On the Privacy-Preserving Properties of Spiking Neural Networks with Unique Surrogate Gradients and Quantization Levels

Ayana Moshruba, Shay Snyder, Hamed Poursiami, Maryam Parsa
Electrical and Computer Engineering
George Mason University
Fairfax, USA
{amoshrub, ssnyde9, hpoursia, mparsa}@gmu.edu

Abstract—As machine learning models increasingly process sensitive data, understanding their vulnerability to privacy attacks is vital. Membership Inference Attack (MIA), which infer whether specific data points were used during training, is one such privacy risk. Previous work suggests that Spiking Neural Networks (SNNs), which rely on event-driven computation and discrete spike-based encoding, exhibit greater resilience to MIAs compared to Artificial Neural Networks (ANNs). This resilience is attributed to their non-differentiable activations and inherent stochasticity, which reduce the correlation between model responses and individual training samples. To further enhance privacy in SNNs, we explore two techniques: Quantization and Surrogate Gradients. Quantization, which reduces model precision to limit information leakage, has been shown to improve privacy resilience in ANNs. Since SNNs exhibit sparse and irregular activations, quantization may have an even stronger effect on disrupting the activation patterns exploited by MIAs. In this study, we compare the vulnerability of SNNs and ANNs to MIAs under *weight* and *activation* quantization across multiple datasets. We evaluate privacy vulnerability using the attack model's Receiver Operating Characteristic (ROC) curve's Area Under the Curve (AUC) metric, where lower values indicate stronger privacy protection, and assess model accuracy to quantify the privacy-accuracy trade-off. Our results show that quantization enhances privacy in both architectures with minimal performance degradation, but full-precision SNNs remain more resilient than even quantized ANNs. Additionally, we examine the impact of surrogate gradients on privacy in SNNs. Among the five surrogate gradients evaluated, Spike Rate Escape provides the best privacy-accuracy trade-off, while Arctangent (aTan) increases vulnerability to MIAs. These findings reinforce SNNs' inherent privacy advantages and demonstrate that both quantization and surrogate gradient selection can further influence privacy-accuracy trade-offs in SNNs.

Index Terms—quantization, surrogate gradients, spiking neural networks, membership inference attack, adversarial machine learning

I. INTRODUCTION

The widespread adoption of Machine Learning (ML) across domains such as healthcare [1], finance [2], and education [3] has raised concerns about privacy risks, particularly from attacks exploiting model behaviors to infer sensitive information [4], [5]. Artificial Neural Networks (ANNs) are widely used across various domains but remain vulnerable to privacy attacks, like Membership Inference Attacks (MIAs), where adversaries attempt to infer whether specific data points were used during training [6]. Prior study suggests that Spiking Neural Networks (SNNs) show lower vulnerability to MIAs compared to ANNs [7], potentially due to their unique spike based computation, which introduces inherent privacy advantages [8]. Specifically, the non-differentiable and discontinuous nature of SNNs reduces the correlation between model responses and individual data points [9], while their stochastic spike-based encoding increases representation diversity, reducing the risk of overfitting and making individual training samples less distinguishable [10].

Given these properties, we explore methods to further enhance the privacy of SNNs. First, we investigate quantization, which has been proven to improve privacy in ANNs by reducing precision and altering learned representations, potentially limiting information leakage [11]. Since SNNs operate in an event-driven manner, where neurons fire only when necessary [12], the combination of spiking sparsity and quantization may further disrupt activation patterns exploited by MIAs. Additionally, we hypothesize that the information loss and noise introduced by quantization [13] could influence the model's susceptibility to privacy attacks.

Second, we examine the role of surrogate gradients in shaping SNN privacy. These functions approximate gradients for non-differentiable spike events, enabling backpropagation-based training. Beyond facilitating learning, surrogate gradients introduce inherent variability into model outputs through spike timing distributions and gradient approximations. This stochasticity resembles the noise injection mechanisms of Differentially Private Stochastic Gradient Descent (DPSGD) [14], suggesting a potential privacy-preserving effect. To explore this, we analyze five different surrogate gradient functions: Fast Sigmoid, Arctangent (aTan), Spike Rate Escape(STE), Triangular, and Straight Through Estimator [15] and evaluate their influence on privacy vulnerability and model accuracy.

In this study, we assess the effects of both *weight* and *activation* quantization on the privacy characteristics of SNNs and ANNs against MIAs, evaluating models across convolutional and fully connected architectures. Our experiments span five datasets: CIFAR-10 [16], MNIST [17], FMNIST [18], Iris [19], and Breast Cancer [20]. Privacy vulnerability is measured using the attack model's Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), where lower values indicate stronger privacy protection. We also assess model accuracy to

quantify the trade-off between privacy and performance under different quantization and surrogate gradient settings.

Based on our experimental analysis, we summarize the key findings regarding the impact of quantization and surrogate gradients on privacy vulnerability and model performance below.

- **Quantization Analysis:** Both weight and activation quantization reduces MIA vulnerability compared to full precision in ANNs and SNNs, while introducing minor degradation in model performance. However, extreme quantization to 2 bits imposes unnecessary performance degradation without substantial privacy benefits, making moderate quantization (4-bit and 8-bit) a more practical choice.
- **Privacy Protection:** While quantization enhances privacy protection in both architectures, full precision SNNs demonstrate inherently superior privacy compared to quantized ANNs across all datasets, highlighting their fundamental privacy advantage without compromising model performance.
- **Surrogate Gradient Evaluation:** Spike rate escape demonstrates superior privacy protection while maintaining high model accuracy, whereas aTan and STE exhibit higher MIA vulnerability.

II. RELATED WORK

Privacy attacks targeting sensitive data have raised concerns about information exposure, driving research into vulnerabilities like MIAs. Li et al. [4] provide an overview of privacy attacks, including MIAs, which were first introduced in genomic studies by Homer et al. [21]. Shokri et al. [22] later formalized MIAs by developing the shadow model framework, where labeled datasets are generated to train an attack model for inferring data membership. Refinements by Salem et al. [23] showed that MIAs could rely on single shadow models and confidence scores, while Nasr et al. [24] demonstrated that white-box access enhances attack precision. While most privacy research has focused on ANNs, studying MIAs in SNNs is valuable due to their potential resilience, which comes from their sparse activations, where neurons fire only when necessary and the randomness in spike timing, both of which make it harder for an attacker to infer data membership. Notably, prior work highlights that SNNs incur less performance degradation when employing DPSGD compared to ANNs [7]. Han et al. [25] and Safronov et al. [26] introduced privacy-preserving techniques like federated learning and differential privacy for SNNs.

While prior research has explored privacy vulnerabilities in SNNs and introduced privacy-preserving techniques, further investigation is needed to understand how architectural modifications can enhance their resilience against MIAs. One such modification is quantization which has become an effective approach for reducing model size and computational demands, particularly in resource-constrained settings [27]. By reducing weight and activation precision, quantization introduces noise, which protect privacy by making adversarial inference more challenging [13]. Studies on ANNs using DoReFa-Net [28], [11], [29] apply both weight and activation quantization,

demonstrating reduced MIA success rates due to added noise and lower overfitting. Applying quantization to SNNs has proven valuable for improving energy efficiency. Yin et al. [30] demonstrated that weight and membrane potential quantization in SNNs reduces memory use by 93.8% and computation energy by 90% with minimal accuracy loss. Schaefer et al. [31], [32] showed that ternary quantization optimizes hardware efficiency by reducing energy and memory costs. Frameworks like Q-SpiNN [33] and QFFS [34] adapt quantization techniques to SNN-specific dynamics, addressing challenges like synaptic weight and membrane potential quantization for low-power neuromorphic systems. Recent findings suggest that quantization noise in SNNs may reduce information leakage in MIAs, aligning with observations in ANNs [13], [11].

On the other hand, surrogate gradients enable effective training in SNNs by approximating gradients for non-differentiable spike events, but their impact on privacy has received little attention. PrivateSNN, introduced by Kim et al., proposes a privacy-preserving approach for converting ANNs to SNNs by incorporating spike-based learning rules to mitigate privacy risks [26]. Similarly, DPSNN, developed by Wang et al., integrates differential privacy with SNNs by leveraging gradient noise and discrete spike sequences, aiming to enhance model robustness against privacy attacks [35]. However, the direct impact of different surrogate gradient functions on the privacy characteristics of SNNs remains unexplored, leaving a gap in understanding their potential role in mitigating privacy risks. Similarly, while prior work has examined quantization for ANN privacy, its implications for SNN privacy have not been thoroughly investigated. Our work bridges these gaps by evaluating how quantization and surrogate gradients impact privacy in SNNs, providing insights into their role in enhancing privacy resilience while maintaining model performance.

III. BACKGROUND

A. Spiking Neural Networks (SNNs)

SNNs are inspired by biological neural activity and represent a fundamental shift from traditional ANNs' continuous outputs. SNNs operate through discrete spikes, occurring only when a neuron's membrane potential surpasses a specific threshold. This mechanism incorporates time as an additional dimension in information processing, where spike timing patterns encode neural representations [12]. This spike-based mode of communication supports asynchronous data processing and aligns well with neuromorphic hardware designed for event-driven computation, offering enhanced energy efficiency and reduced latency [12]. SNNs require alternative training mechanisms to overcome the non-differentiability of their spike function [12].

Surrogate gradients provide a workaround for the non-differentiable spike function by substituting a smooth surrogate function in the backward pass. The function approximates the spiking neuron's activation and is used to compute gradients during backpropagation. Common surrogate gradients include:

- **Fast Sigmoid:** Approximates the gradient using a sharp sigmoid function. It provides precise gradient updates, enhancing feature representation and learning stability.

- **Arctangent:** Employs the derivative of the arctangent function as the gradient, offering smoother updates. However, its less aggressive slope makes it less effective in disrupting predictable patterns.
- **Spike Rate Escape:** A gradient model based on a sigmoid-like function with a decay parameter, effectively introducing noise in spike activations. This makes it useful for improving privacy resilience.
- **Triangular:** Uses a linear approximation for gradients. Its weaker gradient strength often leads to instability in learning.
- **Straight Through Estimator (STE):** Utilizes the gradient of a fast sigmoid function during the backward pass, while maintaining a unit derivative for simplicity. This method balances computational efficiency with gradient approximations.

Surrogate functions enable gradient flow in SNNs despite the non-differentiability of spike events. They introduce variability through spike timing distributions and gradient approximations, a mechanism aligning with DPSGD, where privacy relies on systematic noise injection during training. So unlike DPSGD's explicit noise addition, SNNs inherently generate randomness, which may provide privacy benefits without incurring the performance penalties of DPSGD..

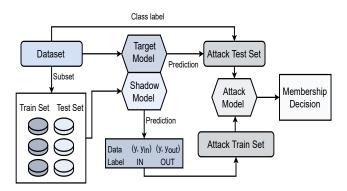### B. Membership Inference Attack (MIA)

Fig. 1: Membership Inference Attack (MIA) Framework

MIA is a privacy attack that enables adversaries to infer the presence of individual data samples in a ML model's training dataset [22]. This attack exploits differences in the model's behavior on training and non-training data. [36]. Models tend to exhibit higher confidence or different error patterns for samples they have encountered during training, compared to new unseen data [37]. By analyzing these patterns, attackers can infer sensitive information about training data, which can lead to privacy breaches [38].

MIA involves training attack models or employing statistical tests to distinguish between the model responses on training and non-training data. The success of these attacks largely depends on the model's overfitting to its training data and the distinctiveness of the model's responses to individual samples. MIAs not only breach data privacy but also expose weaknesses in a model's ability to generalize [39].

The MIA framework in our experiments involves a two-model approach (Figure 1) consisting of a target model and a shadow model, each of the same architecture described in Section IV-A. The process begins with a target model trained on the dataset of interest, which the adversary aims to analyze. To approximate the target model's behavior, a shadow model is trained on an 80% subset of the same dataset, mimicking the target's architecture and learning parameters. The attack training dataset is generated using the shadow model, where predictions on its training data are labeled as 'IN' and those on its test data as 'OUT.' The attack test dataset is then constructed from the target model's predictions, following the same labeling scheme. An SVM with a Radial Basis Function (RBF) kernel serves as the attack model, trained on the constructed attack dataset. For ANNs, logits from the fully connected layer are used as input features, while for SNNs, membrane potentials from the final time step are utilized. To mitigate class imbalance, undersampling is applied during training. The trained attack model is then tested on the target model's predictions, determining whether a given sample was part of the training set ('IN') or not ('OUT'), quantifying its vulnerability to MIAs.

C. Quantization

Quantization reduces the precision of model parameters (weights) and layer activations, typically from 32-bit floating point to lower bit integers [40]. This technique lowers memory usage, accelerates inference, and reduces energy consumption, facilitating the deployment of Deep Neural Networks (DNNs) on resource-constrained devices [41]. Additionally, the noise inherently introduced by quantization acts as an implicit regularization mechanism, mitigating overfitting and improving generalization [13].

In this work, we investigate weight quantization and activation quantization and their impact on privacy in ANNs and SNNs.

Activation Quantization: This method reduces the precision of neuron outputs, constraining activations to discrete levels:

$$a_q = \text{round}\left(\frac{\text{clip}(a, a_{\min}, a_{\max}) - a_{\min}}{\Delta}\right) \cdot \Delta + a_{\min}. \quad (1)$$

Here, $\Delta = \frac{a_{\max} - a_{\min}}{2^k - 1}$ ensures uniform quantization within the activation range $[a_{\min}, a_{\max}]$, where k represents the bit width of quantization.

Weight Quantization: Weight quantization maps weights w to a discrete range, optimizing storage and computation:

$$Q(w) = \Delta \cdot \text{clip}\left(\text{round}\left(\frac{w}{\Delta}\right), 0, 2^k - 1\right), \quad (2)$$

where $\Delta = \frac{w_{\max} - w_{\min}}{2^k - 1}$ defines the step size for a bit-width k, ensuring both positive and negative weights are accounted for.

Quantized models require gradient approximation during training. In ANNs, Quantization-Aware Training (QAT) updates quantized parameters while maintaining gradient flow. In SNNs, non-differentiable spike functions necessitate surrogate gradients approximations. This study applies QAT with *fast sigmoid* to examine their impact on privacy vulnerability while maintaining accuracy.
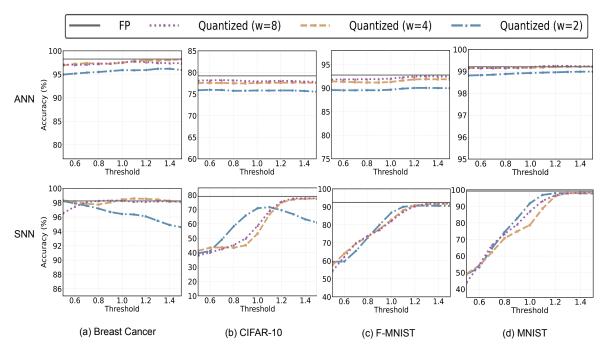
Fig. 2: Activation Quantization impact on Model Accuracy on (a) Breast Cancer (b) CIFAR-10, (c) F-MNIST, and (d) MNIST. The grey solid line represents the Full Precision (FP) model, while the purple dotted, orange dashed, and blue dash-dotted lines correspond to the quantized models with bit precisions of w=8, w=4, and w=2 respectively.

TABLE I: Model Architecture and Configuration

Dataset	Network	Model	Structure
CIFAR10 MNIST FMNIST	ConvNet	ANN	2 Conv layers (32, 64 filters) 2 MaxPool layers 2 FC layers (1000, num_classes) ReLU activations
		SNN	2 Conv layers (32, 64 filters) 2 MaxPool layers LIF neurons ($\beta = 0.95$) Temporal processing ($T = 25$)
Iris Breast-Cancer	FCNet	ANN	2 FC layers (Input, 1000, num_classes) ReLU activations
		SNN	2 FC layers (Input, 1000, num_classes) LIF neurons ($\beta = 0.95$) Temporal processing ($T = 25$)

IV. EXPERIMENTAL SETUP

A. Model Architectures and Datasets

This study evaluates CIFAR10 [16], MNIST [17], FM-NIST [18], Iris [19], and Breast Cancer [20], implemented using PyTorch [42] and SNNtorch [15]. For CIFAR10, MNIST, and FMNIST we follow their standard splits (50,000 training, 10,000 testing), while for Iris and Breast Cancer we apply an 80-20 stratified split.

We preprocess CIFAR10 by normalizing the data to a mean of [0.5, 0.5, 0.5], applying random cropping with a 4-pixel padding and 50% horizontal flipping. For MNIST and FMNIST, we resize the images to 28×28 pixels, convert them to grayscale, and normalize them with a mean of 0 and standard deviation of

1, without augmentations. For Iris and Breast Cancer datasets, we standardize the features using `StandardScaler` [43].

We use Convolutional Neural Networks (CNNs) (ConvNet) for CIFAR10, MNIST, and FMNIST , and Fully Connected Networks (FCNets) for Iris and Breast Cancer. We summarize the detailed configurations for both ANNs and SNNs in Table I. For activation functions, we use ReLU for ANNs and Leaky Integrate and Fire (LIF) neurons ($\beta = 0.95$) for SNNs, which operate over a time step of 25 to propagate spikes.

B. Quantization Methods

We implement activation quantization differently for ANNs and SNNs. In ANNs, we quantize activations using `brevitas.nn QuantReLU` [44], while in SNNs, we develop a custom `state_quant` function within snnTorch [15] to quantize membrane potentials. To assess the effect of activation precision, we apply thresholds ranging from 0.5 to 1.5 before quantization, clipping activations and setting an upper bound on magnitudes included in the quantized representation. Lower thresholds impose stricter clipping, reducing the range of preserved activations and potentially limiting representational capacity. In contrast, higher thresholds retain a broader range of activations, which may enhance expressiveness but could also increase vulnerability to privacy attacks.

We apply weight quantization using `brevitas.nn` for both ANNs and SNNs, reducing parameter precision to 2-bit, 4-bit, and 8-bit levels. This process discretizes model weights, minimizing storage requirements and computational cost while potentially influencing privacy vulnerability.
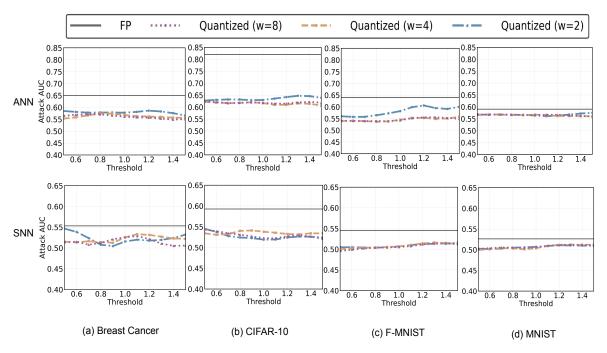
Fig. 3: Activation Quantization impact on Privacy Vulnerability on (a) Breast Cancer (b) CIFAR-10, (c) F-MNIST, and (d) MNIST. The grey solid line represents the Full Precision (FP) model, while the purple dotted, orange dashed, and blue dash-dotted lines correspond to the quantized models with bit precisions of w=8, w=4, and w=2 respectively.

C. Surrogate Gradient Configurations

We use five surrogate gradients in our experiments: *Fast Sigmoid*, *Straight Through Estimator*, *Arctangent*, *Spike Rate Escape*, and *Triangular*. Among these, we specifically employ *Fast Sigmoid* for quantization studies [15]. We set the *Fast Sigmoid* gradient slope to 25, while *Arctangent (atan)* uses an alpha value of 2 which controls the steepness of its curve. For *Spike Rate Escape*, we apply a beta parameter of 1 which regulates the escape rate with a slope of 25. We use *Triangular* and *Straight Through Estimator (STE)* with their default configurations in the framework.

V. RESULTS & DISCUSSION

This section presents the evaluation of how activation and weight quantization as well as surrogate gradients impacts SNN privacy, focusing on MIA vulnerability and performance trade-offs in comparison to ANNs. Privacy vulnerability is measured via ROC-AUC, while accuracy captures performance impact.

A. Activation Quantization

Model Performance: Accuracy degradation due to activation quantization is observed in both ANNs and SNNs, but the overall drop from full precision to quantized models with 8-bits, 4-bits and 2-bits remains relatively small (Figure 2). In ANNs, accuracy decreases more noticeably at lower bit-widths, particularly at 2-bit, where reduced precision compresses activation ranges, limiting representational capacity and reducing the distinction between logits. However, across each bit-width, accuracy remains relatively stable across different threshold

levels, indicating that thresholding has minimal impact on model performance. This is because ReLU activations, operate in a continuous space where moderate clipping does not significantly alter feature representations, allowing the network to maintain performance despite threshold variations.

In SNNs, accuracy trends vary across datasets, as shown in Figure 2. For 4-bit and 8-bit quantized models, higher thresholds consistently improve accuracy across all datasets by retaining a broader activation range, preserving useful signal information. The stochastic nature of SNNs further mitigates quantization noise, ensuring stable performance regardless of dataset complexity. However, 2-bit quantized models show different behaviors. In Breast Cancer, accuracy starts near full precision at lower thresholds but degrades as thresholds increase due to amplified noise overwhelming simpler patterns. In CIFAR-10, accuracy initially improves with thresholding but drops at higher thresholds as excessive noise disrupts feature representation, reducing classification performance.

MIA Vulnerability: Quantization reduces attack vulnerability in ANNs compared to full precision as seen in Figure 3. Quantized models with 2-bit, 4-bit, and 8-bit precision all exhibit lower vulnerability than the full-precision model. However, among them, 2-bit quantized model is the most vulnerable across all datasets, while 4-bit and 8-bit quantized model provide better privacy protection. This occurs because extreme quantization (2-bit precision) severely limits representational capacity, making activations more uniform and easier to infer in ANNs, thereby increasing susceptibility to attacks.

In SNNs, activation quantization minimizes attack vulnerability compared to full precision across all datasets as well. However, unlike in ANNs, 2-bit, 4-bit, and 8-bit quantized

Dataset	Method	ANN			SNN		
		Train Accuracy	Test Accuracy	MIA AUC	Train Accuracy	Test Accuracy	MIA AUC
MNIST	Full precision	99.96(±0.03)%	99.21(±0.01)%	0.5900(±0.008)	99.95(±0.05)%	99.22(±0.02)%	0.5264(±0.011)
	quantized(w=8)	99.88(±0.05)%	99.16(±0.02)%	0.5674(±0.007)	99.93(±0.03)%	99.16(±0.03)%	0.5101(±0.012)
	quantized(w=4)	99.85±(0.04)%	99.12(±0.04)%	0.5800(±0.003)	99.91(±0.02)%	99.21(±0.05)%	0.5117±(0.009)
	quantized(w=2)	99.57(±0.02)%	99.07(±0.03)%	0.5667(±0.009)	99.88±(0.03)%	99.11(±0.04)%	0.4993(±0.005)
FMNIST	Full precision	99.52(±0.13)%	92.77(±0.20)%	0.6400(±0.011)	99.42(±0.32)%	92.44(±0.19)%	0.5453(±0.008)
	quantized(w=8)	95.98(±0.08)%	92.30(±0.10)%	0.5397(±0.010)	96.49(±0.91)%	92.00(±0.10)%	0.5022(±0.003)
	quantized(w=4)	95.47(±0.72)%	92.15(±0.09)%	0.5406(±0.005)	96.44(±0.53)%	91.97(±0.17)%	0.4970(±0.010)
	quantized(w=2)	92.61(±0.20)%	90.69(±0.18)%	0.5295(±0.004)	93.37(±0.51)%	91.36(±0.18)%	0.5024(±0.005)
CIFAR10	Full precision	99.24(±0.08)%	79.20(±0.43)%	0.8200(±0.095)	99.13(±0.45)%	78.99(±0.33)%	0.5927(±0.005)
	quantized(w=8)	79.61(±0.20)%	77.78(±0.28)%	0.6800(±0.015)	72.37(±0.44).%	73.39(±0.13)%	0.5250(±0.022)
	quantized(w=4)	79.56(±0.41)%	77.71(±0.39)%	0.7000(±0.021)	72.54(±0.29)%	72.92(±0.51)%	0.5405(±0.011)
	quantized(w=2)	71.35(±0.24)%	73.37(±0.15)%	0.6200(±0.015)	66.41(±1.09)%	68.59(±0.55)%	0.5280(±0.002)
Iris	Full precision	98.33(±2.34)%	100.0(±0.00)%	0.7700(±0.13)	100.0(±0.00)%	100.0(±0.00)%	0.5728(±0.020)
	quantized(w=8)	82.78(±2.02)%	93.33(±2.94)%	0.6500(±0.070)	79.44(±7.33)%	95.56(±1.92)%	0.5234(±0.012)
	quantized(w=4)	85.56(±1.05)%	91.11(±1.92)%	0.7163(±0.160)	84.17(±2.47)%	94.44(±3.06)%	0.5304(±0.045)
	quantized(w=2)	84.06(±2.20)%	91.11(±3.06)%	0.6400(±0.013)	81.94(±4.18)%	92.22(±6.71)%	0.5380(±0.020)
Breast Cancer	Full precision	99.34(±0.04)%	98.25(±0.00)%	0.6500(±0.008)	100.0(±0.00)%	98.25(±0.24)%	0.5534(±0.017)
	quantized(w=8)	97.51(±0.13)%	97.96(±0.04)%	0.5500(±0.006)	99.09(±0.64)%	98.25(±0.00)%	0.4983(±0.003)
	quantized(w=4)	97.58(±0.57)%	97.96(±0.49)%	0.5404(±0.010)	97.69(±0.36)%	97.66(±0.51)%	0.5099(±0.021)
	quantized(w=2)	97.80(±0.46)%	97.54(±0.40)%	0.5800(±0.013)	97.87(±0.48)%	97.37(±0.00)%	0.5088(±0.021)

TABLE II: Weight Quantization Impact on Privacy Vulnerability and Model Performance: Comparison between ANN and SNN models across different datasets.

SNN models showed somewhat similar privacy vulnerability, specially in F-MNSIT and MNIST datasets. This suggests that the stochastic nature of spike-based encoding inherently limits the granularity of information available to attackers, irrespective of precision levels. The inherent randomness in spike timing and event-driven processing disrupts predictable activation patterns, making the additional noise introduced by quantization less impactful in distinguishing training from non-training data. This means that while quantization is still effective in lowering attack success rates, further reducing bit width does not provide additional privacy benefits for these datasets.

B. Weight Quantization

Model Performance: Weight quantization leads to accuracy degradation in both ANNs and SNNs from full precision models as shown in Table II. In ANNs, the most pronounced drop is observed in 2-bit quantized models, similar to the impact of activation quantization. Extreme quantization severely limits weight precision, resulting in coarser updates that reduce representational capacity and hinders the model's ability to learn fine-grained patterns. This effect is evident in CIFAR-10, where accuracy declines more sharply at lower bit width of 2. The higher complexity of CIFAR-10 requires finer weight precision to capture intricate features, and 2-bit quantization struggles to retain the necessary detail for accurate classification.

In SNNs, weight quantization similarly affects accuracy as shown in Table II. The most pronounced drop is observed in 2-bit quantized models, where limited weight precision reduces the model's capacity to represent detailed spike-based patterns. Like ANNs, this is especially evident in CIFAR-10 where, 2-bit quantized model struggles to encode intricate features

effectively, resulting in a sharper accuracy decline compared to 4-bit and 8-bit models.

MIA Vulnerability: Weight quantization consistently reduces MIA vulnerability in ANNs compared to full-precision models across all datasets (Table II). In most cases, models quantized to 2-bit, 4-bit, and 8-bit levels demonstrate comparable AUC values, effectively obscuring training data from attackers through noise introduced by quantization. Notably, 2-bit quantized models exhibit the lowest attack AUC values across datasets, representing a trend opposite to that observed with activation quantization. This difference arises from how different quantization noise affects model representations. For weight quantization, 2-bit precision disrupts parameter-level patterns, introducing randomness that hampers an attacker's ability to infer sensitive training data. By contrast, activation quantization with 2-bit precision compresses activation ranges excessively, resulting in uniform outputs that are easier to predict, thereby increasing vulnerability.

In SNNs, weight quantization reliably lowers MIA vulnerability across all datasets compared to full-precision models though reductions are smaller due to the already lower baseline AUC. Models quantized to 2-bit, 4-bit, and 8-bit levels demonstrate similar AUC values which indicates that weight quantization has a uniformly beneficial impact on reducing vulnerability across all levels of quantization. This trend is similar to the impact observed in activation quantization, where different bit widths also produced closely aligned AUC values. The underlying reason for this consistency lies in the stochastic nature of spike-based computation, which inherently disrupts predictable activation patterns. In SNNs, weight quantization further amplifies this effect by introducing additional noise to synaptic weights, but due to the already high variability in spike timing and membrane dynamics, the additional

Dataset	Surrogate Gradients	SNN		
		Train Accuracy	Test Accuracy	MIA AUC
MNIST	Fast Sigmoid(slope,k=25)	99.88(±0.01)%	99.22(±0.04)%	0.518(±0.001)
	aTan(alpha=2)	99.96(±0.02)%	99.25(±0.06)%	0.547(±0.006)
	Spike Rate Escape(beta=1, slope=25)	99.97(±0.01)%	99.25(±0.03)%	0.508(±0.008)
	Triangular	75.16(±0.09)%	76.34(±0.14)%	0.503(±0.003)
	Straight Through Estimator	99.57(±0.04)%	98.79(±0.07)%	0.528(±0.005)
FMNIST	Fast Sigmoid(slope,k=25)	99.45(±0.04)%	91.97(±0.12)%	0.518(±0.008)
	aTan(alpha=2)	99.58(±0.03)%	92.18(±0.14)%	0.547(±0.002)
	Spike Rate Escape(beta=1, slope=25)	99.74(±0.09)%	92.23(±0.11)%	0.523(±0.016)
	Triangular	78.79(±0.24)%	79.49(±0.20)%	0.498(±0.011)
	Straight Through Estimator	91.55(±0.13)%	89.69(±0.21)%	0.512(±0.003)
CIFAR10	Fast Sigmoid(slope,k=25)	82.90(±0.46)%	78.04(±0.40)%	0.535(±0.010)
	aTan(alpha=2)	87.56(±0.41)%	78.99(±0.38)%	0.561(±0.010)
	Spike Rate Escape(beta=1, slope=25)	89.92(±0.39)%	79.95(±0.33)%	0.567(±0.013)
	Triangular	21.43(±0.76)%	23.92(±0.86)%	0.550(±0.036)
	Straight Through Estimator	53.10(±0.66)%	62.49(±0.54)%	0.644(±0.034)
Iris	Fast Sigmoid(slope,k=25)	82.50(±2.46)%	90.00(±6.34)%	0.654(±0.022)
	aTan(alpha=2)	96.67(±1.46)%	93.33(±0.50)%	0.563(±0.095)
	Spike Rate Escape(beta=1, slope=25)	96.67(±1.00)%	100.00(±0.00)%	0.543(±0.036)
	Triangular	100.00(±0.00)%	100.00(±0.00)%	0.510(±0.029)
	Straight Through Estimator	97.50(±0.04)%	100.00(±0.00)%	0.542(±0.108)
Breast Cancer	Fast Sigmoid(slope,k=25)	100.00(±0.00)%	100.00(±0.00)%	0.494(±0.013)
	aTan(alpha=2)	100.00(±0.00)%	97.37(±0.11)%	0.538(±0.015)
	Spike Rate Escape(beta=1, slope=25)	100.00(±0.00)%	97.37(±0.11)%	0.512(±0.026)
	Triangular	97.58(±0.12)%	98.25(±0.00)%	0.497(±0.022)
	Straight Through Estimator	99.78(±0.12)%	98.25(±0.00)%	0.482(±0.024)

TABLE III: Impact of training SNNs with different surrogate gradients on Privacy Vulnerability and Model Performance across different datasets.

perturbations from quantization do not significantly alter the model's susceptibility to MIAs.

From both accuracy and vulnerability perspectives, extreme activation quantization (2-bit precision) provides no tangible benefit in neither ANNs nor SNNs. Thus further reducing activation bit width beyond moderate levels (4-bit or 8-bit) is ineffective and unnecessary. When comparing the impact of activation and weight quantization in MIA vulnerability for ANNs and SNNs, it is evident that SNNs consistently demonstrate lower MIA vulnerability compared to their ANN counterparts, regardless of quantization levels. Remarkably, even fully precise SNN models, which theoretically should be more vulnerable than quantized models, exhibit lower privacy vulnerability than quantized ANNs at any bit width. This highlights the inherent privacy advantages of SNNs due to their stochastic spike-based encoding and temporal dynamics, which naturally disrupt predictable patterns that attackers exploit.

C. Surrogate Gradients

Surrogate gradients are evaluated from three angles: privacy vulnerability, performance, and trade-off between the two as depicted in Table III.

From the perspective of vulnerability, Spike Rate Escape stands out as the most resilient surrogate gradient across most datasets, effectively lowering attack success rates. Its decay parameter introduces sufficient noise in spike activation patterns, making it harder for attackers to infer training data. In contrast, arctangent often demonstrates the highest vulnerability, likely due to its smoother gradient approximations that fail to sufficiently disrupt predictable patterns. Straight Through Estimator (STE) also exhibits high vulnerability in complex datasets, such as CIFAR-10, where its simplistic gradient approximation inadequately masks sensitive patterns.

In terms of performance, Fast Sigmoid consistently delivers the best results across datasets due to its sharp gradient slopes, allowing precise updates and better feature representation. In contrast, Triangular struggles significantly, performing poorly across datasets. Its weaker gradient approximations may provide insufficient feedback for parameter updates, especially in feature-rich datasets. Additionally, its inability to leverage the stochastic nature of SNNs effectively may contribute to both lower accuracy and a failure to disrupt predictable patterns.

When balancing vulnerability and performance, Spike Rate Escape offers the best trade-off, combining strong accuracy with consistently lower vulnerability. Its ability to integrate noise effectively complements its robust feature learning capabilities. In contrast, arctangent and STE fail to strike this balance, as their high vulnerability undermines their moderate performance, making them less effective for scenarios prioritizing both privacy and accuracy.

VI. CONCLUSION

The growing implementation of machine learning across sectors like healthcare, finance, and education has raised concerns about potential privacy breaches through inference attacks, particularly when models process sensitive data. Previous research examining privacy characteristics of neural networks has shown that SNNs exhibit better resilience against MIAs compared to traditional ANNs. This study explores how the privacy-preserving characteristics of SNNs can be further enhanced through quantization and by leveraging various surrogate gradient training methods. Our results show that while quantization reduces MIA vulnerability in both SNNs and ANNs, the privacy advantage of SNNs remains inherent. Notably, even full-precision SNNs exhibit lower vulnerability than quantized ANNs at any bit width, reinforcing the fundamental privacy

resilience of spike-based computation over traditional neural architectures. The training of SNNs with different surrogate gradients further highlights their impact on balancing accuracy and privacy. Spike Rate Escape provides the best privacy protection while maintaining strong performance, whereas Arctangent and Straight Through Estimator (STE) exhibit higher vulnerability. Looking forward, we will investigate the energy efficiency of full-precision SNNs compared to quantized ANNs, as the privacy advantage of SNNs holds greater significance when evaluated alongside their energy savings. Additionally, future work will integrate Differentially Private Stochastic Gradient Descent (DPSGD) with quantized ANN and SNN models to explore the synergistic impact of these techniques on both model performance and privacy.

References

[1] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data security and privacy in healthcare: A review," *Procedia Computer Science*, vol. 113, pp. 73–80, 2017.

[2] M. Tripathi and A. Mukhopadhyay, "Financial loss due to a data privacy breach: An empirical analysis," *Journal of Organizational Computing and Electronic Commerce*, vol. 30, no. 4, pp. 381–400, 2020.

[3] D. Florea and S. Florea, "Big data and the ethical implications of data privacy in higher education research," *Sustainability*, vol. 12, no. 20, p. 8744, 2020.

[4] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[5] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! a survey of attacks on private data," *Annual Review of Statistics and Its Application*, vol. 4, no. 1, pp. 61–84, 2017.

[6] G. Golla, "Security and privacy challenges in deep learning models," *arXiv preprint arXiv:2311.13744*, 2023.

[7] A. Moshruba, I. Alouani, and M. Parsa, "Are neuromorphic architectures inherently privacy-preserving? an exploratory study," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273962803

[8] H. Poursiami, I. Alouani, and M. Parsa, "Brainleaks: On the privacy-preserving properties of neuromorphic architectures against model inversion attacks," *arXiv preprint arXiv:2402.00906*, 2024.

[9] Q. Meng, M. Xiao, S. Yan, Y. Wang, Z. Lin, and Z.-Q. Luo, "Training high-performance low-latency spiking neural networks by differentiation on spike representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 444–12 453.

[10] W. Olin-Ammentorp, K. Beckmann, C. D. Schuman, J. S. Plank, and N. C. Cady, "Stochasticity and robustness in spiking neural networks," *Neurocomputing*, vol. 419, pp. 23–36, 2021.

[11] A. Famili and Y. Lao, "Deep neural network quantization framework for effective defense against membership inference attacks," *Sensors*, vol. 23, no. 18, p. 7722, 2023.

[12] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science*, vol. 2, no. 1, pp. 10–19, 2022.

[13] T. Kang, L. Liu, H. He, J. Zhang, S. Song, and K. B. Letaief, "The effect of quantization in federated learning: Ar\'enyi differential privacy perspective," *arXiv preprint arXiv:2405.10096*, 2024.

[14] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 245–248.

[15] J. K. Eshraghian, M. Ward, E. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1016–1054, 2023.

[16] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," 2009, accessed: 2024-05-28. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[17] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[18] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[19] L. Omelina, J. Goga, J. Pavlovicova, M. Oravec, and B. Jansen, "A survey of iris datasets," *Image and Vision Computing*, vol. 108, p. 104109, 2021.

[20] M. Zwitter and M. Soklic, "Breast Cancer," UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C51P4M.

[21] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS genetics*, vol. 4, no. 8, p. e1000167, 2008.

[22] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[23] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[24] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[25] B. Han, Q. Fu, and X. Zhang, "Towards privacy-preserving federated neuromorphic learning via spiking neuron models," *Electronics*, vol. 12, no. 18, p. 3984, 2023.

[26] Y. Kim, Y. Venkatesha, and P. Panda, "Privatesnn: privacy-preserving spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1192–1200.

[27] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X.-s. Hua, "Quantization networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7308–7316.

[28] C. Kowalski, A. Famili, and Y. Lao, "Towards model quantization on the resilience against membership inference attacks," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3646–3650.

[29] W. Wei, Y. Liang, A. Belatreche, Y. Xiao, H. Cao, Z. Ren, G. Wang, M. Zhang, and Y. Yang, "Q-snns: Quantized spiking neural networks," *arXiv preprint arXiv:2406.13672*, 2024.

[30] R. Yin, Y. Li, A. Moitra, and P. Panda, "Mint: Multiplier-less integer quantization for energy efficient spiking neural networks," in *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2024, pp. 830–835.

[31] C. J. Schaefer and S. Joshi, "Quantizing spiking neural networks with integers," in *International Conference on Neuromorphic Systems 2020*, 2020, pp. 1–8.

[32] C. J. Schaefer, P. Taheri, M. Horeni, and S. Joshi, "The hardware impact of quantization and pruning for weights in spiking neural networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 5, pp. 1789–1793, 2023.

[33] R. V. W. Putra and M. Shafique, "Q-spinn: A framework for quantizing spiking neural networks," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[34] C. Li, L. Ma, and S. Furber, "Quantization framework for fast spiking neural networks," *Frontiers in Neuroscience*, vol. 16, p. 918793, 2022.

[35] J. Wang, D. Zhao, G. Shen, Q. Zhang, and Y. Zeng, "Dpsnn: A differentially private spiking neural network with temporal enhanced pooling," *arXiv preprint arXiv:2205.12718*, 2022.

[36] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[37] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[38] E. De Cristofaro, "An overview of privacy in machine learning," *arXiv preprint arXiv:2005.08679*, 2020.

[39] R. Gomm, M. Hammersley, and P. Foster, "Case study and generalization," *Case study method*, pp. 98–115, 2000.

[40] Y. Guo, "A survey on methods and theories of quantized neural networks," *arXiv preprint arXiv:1808.04752*, 2018.

[41] O. Krestinskaya, L. Zhang, and K. N. Salama, "Towards efficient in-memory computing hardware for quantized neural networks: state-of-the-art, open challenges and perspectives," *IEEE Transactions on Nanotechnology*, 2023.

[42] N. Ketkar, J. Moolayil, N. Ketkar, and J. Moolayil, "Introduction to pytorch," *Deep learning with python: learn best practices of deep learning models with PyTorch*, pp. 27–91, 2021.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] A. Pappalardo, "Xilinx/brevitas," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.3333552