Spike learning based privacy preservation of Internet of Medical Things in Metaverse

KHOWAJA,, Sunder Ali, DAHRI, Kamran, JARWAR, Muhammad Aslam <http://orcid.org/0000-0002-5332-1698> and LEE, Ik Hyun

Spike learning based Privacy Preservation of Internet of Medical Things in Metaverse

Sunder Ali Khowaja, *Senior Member, IEEE*, Kamran Dahri, Muhammad Aslam Jarwar, *Senior Member, IEEE*, and Ik Hyun Lee

***Abstract*— With the rising trend of digital technologies, such as augmented and virtual reality, Metaverse has gained a notable popularity. The applications that will eventually benefit from Metaverse is the telemedicine and e-health fields. However, the data and techniques used for realizing the medical side of Metaverse is vulnerable to data and class leakage attacks. Most of the existing studies focus on either of the problems through encryption techniques or addition of noise. In addition, the use of encryption techniques affects the overall performance of the medical services, which hinders its realization. In this regard, we propose Generative adversarial networks and spike learning based convolutional neural network (GASCNN) for medical images that is resilient to both the data and class leakage attacks. We first propose the GANs for generating synthetic medical images from residual networks feature maps. We then perform a transformation paradigm to convert ResNet to spike neural networks (SNN) and use spike learning technique to encrypt model weights by representing the spatial domain data into temporal axis, thus making it difficult to be reconstructed. We conduct extensive experiments on publicly available MRI dataset and show that the proposed work is resilient to various data and class leakage attacks in comparison to existing state-of-the-art works (1.75x increase in FID score) with the exception of slightly decreased performance (less than 3%) from its ResNet counterpart. while achieving 52x energy efficiency gain with respect to standard ResNet architecture.**

***Index Terms*— Privacy Preservation, Medical Images, Metaverse.**

I. Introduction

Metaverse has gained huge popularity and attention in recent times from industry and research alike. Metaverse transforms the physical world into the virtual one while providing a sense of realism. The history of metaverse dates back to 1992, however it was the Facebook's name to Meta that kind of made the metaverse a household name. Metaverse combines multiple emerging technologies such as cloud/Edge computing, Internet of Everything (IoE), computer vision, robotics, blockchain, and artificial intelligence to provide the promised immersive experience [1]. Although metaverse is at nascent stage now but its incessant evolution will help extend its boundaries to the field of entertainment, tourism industry, agriculture, education, finance, and healthcare [1].

Since the inception of COVID-19, researchers have been working on the adoption of Metaverse for healthcare applications as the metaverse can be integrated with the wearable sensors to provide continuous monitoring services in both real and virtual worlds. One of the potential applications of medical informatics in metaverse is the cancer detection. Cancer is considered as one of the leading causes of death throughout the world [2]. In 2018 alone, cancer was responsible for 9.6 million casualties [2]. The likelihood of survival from cancer heavily depends on the early diagnosis and detection, which also helps in reducing treatment expenses and morbidity, respectively. A definite, reliable, and precise cancer prognosis at its early stage is quite challenging due to the ambiguity in symptoms and images acquired from varying modalities. Therefore, providing an accurate cancer prognosis is required to reduce the morbidity rate and increase the chance of survival [2].

Metaverse provides a lot of benefits in healthcare domain, however, researchers have pointed out that it is vulnerable to security and privacy attacks [1]. Moreover, due to the potential of metaverse for collecting sensitive data, the stakes are significantly higher for the breach of security. The malicious user can manipulate the data or label that can cause misdiagnosis which can result in life altering situation. Therefore, the security and privacy concerns in metaverse, specifically for the medical data, has been the top priority for the researchers to consider.

Existing works have mostly focused on the image encryption techniques to preserve the data privacy [3], [5]. The process first encrypts the images at the user end and is decrypted at the receiver end for performing inference. The problem with the encryption techniques is that the stronger they are, the more computationally complex they get, which will hinder the metaverse experience eventually. Furthermore, studies have proved that the use of encryption techniques degrades the image quality to a certain extent, which is highly spurious for medical images that are quite sensitive and minor changes could result in false prognosis [4]. Recent methods are also capable of extracting and recreating data from pre-trained model weights, which is termed as model inversion attacks [6]. The aforementioned attack help in the designing of model poisoning attack [7]. To divulge such private information along with the class (prognosis) is considered to be a severe threat in the field of privacy preserving machine learning. Such information can be leveraged by the malicious attacker to be misused in physical world or generation of strong adversarial attack. Therefore, it is essential to develop a privacy preservation machine learning model that not only addresses data leakage but also the class leakage problem in metaverse, accordingly.

Recently, spiking neural networks have gained a lot of interest from researchers due to its capability of preserving privacy as well as lower power consumption [34]. Spike neural networks get activated only when certain events occur, therefore, researchers have shown that it yields 2x times less latency than conventional methods [34], which is quite beneficial specifically with the domain of metaverse. Researchers have also considered using CNNs and spike neural networks together in order to overcome the performance issues, especially in the domains of object detection, emotion classification, and cybersecurity [25]. Although, spike neural networks have been explored in existing studies for securing the weights, it exhibit the problems including reduced performance, weight conversion issues, and is mostly applied to traffic signs or moving objects.

In this regard, we propose generative adversarial networks (GAN) -

This work was supported by the Priority Research Centers Program (NRF2017R1A6A1A03015562) through the National Research Foundation of Korea (NRF) funded by the ministry of Education.

Sunder Ali Khowaja and Kamran Dahri are with Faculty of Engineering and Technology, University of Sindh, Pakistan. sandar.ali@usindh.edu.pk, kamran.dahri@usindh.edu.pk

Muhammad Aslam Jarwar is with Department of Computing, Sheffield Hallam University, Sheffield, United Kingdom. a.jarwar@shu.ac.uk

Ik Hyun Lee is with Department of Mechatronics Engineering, Tech University of Korea, Siheung-Si, Republic of Korea and IKLAB, Tech University of Korea, Siheung-Si, Republic of Korea. ihlee@tukorea.ac.kr

spike learning based convolutional neural network (CNN) transformation paradigm (GASCNN) to address the problems concerning class and data leakage in medical images. We address the data leakage problem by using GANs to generate synthetic data and then apply GAN-CNN transformation to camouflage the sensitive information. The class leakage problem is addressed by using encryption on weight parameters using spike learning strategy in a temporal manner. To be specific, we use spiking leak-integrate-and-fire (LIF) activation to optimize the parameters, which resists the calculation of backward gradients. The transformation from GAN-CNN also helps in reducing the computational complexity, as the spike based learning is computationally intensive for data generation. The transformation paradigm kind of regularizes the training process so that it can be used for edge devices as well. The contributions of this work are summarized below:

- To the best of our knowledge, this is the first work to address data and class leakage for medical images in the context of Metaverse.
- We proposed spike learning strategy to protect the parameter weights of CNN, which resists the recreation of data from model parameters.
- We propose the transformation paradigm from GAN-CNN to reduce the computational cost and enable its usage for edge devices.
- We carry out extensive analysis on publicly available dataset to show efficacy of GASCNN from privacy as well as energy-efficiency point of view.

The rest of the paper is structured as follows: Section 2 provides a consolidated literature review of existing works considering privacy and security of medical images. Section 3 presents the proposed methodology for GASCNN. Section 4 illustrates the experimental setup along with the analysis carried out to prove the efficiency of proposed work. Section 5 concludes the work and provides future research directions, accordingly.

II. Related Works

This section consolidates a brief literature review on the methods presented for data leakage, visual privacy, and model and feature inversion. The studies related to data leakage will highlight the works concerning privacy preservation machine learning models. The visual privacy section highlights the studies that focus on encryption techniques. The GANs and learnable encryption emphasize on the use of GANs for the adopting privacy preservation within the learning process and lastly, the model and feature inversion provides insights for the studies that focus on differential privacy attacks. This work proposes defense against data and class leakage attacks through GANs and encryption, which concerns all the aforementioned categories, respectively. The last subsection will provide a summarization as to how the proposed work is different and unique in comparison to existing ones.

A. Data Leakage

Existing studies on data privacy and leakage have been centered around learnability, statistics, information theory, anonymization, shift-keying methods, and closeness [8]. The aforementioned studies are good for static and small-scale datasets, however their performance falters with increasing scale and dynamic environments. The concept of differential privacy was introduced in [9] that prevents a malicious attacker to gain access to the user or data, respectively. Over the years, many encryption techniques have been proposed to prevent the data from being leaked. Romi et al. [10] proposed the multi-round encryption technique that used secure multichannel approach. The method was prone to weak security and large computation times. Yinan et al. [11] presented a selective encryption strategy with small key sizes and termed it as DNA origami cryptography. The method was effective in terms of computational complexity, but was vulnerable to brute force attacks. Kaur et al. [12] proposed a high-dimensional chaotic map by applying piecewise linear operations, which performed quite well in comparison to the aforementioned studies. The trend has been continued to recent times when variants of aforementioned encryption techniques and hyperchaotic maps are generated, nevertheless the problem of increased computation time still remains.

Similar methods for the prevention of data leakage are applied to medical images as well. A recent study [14] combined the characteristics of frequency domain and discrete cosine transform to encrypt images, which takes less computation time and is robust against many security attacks, but performs poorly against crop and noise attacks. Rehman et al. [13] proposed substitution boxes based on chaotic maps for encrypting X-ray and MRI images. They used a large key space, i.e., 2^{100} that could resist the brute force attacks, therefore the problem of computational complexity at the time of inference is prominent. Khowaja et al. [5] proposed the use of chaotic maps along with a secret key image and noise level addition. The encryption method showed better results than many chaotic map-based techniques both in terms of performance as well as computation times, but still it was not suitable for real-time inference systems. Another recent study [7] proposed the combination of discriminative and generative networks to simulate model inversion and poisoning attacks to show its effect on the recognition performance. The study showed that the data leakage and its subsequent poisoning quite reduce the recognition performance by upto 20%, which is quite significant in the domain of medical imaging. Cameiro et al. [15] proposed the use of hyperchaotic maps by fusing multiple grayscale images using bifurcation, phase, and Lyapunov diagrams. The method was quite effective, however, resulted in large computation times.

B. GANs and learnable encryption

Recently, some studies have shifted towards the use of generative adversarial networks (GANs) to perform data encryption and decryption. A study DeepEDN [16] was proposed to encrypt and decrypt using cycle-generative adversarial network (Cycle-GAN) on medical images. The study also showed that a specific region of interest instead of the whole image can also be decrypted and extracted using such methods. The term learnable encryption was introduced in Tanaka scheme presented in the study [17]. The idea behind Tanaka scheme is that the encryption will only be applied for humans rather than machines, therefore, the encrypted images would be directly learned by machines. An improvement over Tanaka scheme was proposed in [18] with the name SKK scheme. The method used independent keys for encryption rather than the same ones as performed in Tanaka scheme. The usage of individual keys will lead to efficient shuffling operation along with a large key-space that is resilient to both the brute-force and DNN based attacks, respectively. The SKK scheme would also require extensive data augmentation tasks in order to provide a reasonable amount of accuracy. The use of GANs eliminate the need for encryption key management, as shown in one of the recent works proposed in [19]. The method also uses cycle-GAN to transform the plain images to encrypted ones using the encoder part. Tang et al. [35] proposed the use of Markov-GAN method that used Simpson index to perform encryption using Markov images. The method was performed on traffic sign recognition dataset.

C. Visual Privacy

Visual privacy preservation techniques aim to camouflage the image data with certain operations such that the information cannot be determined by the attacks. The operations performed for visual privacy preservation include adversarial image perturbation, identity obfuscation, down scaling, mean shift filtering, and Gaussian blur [20]. Some recent studies have focused more on adding adversarial perturbations, such as one-pixel attack, multi-pixel attack, jacobian saliency map attack, carlini&wagner attack, fast gradient sign method attack, and so forth [21]. Studies perform these attacks intentionally in order to cope with the adversarial and differential privacy attacks on the image data. Ryoo et al. [22] proposed the learnable machines on downsampled data to cope with the attacks, however, such methods require a trade-off between the security and performance accuracy. Liao et al. [36] combined the characteristics of CNN and spiking neural networks such that the feature extraction is performed using the former network architecture while the latter learns to encrypt the data via surrogate gradient learning. The work was performed on motor imagery acquired using EEG signals.

D. Model and Feature Inversion

With the advancement in adversarial deep learning and machine learning technologies, researchers have shown significant interest in the ability of methods that can intercept a classifier for stealing data from model weights and parameters [6]. First, such kind of the model and feature inversion methods were designed around optimization techniques that were able to invert mid to low level features extracted from convolutional neural networks (CNN). Following these techniques, other methods used up-CNN for improving the inversion of mid-level representations [7]. These methods were good for low-level feature inversion, however, the results were not effective for inversion of high-level features, accordingly. Recently, methods based on GANs were proposed that were not only able to invert high-level features but also reconstruct the data from model parameters [6]. The study deep leakage from gradients started the conversation for the reconstruction of data from model weights, since then studies like improved data leakage from gradients [23], PGSL [21], Industrial Private AI [5], and SPIN [7] were proposed that leverage model inversion attacks for specific applications. Following the trends, many studies considered derivatives of the aforementioned works to perform data or class label leakages from gradients such as WDLG [37], DEFEAT [38], and GLAUS [39].

E. Difference between existing methods and ours

In comparison to existing works, our method can scale up to high-dimensional data while encrypting the medical images using learnable machines. Our work is different from existing works in terms of transformation paradigm, i.e., from GANs to spike learning-based CNNs in order to reduce the computational complexity of the encryption system. To the best of our knowledge, this is the first work to perform transformation of GANs to spike learning-based CNNs for medical images in the context of metaverse. Furthermore, we show that the proposed work is capable of achieving the best trade-off in terms of recognition performance and computational complexity.

III. Preliminary for Leakage problems

The conversion in the proposed work from ResNet to spike learning based CNN is leveraged from the study [24]. The conversion assumes that the model uses leak-integrate-and-fire (LIF) neuron, which is formulated in equation 1.

$$\mathcal{M}_p^t = \sum_q w_{pq} r_q^t + \vartheta \mathcal{M}_p^{t-1} \tag{1}$$

In the above equation $\mathcal{M}$ refers to the membrane potential, w_{pq} represent the weights between post-synaptic neuron p and pre-synaptic neuron q, and ϑ is the leak factor. The spike r_p^t is generated when the $\mathcal{M}$'s crosses the firing threshold ς. A soft reset is then performed to lower the value of $\mathcal{M}_p^t$ by ς. The transformation paradigm process is carried out by normalizing the firing threshold ς or the weights. The weights of pretrained ResNet are replicated to the attributed CNN or ResNet, followed by the consideration of maximum activation value to set the firing threshold. Subsequently, the maximum activation is computed for each layer and for each time step. The transformation paradigm is then processed from the first layer to all subsequent layers, respectively. Batch normalization is not used during spike learning, as spikes are zero mean values. However, the dropout layer is used as suggested in the existing studies [24], [25].

The transformation paradigm requires the whole training data to be converted from ResNet to spike learning based CNNs, which leads to the privacy issues and data leakage. Existing studies have suggested to use synthetic data rather than the actual data to perform the transformation paradigm, which is similar with the studies performing domain adaptation, model distillation, model compression, and quantization [24]–[26]. But by doing so, the performance of the recognition system is compromised to achieve an acceptable level of trade-off.

Other than the data leakage, the attacker can also target shape of the objects or patterns, which corresponds to the class leakage problem. The algorithm in Table 1 provides a straightforward pseudocode to generate image based on class representation. First, an input tensor is initialized with uniform distribution. Pre-softmax logits are then maximized via input noise using an iterative optimization strategy for a specific target class. The image is then smoothed using Gaussian blur to reduce gradients.

The algorithm makes a basic assumption that the gradients can be calculated in an exact manner for all layers. However, when it comes to spike neural networks, the gradients cannot be computed exactly due to the nature of LIF activation, which is non-differentiable [24], [25]. The only way to generate images representative of a specific class is to apply the reverse transformation paradigm, i.e., from CNN to ResNet. Generally, two types of techniques such as threshold scaling and weight scaling are used for reverse transformation. Even though, with the reverse transformation, it is not easy for the attacker to obtain weights for the ResNet as the attacker has to apply several combinations of scaling factor for each layer. The state-of-the-art transformation methods operate on threshold scaling mechanism [24], [25], that vary the threshold while ensuring the uniformity of weight parameters. In this study, the class leakage problem and transformation is addressed using threshold scaling technique. This is made possible by replacing the LIF to rectified linear unit (ReLU) neuron. By doing so, the attacker can simply use the algorithm in Table 1 to reconstruct the image of a specific class.

IV. GASCNN

This section provides details regarding the workflow and methodology of the proposed work. The overall system workflow is shown in Figure 1. We use encoder-decoder style generative adversarial networks (GAN) that would learn to generate images through Residual Network (ResNet) architecture. The pretrained ResNet is then used to perform the transformation paradigm, followed by the encryption of weights using spike learning rule.

A. GANs for Image Generation

To create a synthetic dataset, we leverage conditional GANs (CGAN) [27], which will generate brain MRI images while pre-

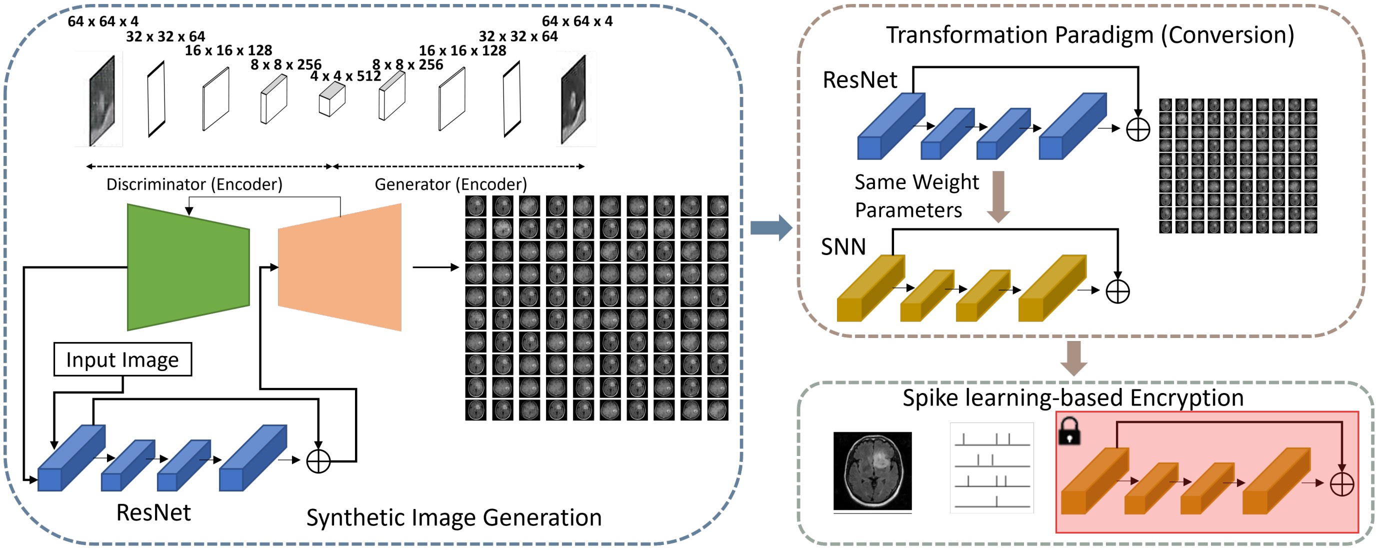

Fig. 1. The proposed Generative Adversarial Network - Spike Learning based CNN Transformation Paradigm Framework (GASCNN).

TABLE I
ALGORITHM FOR IMAGE GENERATION THROUGH CLASS LEAKAGE

```
Input: Number of Iterations ($N$), learning rate ($\eta$), target class ($\mathcal{L}$), blur frequency ($\mathcal{B}$)
Output: Image representing specific class $x$
Initialize uniform random distribution as input U(0,1) $\rightarrow x$
for 1 $\rightarrow n$ to $N$:
    if $\mathcal{B}$ % $n$ == 0:
        $GaussianBlur(x) \rightarrow x$
    end if
    Compute pre-softmax output $network(x) \rightarrow y$
    $\eta \frac{\partial y_{\mathcal{L}}}{\partial x} + x \rightarrow x$
end for
```

serving the patients' identity. The reason for using CGAN for the synthetic image generation is that even though the training data belongs to different sources, the inherent features can be modeled for different Alzheimer's stages. The synthetic image generation in this study is composed of three modules, i.e., attention-based GAN, discriminator, and a ResNet module, respectively.

Let us denote the 2D attention-based generator as Gen_a, synthetic image slices as SL_{2D}, target condition for disease stage as SC, and 2D discriminator as DIS, respectively. The generator architecture first combines SC and ResNet generated feature map Res_{2D} along channel axis through concatenation. We used categorical cross-entropy loss to train the ResNet and update feature maps, accordingly. The input image is fed to the ResNet which generates a feature map, which is then sent to the generator module. It should be noted that SC and Res_{2D} are of same dimensions, however, the SC comprises a single value, i.e. $[0, 0.5, 1]$, across all pixels representing the classes AD, MCI, or Normal. The weighted sum of the generator Gen_a and input Res_{2D} results in synthetic image slices as shown in equation 2.

$$SL_{2D} = g_a - ga \cdot \varrho + Res_{2D} \cdot \varrho \tag{2}$$

where ϱ and g_a are the outputs from Gen_a. The former represents the pixel-wise attention mask generated through sigmoid function, while the latter is the generated image through tangent hyperbolic function. The mask ϱ in this case helps in preserving features and image quality to help in aiding comparison, while the g_a maintains unrelated regions but at the same time modifies the brain areas considering the conditions instigated by Alzheimer's disease (AD) contortion. The Gen_a architecture comprises an encoder, decoder, and transition layer. Three convolutional layers are used in the encoder layer to extract feature maps. Six residual blocks are used in the transition layer to modify the feature maps in accordance with the target condition. The decoder generates two outputs. The first is the mask bounded by a sigmoid function that needs to be transformed outside the region, and the second is the transformed image that is generated using the tangent hyperbolic function and a transposed convolutional layer.

The discriminator DIS is designed to increase the quality of generated synthetic slice such that it reciprocates realism along with the verification of target condition in the generated image. The DIS comprise of 5 convolutional layers and 2 Fully connected layers that ensure the realism by verifying if the generated image is fake or real $\mathcal{P}_a$ and that the generated image either meets the target condition or not $\mathcal{P}_{sc}$. The input to the DIS can be either a generated synthetic image from generator or real image, accordingly.

Wasserstein GAN (WGAN) has been used to train the GAN in the proposed study with adversarial losses and gradient penalty [28] in order to satisfy Lipschitz constraint. The reasons for choosing WGAN and gradient penalty are three-fold. The first is the training stability, the second is the improved quality of the generated synthetic images, and the third is the fast convergence. The adversarial loss for the GAN is defined in equation 3.

$$\begin{aligned} Loss_{adv} = {} & (\|\nabla_z \mathcal{P}_a(z)\|_2 - 1)^2 \cdot \alpha - \mathbb{E}_{Res_{2D}}[\mathcal{P}_a(Res_{2D})] \\ & + \mathbb{E}_{Res_{2D},SC}[\mathcal{P}_a(SL_{2D})] \end{aligned} \tag{3}$$

The first term in the loss refers to the gradient penalty, while the remaining terms represent Wasserstein distance. The notation z represents the weighted sum of Res_{2D} and SL_{2D} with the assumption that it follows a uniform distribution $U[0, 1]$. The aforementioned assumption is in compliance with the existing works [21], [25], [34]. This weighted sum can be defined as shown in equation 4.

$$z = (1 - \mathcal{U}) \cdot SL_{2D} + \mathcal{U} \cdot Res_{2D} \tag{4}$$

The notation $\mathcal{U}$ refers to the uniform distribution [0,1]. The aforementioned loss ensures the quality of generated synthetic images, however, we also need to ensure the target conditions, which is carried out using regression loss. The regression loss is defined in equation 5.

$$Loss_{reg} = \mathbb{E}_{Res_{2D}}[\|\mathcal{P}_{sc}(Res_{2D}) - SC\|_2^2] \tag{5}$$

We also adopt an attention-based adaptive identity loss [29], which controls the degree of transformation via per pixel weighted combination between the Res_{2D} and SL_{2D}. The loss also ensures that the boundary artifacts are not added and the transformation is not excessive, which is essential to keep the realistic features intact in the generated images. The said loss is defined in equation 6.

$$Loss_{AAID} = \mathbb{E}_{Res_{2D},SC}[\|(Res_{2D} - SL_{2D}) \cdot (1 - \varrho)\|] \tag{6}$$

The reason for using attention-based adaptive identity loss is that the generated image may represent artifacts such as unnatural boundaries and instability in the training process that results in sharp changes concerning attention maps. The aforementioned loss ensures that the transformation is smoother and the boundary artifacts are reduced. The generator also uses attention loss to avoid saturation in the attention mask, i.e. $Loss_{att} = \mathbb{E}_{Res_{2D},SC}[\|\varrho\|]$. The total variation loss also ensures spatial smoothness in the generated image, and is defined in equation 7.

$$\begin{aligned} Loss_{TVR} =& \mathbb{E}_{Res_{2D},SC} \\ & \left[\sum_{u,v}^{H,W} [|\varrho_{u+1,v} - \varrho_{u,v}| + |\varrho_{u,v+1} - \varrho_{u,v}|] \right] \end{aligned} \tag{7}$$

Finally, the loss optimization functions for training the Generator and Discriminator are given in equation 8 and 9.

$$\begin{aligned} Loss_{Gen} &= \mathbb{E}_{Res_{2D},SC}[\mathcal{P}a(SL_{2D})] + \alpha_{TVR} \cdot Loss_{TVR} \\ &+ \alpha_{att} \cdot Loss_{att} + \alpha_{AAID} \cdot Loss_{AAID} + \alpha_{reg} \cdot Loss_{reg} \end{aligned} \tag{8}$$

$$Loss_{DIS} = \alpha_{reg} \cdot Loss_{reg} + Loss_{adv} \tag{9}$$

B. Transformation of ResNet - SNN

This study assumes that once the generator network and ResNet is trained from the previous stage, we do not have access to the original dataset. However, we presume that the trained ResNet is able to generate suitable features, which are essential for the generating realistic images as the ResNet is trained in accordance with the generator network in the previous stage. In order to go through the transformation process, we need to use the maximum activation value from each layer to compute the threshold value from synthetic images [26]. It has been shown by existing studies that synthetic images from transformed SNN result in the same performance when compared to the ResNet. We follow the process of transformation from ResNet to SNN as suggested in [24]–[26]. However, first we need to convert the synthetic data as well. Following are the steps to perform data free conversion.

- Initialize the synthetic dataset.
- For each class set compute the class similarity concerning ResNet's fully connected layer weights. Within the loop compute the following.
- For number of samples per class, use Dirichlet distribution to sample the soft labels and initialize input using uniform distribution.
- Also, find the sample that minimizes cross entropy loss concerning the feature maps of ResNet and associated label.
- Append the sample to converted dataset.
- Perform the transformation from ResNet to SNN using [24]–[26].

Once the transformation is completed, the SNN's are still vulnerable to class leakage attacks, as one can recover the original ResNet by accessing the weights of the transformed SNN and changing the neuron activation from LIF to ReLU. As medical images are considered to be private and sensitive data, it is therefore necessary to encrypt the SNN parameters using spike learning. For the conversion, the study uses small-time steps such as 70 $\sim$ 100 as it reduces the memory for post-transformation learning and training time, respectively. Furthermore, studies have shown that using small steps results in inferential energy efficiency. In addition, the performance loss is also reduced due to the small time-steps, accordingly.

C. Spike learning based Encryption

The idea of spike based learning encryption is built upon the assumption that it's difficult to elucidate class information from temporal data representation. As we have static images from our synthetic dataset, a mapping to temporal representation needs to be carried out. In this regard, we use rate coding, which generate spikes within a specified time window. It should be noted that the number of generated spikes are correlated with pixel intensity. Let's suppose that the maximum and minimum pixel intensity values are represented as $\imath_{max}$ and $\imath_{min}$. Random numbers with normal distribution within the range $[\imath_{min}, \imath_{max}]$ are generated for each pixel $[u, v]$ and time-step t. Each pixel intensity is then compared with the generated number, if the pixel intensity is lower than the generated random number, the rate coding will generate a spike. If the aforementioned condition is not met, the spike will not be generated. This process spans the pixels in spatial domain to temporal axis, accordingly.

The spikes from the temporal axis along with gradient optimization are then used to train the transformed SNNs. Presynaptic spikes are accumulated via LIF neurons to generate output spikes via the computation shown in equation 1. The information incorporated in spikes is propagated to all layers and accrued at the prediction layer. In this way, the spikes from the temporal axis when passed through softmax function can be represented in the form of probability distribution. In order to train the SNN with spike learning, we use cross entropy loss. However, studies have shown that training SNNs need more than one feed-forward steps for a single input modality, which increases the training time. Common practices for reducing training time concerning SNN training is to reduce the volume of synthetic data. However, reducing the number of samples might result in overfitting. In this regard, a study [30] proposed the use of knowledge distillation that helps in improving the ability of model to generalize. In this regard, the loss function for training SNNs incorporate both the distillation $Loss_{dist}$ and cross entropy loss $Loss_{ent}$ as shown in equation 10.

$$\begin{aligned} Loss_{SNN} =& \delta \left(Res_{2D}(X, \beta), Snn_{2D}(X, \beta) \right) \cdot Loss_{dist} \\ & + Loss_{ent} \cdot (1 - \delta) \end{aligned} \tag{10}$$

In above equation, δ refers to the balancing coefficient, distillation temperature is represented as β, and X is the subset of converted synthetic dataset. The knowledge distillation loss in equation 11 is associated with better generalization and conversion of synthetically generated data, accordingly. The study uses spatio-temporal back-propagation to compute each layer's gradients and accumulate them throughout the t [25].

V. Results and Analysis

In this section, we present the details regarding dataset, network parameters, experimental setup, and analytical results.

A. Dataset

For the validation of the GASCNN method, we have considered open source magnetic resonance imaging (MRI) brain tumor datasets. We have combined three publicly available MRI image datasets. The first is from Kaggle[1] having 155 and 98 images for brain tumor and

[1] https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection

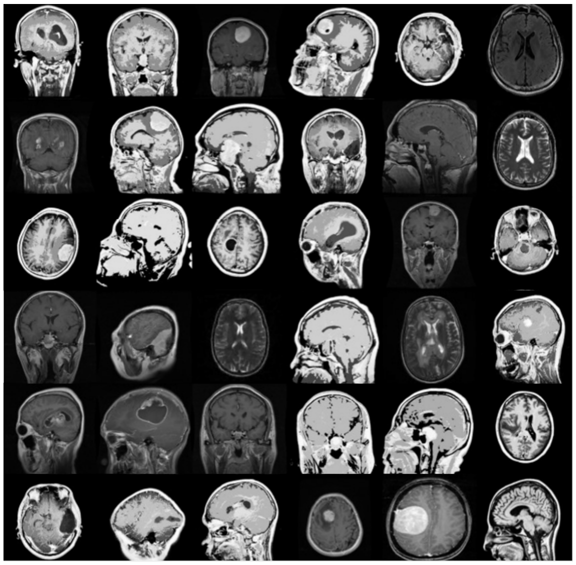

Fig. 2. Example images from MRI dataset

normal categories. The second dataset is from GitHub[2] that comprise 3264 images with four categories, i.e., pituitary, glioma, meningioma tumor, and no tumor, respectively. Lastly, the third dataset is a publicly available[3] having 3064 images with three categories, i.e., pituitary, glioma, and meningioma tumor from 233 patients. We added 98 images with no tumor category with the second and third dataset, which resulted in a total of 6426 images, respectively. Furthermore, the images in third dataset were of 16-bits. In this regard, we performed conversion to 8-bit followed by a histogram equalization process to make it compatible with other images in the dataset. Some images from the combined dataset are shown in figure 2.

B. Network Parameters

The proposed method consists of two sections, i.e., GAN and SNN. Both of the phases were developed using Python and PyTorch framework. For the GAN part, specifically, the hyperparameters used in equation 7 are assigned the values as follows: $\alpha_{TVR} = 10^{-5}$, $\alpha_{att} = 0.1$, $\alpha_{AAID} = 7$, and $\alpha_{reg} = 3$, respectively. The values of α_{att} and α_{TVR} are chosen based on the study [29], while the values of α_{AAID} and α_{reg} are chosen on empirical basis. The learning rate was set to $1e^{-4}$ along with constant decay. The model was trained with ADAM optimizer for 150k iterations.

For ResNet conversion to SNN, threshold scaling [31] was applied. For spike learning, the training was performed with a learning rate of $1e^{-3}$ with a scheduling rate decay factor of 5 at 20% of number of epochs. The network was trained on 5000 synthetically generated images. The hyperparameters in equation 10, i.e., δ and β were set to 0.6 and 15, respectively. Lastly, the values of $\mathcal{B}$ and η in Table 1 were set to 3 and 5. The network post transformation was trained for 30 epochs.

C. Performance Comparison

In this subsection, we show the experimental analysis to reveal the effectiveness of the GASCNN as the network encrypts the images without yielding notable recognition loss. The comparison is performed with SKK scheme [18], learnable image encryption scheme (LIS) [8], and Chaotic-map based image encryption (CMI) [5], and the proposed method. The comparison is performed using F1-score. The results are reported in Table 2. We used ResNet18 network architecture for reporting the accuracy on plain images. The results reveal that the proposed method, even though uses synthetic data, can encrypt the images without significant drop in performance. We also provide a qualitative comparison on the encrypted images with SKK, LIS, and CMI methods, accordingly. A qualitative comparison is shown in Figure 3. For the visualization, we encrypt the spike map with SKK scheme, respectively. It can be visualized that the SKK encrypted image results cannot hide important features that represent original class. Meanwhile, LIS encryption also provides an emergent pattern that can be leveraged for data reconstruction. On the contrary, the GASCNN does not provide any patterns or information that could be helpful in reconstruction of data. Considering the qualitative and quantitative comparison, it is safe to assume that the propose work shows better resiliency to attacks in comparison to existing works and exhibits a better coping mechanism to class leakage attacks.

TABLE II
COMPARATIVE ANALYSIS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART WORKS USING F1-SCORE

Method	Data Required	F1-score
ResNet	Training	97.52
SKK [18]	Training	93.36
LIS [8]	Training	89.30
CMI [5]	Training	74.24
GASCNN	Synthetic	94.88

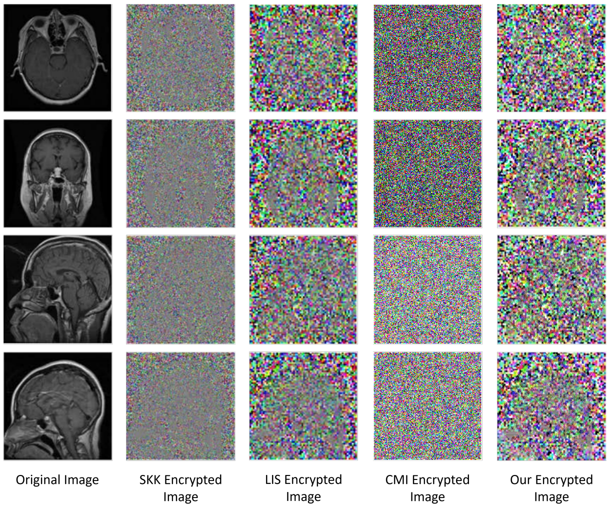

Fig. 3. Qualitative comparison of the images encrypted using state-of-the-art methods

D. Performance against security attacks

We validate the effectiveness of the proposed method against five security attacks, i.e., minimum difference attack [32], leading bit attack [32], model inversion attack [6], reverting SNN to ResNet attack, and generating class representation attack. The initial three attacks are used to reconstruct the original images from the encrypted ones. The study [8] showed that the values of color component are similar between edges for certain areas. The study also suggested that the recovery of encrypted image is mainly dependent on the negative-positive transformation, as the combined magnitude of each pixel's color component does not incur changes while shuffling them.

Similarly, the study [32] also revealed that the in order to recover the

[2] https://github.com/sartajbhuvaji/brain-tumor-classification-dataset

[3] Cheng, Jun (2017): brain tumor dataset. figshare. Dataset. https://doi.org/10.6084/m9.figshare.1512427.v5

original image, the change should be minimized for color component values. The method operates at the principle $\sum_i |p_i - q_i|$. In the above expression q is the pixel that has yet to be decrypted, while p is a decrypted nearby pixel. In order to decrypt pixel q, 48 permutations of negative-positive transformation and color shuffling needs to be carried out. The numbers of each option are computed and the one which minimizes the steps towards decrypting pixel q is selected.

The model inversion attacks intend to recreate the original image with differentiable model that minimize the distance between the differential model and trained model weights, respectively. The study [6] proposed deep leakage from gradients that showed if an attacker gets their hands on the trained model, the weights can be leverage to recreate the data.

The latter two attacks are concerned with class leakage attacks. If the attacker gets their hands on the SNN, it might leverage the model weights to recover the ResNet. The algorithm in Table 1 can then be used to optimize the input. We believe that without spike learning strategy, the attacker can use back-propagation to reconstruct the class representation, however, the GASCNN uses spike learning which presumably would help in coping with such attack as the weights in spatial domain are encrypted.

The last attack assumes that the attacker wants to reconstruct the class representation by directly back-propagating the SNN using the algorithm in Table 1. The attacker might face problems in this case, as the LIF activation function is non-differentiable [24], [25]. In this regard, approximate gradient functions [25] are used. We assume that the approximate gradients would have a strong deviation from the real gradients due to the fact that the gradients needs to be accumulated at first layer for converting temporal axis back to spatial domain.

To validate the robustness against the aforementioned attacks, i.e., minimum difference attack (MDA), leading bit attack (LBA), model inversion attack (MIA), conversion-based class representation attack (CBCRA), and direct class representation attack (DCRA), we perform the experiment using Frechet inception distance (FID) metric [33] on the plain images and encrypted images using SKK scheme, CMI, and the proposed method, respectively. The FID scores are commonly computed for the studies where GANs are used and the outputs needs to be evaluated based on the similarity. The FID score undertakes embedded features to compare the statistics. The lower FID score suggests that the reconstructed image has original image-like features, while the larger FID score suggests that the reconstructed image has large deviations in comparison to original image feature space. The results are reported in Table 3.

For the data leakage attacks, the model inversion attack cannot be applied to CMI as it's not machine learnable encryption. The results reveal that the proposed method outperforms SKK and CMI encryption techniques on all data leakage attacks. Furthermore, it was observed that the model inversion and leading bit attacks can recreate the original image from SKK scheme quite well. On the contrary, minimum difference attack does not perform well on any of the attacks and also introduces false colors during the reconstruction process. Similar is the case with class leakage attacks, our proposed method outperforms the SKK encryption schemes. It should be noted that we directly used algorithm in Table 1 on SKK scheme. It was observed that the SKK is quite vulnerable to CBCRA as it reveals important features of the actual class label. The DCRA is only applied to the proposed method as it is a direct attack on SNN. Our findings concluded that the intrinsic nature of SNN (transformation paradigm from ResNet to SNN) is effective against the DCRA, and is more robust against this attack than the encryption version. The FID for DCRA on transformed SNN was noted to be 417.7, respectively.

In addition, we also performed analysis concerning energy efficiency using the estimation model proposed in [40], which considers multiply and accumulate operations. The energy efficiency experiment undertakes the inference stage only and compares it with standard ResNet. Average number of spikes, i.e. Spike rate, across the layers was used to calculate the energy. Our analysis showed that the proposed work achieves a bit lower accuracy in comparison to standard ResNet but with more than 52x energy gains suggesting that the proposed approach lowers the spike rate, thus by extension, is more energy efficient than the standard ResNet architectures.

TABLE III

COMPARATIVE ANALYSIS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART WORKS ON DATA AND CLASS LEAKAGE ATTACKS USING FID

Method	Plain Images	SKK	CMI	Ours
MDA	47.9	87.3	106.9	115.4
LBA	36.4	54.3	97.6	103.7
MIA	23.2	36.4	-	94.8
CBCRA	128.6	318.8	-	419.1
DCRA	134.8	-	-	407.3

VI. CONCLUSION

Metaverse is a concept that is gaining a lot of interest in the research community. However, its reachability and digital footprint makes it an easy target for data and class leakage attacks. This paper proposes generative adversarial networks and spike learning based convolutional neural network (GASCNN) to cope with the data and class leakage issues for medical images. We proposed a GAN-ResNet based medical image generation method, and spike learning based encryption techniques for model weights. We have carried out extensive analysis on publicly available MRI dataset to show the efficacy of the proposed approach in terms of recognition as well as performance against security attacks. The results reveal that the proposed work is resilient to both data and class leakage attacks that include minimum difference attack, leading bit attack, model inversion attack, conversion based class representation attack, and direct class representation attacks respectively. Our experiments also reveal that the proposed work can achieved around 52x energy efficiency gains in comparison to the standard ResNet, which is compliant with existing spiking neural network architecture studies. During our experiments, we found a limitation that the GAN must be trained again if the data concerning a specific application is changed. Even though a pre-trained ResNet on X-ray images is used, the GANs cannot generate ResNet converted SNN unless the GAN is first trained on few samples of original X-ray images first. We intend to combine the universal source free domain adaptation with spike learning techniques to make the network able of generating different application oriented medical images while securing them with spike learning based technique. Furthermore, we would like to dwell upon the possible usage of the proposed method against model poisoning attack in the domain of medical imaging, respectively.

REFERENCES

[1] M. Zawish, F.A. Dharejo, S.A. Khowaja, K. Dev, S. Dav, N.M.F. Qureshi, and P. Bellavista. AI and 6G into the Metaverse: Fundamentals, Challenges and Future Research Trends. arXiv preprint arXiv:2208.10921, 2022.

[2] M. U. Rehman et al., "A Novel Chaos-Based Privacy-Preserving Deep Learning Model for Cancer Diagnosis," in IEEE Transactions on Network Science and Engineering, vol. 9, no. 6, pp. 4322-4337, 1 Nov.-Dec. 2022

[3] T. A. Al-Maadeed, I. Hussain, A. Anees, and M. T. Mustafa, "A image encryption algorithm based on chaotic lorenz system and novel primitive polynomial s-boxes," Multimedia Tools Appl., vol. 80, pp. 24801–24822, 2021.

[4] F. A. Dharejo, M. Zawish, F. Deeba, Y. Zhou, K. Dev, S.A. Khowaja, N.M.F. Qureshi. Multimodal-Boost: Multimodal Medical Image Super-Resolution Using Multi-Attention Network with Wavelet Transform. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Early access article, pp. 1 - 14, 2022.

[5] S. A. Khowaja, K. Dev, N. M. F. Qureshi, P. Khuwaja and L. Foschini, ”Toward Industrial Private AI: A Two-Tier Framework for Data and Model Security,” in IEEE Wireless Communications, vol. 29, no. 2, pp. 76-83, April 2022

[6] L. Zhu and S. Han. Deep Leakage from Gradients. Advances in Neural Information Processing Systems, pp.14774-14784, 2020.

[7] S. A. Khowaja, P. Khuwaja, K. Dev, and A. Antonopoulos, “SPIN: Simulated Poisoning and Inversion Network for Federated Learning-Based 6G Vehicular Networks,” IEEE International Conference on Communications (ICC), pp. 1 - 6, May. 2023

[8] Q. -X. Huang, W. L. Yap, M. -Y. Chiu and H. -M. Sun, ”Privacy-Preserving Deep Learning With Learnable Image Encryption on Medical Images,” in IEEE Access, vol. 10, pp. 66345-66355, 2022.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” Theory of Cryptography, pp. 265–284, 2006

[10] R. Audhkhasi and M. L. Povinelli, “Generalized multi-channel scheme for secure image encryption,” Sci. Rep., vol. 11, no. 1, pp. 1–9, 2021.

[11] Y. Zhang et al., “DNA origami cryptography for secure communication,” Nature Commun., vol. 10, no. 1, pp. 1–8, 2019. .

[12] G. Kaur, R. Agarwal, and V. Patidar, “Chaos based multiple order optical transform for 2D image encryption,” Eng. Sci. Technol. Int. J., vol. 23, no. 5, pp. 998–1014, 2020.

[13] M. U. Rehman, A. Shafique, S. Khalid, and I. Hussain, “Dynamic substitution and confusion-diffusion-based noise-resistive image encryption using multiple chaotic maps,” IEEE Access, vol. 9, pp. 52277–52291, 2021.

[14] D. Hyun, L. Abou-Elkacem, R. Bam, L. L. Brickson, C. D. Herickhoff, and J. J. Dahl, “Nondestructive detection of targeted microbubbles using dualmode data and deep learning for real-time ultrasound molecular imaging,” IEEE Trans. Med. Imag., vol. 39, no. 10, pp. 3079–3088, Oct. 2020

[15] G. Carneiro, J. Nascimento, and A. P. Bradley, “Automated analysis of unregistered multi-view mammograms with deep learning,” IEEE Trans. Med. Imag., vol. 36, no. 11, pp. 2355–2365, Nov. 2017.

[16] Y. Ding, G. Wu, D. Chen, N. Zhang, L. Gong, M. Cao, and Z. Qin, ”DeepEDN: A deep-learning-based image encryption and decryption network for internet of medical things,” IEEE Internet Things J., vol. 8, no. 3, pp. 1504–1518, Feb. 2021

[17] M. Tanaka, ”Learnable image encryption,” in Proc. IEEE Int. Conf. Consum. Electron. Taiwan (ICCE-TW), May 2018, pp. 1–2.

[18] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, ”Pixel-based image encryption without key management for privacy-preserving deep neural networks,” IEEE Access, vol. 7, pp. 177844–177855, 2019.

[19] W. Sirichotedumrong and H. Kiya, ”A GAN-based image transformation scheme for privacy-preserving deep neural networks,” in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 745–749.

[20] T. Xiao, Y.-H. Tsai, K. Sohn, M. Chandraker, and M.-H. Yang, “Adversarial Learning of Privacy-Preserving and Task-Oriented Representations,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12434–12441, Apr. 2020

[21] S. A. Khowaja, I. H. Lee, K. Dev, M. A. Jarwar and N. M. F. Qureshi, ”Get Your Foes Fooled: Proximal Gradient Split Learning for Defense Against Model Inversion Attacks on IoMT Data,” in IEEE Transactions on Network Science and Engineering, 2022

[22] M. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, “Privacy-Preserving Human Activity Recognition from Extreme Low Resolution,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, Feb. 2017.

[23] B. Zhao, K. R. Mopuri, and H. Bilen, “iDLG: Improved Deep Leakage from Gradients,” arXiv:2001.02610 [cs, stat], Jan. 2020.

[24] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, “Going Deeper in Spiking Neural Networks: VGG and Residual Architectures,” Frontiers in Neuroscience, vol. 13, Mar. 2019

[25] Y. Kim, Y. Venkatesha, and P. Panda, “PrivateSNN: Privacy-Preserving Spiking Neural Networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 1192–1200, Jun. 2022

[26] G.K. Nayak, K.R. Mopuri, V. Shaj, V.B. Radhakrishnan, and A. Chakraborty. Zero-shot knowledge distillation in deep networks. In International Conference on Machine Learning (pp. 4743-4751). PMLR, May, 2019.

[27] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In International conference on machine learning (pp. 2642-2651). PMLR, July, 2017.

[28] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR, July, 2017.

[29] E. Jung, M. Luna, and S. H. Park, “Conditional GAN with 3D Discriminator for MRI Generation of Alzheimer’s Disease Progression,” Pattern Recognition, p. 109061, Sep. 2022

[30] L. Yuan, F. E. Tay, G. Li, T. Wang and J. Feng, ”Revisiting Knowledge Distillation via Label Smoothing Regularization,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3902-3910, 2020

[31] B. Han, and K. Roy. Deep spiking neural network: Energy efficiency through time based coding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020.

[32] A. H. Chang and B. M. Case, ”Attacks on image encryption schemes for privacy-preserving deep neural networks,” 2020, arXiv:2004.13263.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. ”Gans trained by a two time-scale update rule converge to a local nash equilibrium.” Advances in neural information processing systems, 30, 2017.

[34] Y. Zhang, H. Xu, L. Huang, and C. Chen. ”A storage-efficient SNN–CNN hybrid network with RRAM-implemented weights for traffic signs recognition.” Engineering Applications of Artificial Intelligence, vol. 123, pp. 106232, 2023.

[35] Z. Tang, J. Wang, B. Yuan, H. Li, J. Zhang and H. Wang. ”Markov-GAN: Markov image enhancement method for malicious encrypted traffic classification.” IET Information Security, vol. 16, pp. 442-458, 2022.

[36] X, Liao, Y. Wu, Z. Wang, D. Wang, and H. Zhang. ”A convolutional spiking neural network with adaptive coding for motor imagery classification.” Neurocomputing, vol. 549, pp. 126470, 2023.

[37] Z. Wang, C. Peng, X. He, and W. Tan. ”Wasserstein Distance-Based Deep Leakage from Gradients.” Entropy, vol. 25, no. 5, 810, 2023.

[38] G. Lu, Z. Xiong, R. Li, N. Mohammad, Y. Li, and W. Li. ”DEFEAT: A decentralized Federated Learning against gradient attacks.” High-Confidence Computing, In Press, pp. 100128, 2023.

[39] Y. Yang, Z. Ma, B. Xiao, Y. Liu, T. Li, and J. Zhang. ”Reveal Your Images: Gradient Leakage Attack against Unbiased Sampling-Based Secure Aggregation.” IEEE Transactions on Knowledge and Data Engineering, Early Access Article, pp. 1-14, 2023.

[40] P. Panda, S. A. Aketi, and K. Roy. ”Toward Scalable, Efficient, and Accurate Deep Spiking Neural Networks With Backward Residual Connections, Stochastic Softmax, and Hybridization.” Frontiers in Neuroscience, vol. 14, pp. 653, 2020.