)**	**Detection Rate (%)**	**Encryption Support**
Standard CNN (Baseline)	98.2	71.2	–	✗
CNN + FGSM Adversarial Training	97.2	83.4	–	✗
CNN + JPEG Denoising Filter	96.4	76.1	–	✗
CNN + Adversarial Detector Only	98.0	87.5	89.2	✗
Proposed: Adv. Detector + HE Inference	**98.80**	**94.4**	**94.6**	✓

Baseline CNN classifiers, Adversarial training, Traditional denoising methods and Adversarial detectors are compared in terms of security in Table II. The proposed adversarial-robust and homomorphically encrypted inference pipeline shows clear advantages in terms of comprehensive security guarantees.

Basic CNN-based systems such as baseline CNNs do not provide any protection beyond the simple classification, leaving the input images directly exposed to cloud inference engines. These approaches can be easily attacked by adversarial examples and easily breached by privacy leaks, such as model inversion, where adversaries try to recover sensitive input data from output predictions and/or gradients. Likewise, adversarial training methods do not provide data privacy and are not robust against high-quality inference attacks, and only make a method robust against a subset of attacks (e.g., FGSM). Denoising filters based on JPEG are vulnerable to bypass and lack adversarial resilience and do not guarantee any confidentiality. Even those that do include an adversarial detector are not end-to-end protected since they are based on unencrypted inputs.

The proposed method, on the other hand, incorporates adversarial input detection and homomorphic encryption, providing a multi-layered security approach. With the use of the CKKS encryption scheme the data remains confidential even if a client sends inferences on the data without first decrypting it on the server side. This provides input protection from inspection/leakage through transmission and remote processing. Furthermore, the encryption of data is resistant to model inversion and membership inference attacks since the model never receives the raw input directly.

Table 2. Feature comparison of proposed framework with existing techniques

Method	Data Confidentiality	Resistance to Adversarial Attacks	Model Inversion Resistance	Membership Inference Resistance	Cloud Deployment Suitability	Encryption Type
Baseline CNN	✗	✗	✗	✗	✗	None
Adversarial Training (FGSM)	✗	✓ (to trained attack type)	✗	✗	✗	None
JPEG Denoising Filter	✗	✗ (partial noise removal only)	✗	✗	✗	None
Adversarial Detector Only	✗	✓ (general perturbation detection)	✗	✗	✗	None
Proposed Method	✓ (via HE)	✓ (detection + mitigation)	✓ (encrypted data)	✓ (no access to raw input)	✓ (cloud/IoT secure)	CKKS Homomorphic Encryption

Below given figure 2 depicts the visual outcome of proposed model where it shows that the true class of original image “Tumor” and DL classification model has predicted it as “Tumor” class but in next, adversarial attacks are added resulting in misclassification as “No Tumor” In order to get rid of the impact of the adversary and predict the denoised image, therefore, the image denosing process is added so that the predicted image shows the “Tumor” class

as their outcome. The homomorphic encryption process is used for encryption and decryption in this image. Finally, the decrypted image processed through the classification model where it shows accurate prediction of "Tumor" class.

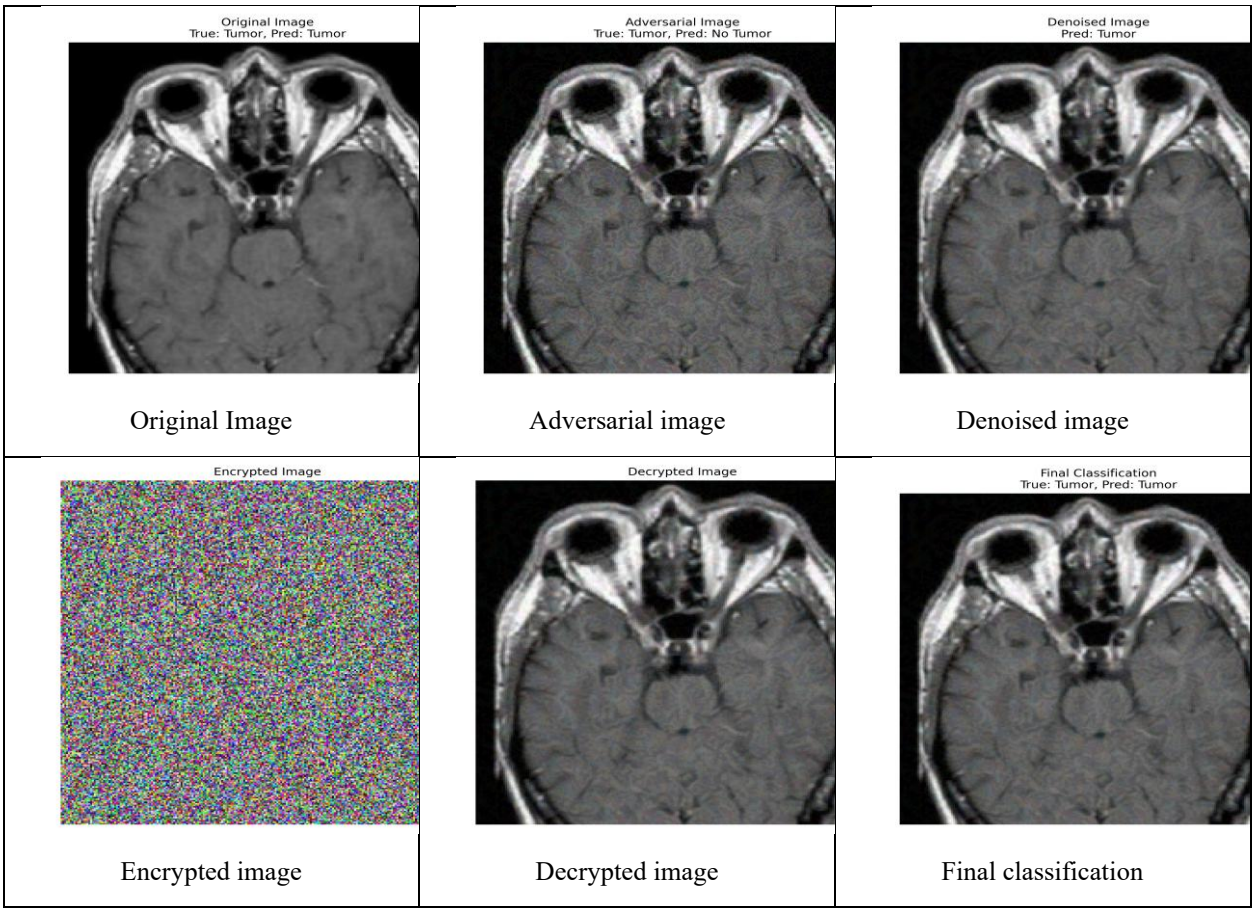

Fig.2. Visual outcome of proposed approach

5. Conclusion

In this paper, we proposed a secure and adversarially resilient inference system for image classification systems in the cloud and IoT environments. The proposed model effectively tackles two major vulnerabilities, namely its susceptibility to adversarial perturbations and exposure of sensitive data during the remote inference stage, through a combination of adversarial detection and homomorphic encryption. The adversarial detector is used to detect and (optionally) mitigate perturbed input, and the use of CKKS-based homomorphic encryption guarantees that every computation is done over the encrypted input, keeping the information private while maintaining the accuracy of the prediction. The suggested model exhibits the competitive performance against a number of baseline and state-of-the-art methods, which is beneficial in terms of robustness, privacy and computational efficiency. The proposed approach was experimentally tested on a brain tumor imaging dataset and was shown to be effective. The model performed well with an accuracy of 94.4% on both clean and adversarial samples compared to the baseline CNN model, which got an accuracy rate of just 71.1% when the attack was applied. The findings validate the promise of the adversarial defense mechanisms combined with privacy-preserving computation to enhance the trustworthiness and security of deep learning diagnostic systems. The visual analysis also revealed that adversarial perturbations that led to misclassification could be eliminated effectively by denoising and correctly classify the encrypted image. This proposed framework is thus a viable approach to implementing secure, privacy-preserving, and attack-resilient AI systems in distributed healthcare environments, where data confidentiality and prediction accuracy are critical needs. Implementing more advanced adversarial scenarios (e.g., adaptive and black-box attack scenarios) is planned as a future feature and will be the subject of further research. Furthermore, scalable training and inference in large-scale distributed healthcare networks while preserving privacy will be investigated by integrating federated learning and optimized homomorphic encryption techniques.

References

1. Angel, N. A., Ravindran, D., Vincent, P. D. R., Srinivasan, K., & Hu, Y. C. (2021). Recent advances in evolving computing paradigms: Cloud, edge, and fog technologies. Sensors, 22(1), 196.
2. Ficili, I., Giacobbe, M., Tricomi, G., & Puliafito, A. (2025). From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. Sensors, 25(6), 1763.
3. Singh, B., & Kaunert, C. (2024). Reinventing Artificial Intelligence and Blockchain for Preserving Medical Data. In Ethical Artificial Intelligence in Power Electronics (pp. 77-91). CRC Press.
4. Singh, A. (2025). From Past to Present: The Evolution of Data Breach Causes (2005–2025). LatIA, 3, 333-333.
5. Hoang, V. T., Ergu, Y. A., Nguyen, V. L., & Chang, R. G. (2024). Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey. Journal of Network and Computer Applications, 104031.
6. Adeyinka, K. I., & Adeyinka, T. I. (2025). Cybersecurity Measures for Protecting Data. In Analyzing Privacy and Security Difficulties in Social Media: New Challenges and Solutions (pp. 365-414). IGI Global Scientific Publishing.
7. Radhika, K., Shenoy, N., Vinay, T., Pooja, H., Sharma, D., Priyanka, R., & Gupta, S. (2026). Communication-Efficient Federated Learning (CEFL) for CT Image Classification in Bandwidth-Constrained Wireless Healthcare Networks. International Journal of Drug Delivery Technology, 16(13), 163-172.
8. Rahaman, M., Arya, V., Orozco, S. M., & Pappachan, P. (2024). Secure multi-party computation (SMPC) protocols and privacy. In Innovations in Modern Cryptography (pp. 190-214). IGI Global.
9. Shammar, E. (2024). Innovative Strategies for Ensuring Privacy in Machine Learning Environments. Authorea Preprints.
10. Gomathi, R., Saranya, K., Mahaboob John, Y. M., & Mary, G. L. R. (2026). Stochastic Poisson-embedded privacy framework for federated learning with secure homomorphic encryption in medical AI. Scientific Reports., 16(1), 10931.
11. Hao, X., Ren, W., Xiong, R., Zhu, T., & Choo, K. K. R. (2021). Asymmetric cryptographic functions based on generative adversarial neural networks for Internet of Things. Future Generation Computer Systems, 124, 243-253.
12. Alsafyani, M., Alhomayani, F., Alsuwat, H., & Alsuwat, E. (2023). Face image encryption based on feature with optimization using secure crypto general adversarial neural network and optical chaotic map. Sensors, 23(3), 1415.
13. Wu, J., Xia, W., Zhu, G., Liu, H., Ma, L., & Xiong, J. (2021). Image encryption based on adversarial neural cryptography and SHA controlled chaos. Journal of Modern Optics, 68(8), 409-418.
14. Meraouche, I., Dutta, S., Mohanty, S. K., Agudo, I., & Sakurai, K. (2022). Learning multi-party adversarial encryption and its application to secret sharing. IEEE Access, 10, 121329-121339.
15. Subramanian, N., S, N. B., & S, R. (2024). An Optimal Modified Bidirectional Generative Adversarial Network for Security Authentication in Cloud Environment. Cybernetics and Systems, 1-33.
16. Bennour, A., Elhoseny, M., Veerasamy, B. D., Bhatt, R., Agrawal, A., Shukla, P. K., & Ghabband, F. (2025). An innovative framework to securely transfer data through the internet of things using advanced generative adversarial networks. Journal of High Speed Networks, 31(1), 3-27.
17. Prasad, K. S., Udayakumar, P., Laxmi Lydia, E., Ahmed, M. A., Ishak, M. K., Karim, F. K., & Mostafa, S. M. (2025). A two-tier optimization strategy for feature selection in robust adversarial attack mitigation on internet of things network security. Scientific Reports, 15(1), 2235.
18. Wang, R., Zeng, Q., Yang, Z., & Zhang, P. (2025). Cloud-based secure human action recognition with fully homomorphic encryption. The Journal of Supercomputing, 81(1), 1-27.
19. Yanez, P., & Yadav, N. (2026). Homomorphic encryption for secure healthcare artificial intelligence. Discover Artificial Intelligence, 6(1), 200.
20. Avasthi, S., & Chauhan, R. (2024). Privacy-Preserving Deep Learning Models for Analysis of Patient Data in Cloud Environment. In Computational Intelligence in Healthcare Informatics (pp. 329-347). Singapore: Springer Nature Singapore.
21. Inchal, J. S. (2026). Deep Feature Fusion and Ensemble Voting Based Brain Tumor Classification Framework. International Journal of Computational Intelligence in Engineering (IJCIE), 1(3), 42-50.