Hardware Support for Trustworthy Machine Learning: A Survey

Md Shohidul Islam*§, Ihsen Alouani †‡, Khaled N. Khasawneh*
*ECE Dept., George Mason University, Fairfax, VA, USA
§CSE Dept., Dhaka University of Engineering and Technology, Gazipur, Bangladesh
† IEMN CNRS 8520, INSA Hauts-de-France, UPHF, Valenciennes, France
‡CSIT, Queen's University Belfast, UK
Email: {mislam20, kkhasawn}@gmu.edu, i.alouani@qub.ac.uk

***Abstract*—Machine Learning (ML) are used in an increasing number of applications as they continue to deliver state-of-the-art performance across many areas including computer vision natural language processing (NLP), robotics, autonomous driving, and healthcare. While rapid progress in all aspects of ML development and deployment is occurring, there is a rising concern about the trustworthiness of these models, especially from security and privacy perspectives. Several attacks that jeopardize ML models' integrity (e.g. adversarial attacks) and confidentiality (e.g. membership inference attacks) have been investigated in the literature. This, in return, triggered substantial work to protect ML models and advance their trustworthiness. Defenses generally act on the input data, the objective function, or the network structure to mitigate adversarial effects. However, these proposed defenses require substantial changes to the architecture, retraining procedure, or incorporate additional input data processing overheads. In addition, often these proposed defenses require high power and computational requirements, which make them challenging to deploy in embedded systems and Edge devices. Towards addressing the need for robust ML at acceptable overheads, recent works have investigated hardware-emanated solutions to enhance ML security and privacy. In this paper, we summarize recent works in the area of hardware support for trustworthy ML. In addition, we provide guidelines for future research in the area by identifying open problems that need to be addressed.**

I. Introduction

Machine Learning (ML) has become the linchpin of technological progress, permeating an expanding array of applications and consistently delivering state-of-the-art performance across diverse domains [1]. From revolutionizing computer vision and natural language processing (NLP) to steering advancements in robotics, autonomous driving, and healthcare, ML has ushered in a new era of possibilities [2]. However, amid the swift progress in ML development and deployment, a formidable challenge looms large—the susceptibility of these systems to security and privacy attacks [3], [4].

One prominent facet of this challenge is the emergence of adversarial examples, cunningly designed imperceptible perturbations injected into input data with the malicious intent of causing ML classifiers to misclassify [5], [6]. The potential fallout from such mispredictions can be profound. Take self-driving cars as an example, where misclassifying a 'stop' sign as a 'yield' sign or a 'speed limit' sign could result in life-threatening situations or significant material damage. Recently, real-world scenarios have witnessed various adversarial attacks, presenting a concerning menace to the safety and security dimensions of ML-powered applications [1].

Simultaneously, ML models are now undergoing training with datasets of increasing sensitivity, encompassing clinical and biomedical records, personal photos, genome data, financial details, social interactions, location traces, and more [3]. Given the intricate nature of machine learning (ML) models and their substantial computational demands, training often occurs on cloud providers offering ML-as-a-Service, such as Amazon AWS, Microsoft Azure, and Google API [3]. This approach allows both novices and professionals to train models that may contain *personally identifiable information (PII)* or potentially sensitive personal data [7]. Safeguarding data privacy in these systems, particularly preventing any leakage of training data, is imperative for establishing trustworthy ML systems.

One of the initial vulnerabilities related to privacy in ML is the membership inference attack (MIA). MIA involves the unauthorized extraction of sensitive information about the private training data solely by accessing the model during the inference phase [8]. Specifically, MIAs can discern whether a particular sample has been employed in training a targeted model. Due to the phenomenon of overfitting to the training data, ML models exhibit biases and manifest distinct behavior on training data (members) compared to test data (non-members) [9]. This bias becomes apparent through a statistically higher confidence of models in classifying members as opposed to non-members. Attackers exploit this bias to execute membership inference attacks effectively. The confluence of these security and privacy threats underscores an urgent imperative to advance the science of machine learning, security, and privacy in tandem, striving for holistic solutions that fortify the robustness of ML frameworks [10], [11].

Since the initial unveiling of adversarial attacks, an extensive body of work has been dedicated to advancing potential defense mechanisms. These defenses commonly revolve around altering the input data, tweaking the objective function, or modifying the network structure to mitigate the adverse effects of attacks [12]. However, a notable drawback of these proposed defenses is their propensity to demand substantial changes to the architecture, retraining procedures, or the

integration of additional processing overheads in the input data [13]. Furthermore, the high power and computational requirements often associated with these defenses pose a significant deployment challenge, especially in embedded systems and applications with stringent latency constraints [14].

In response to the need for robust ML with acceptable overheads, recent research endeavors have pivoted toward exploring the role of hardware in fortifying ML security and privacy [15] [16] [17] [18] [19] [20] [21] [22]. This recent line of research started in 2017 and try to propose hardware-emanated defenses towards having both gains in resources and robustness. This paper provides a comprehensive synthesis and characterization of recent works in the burgeoning field of hardware support for trustworthy ML. By delving into these advancements, the goal of this review paper is not only to illuminate the current landscape but also to offer valuable insights that guide future research. The identification of intriguing open problems within the intersection of hardware and ML security and privacy serves as a compass, pointing the way toward innovative solutions that can reshape the future of secure and private machine learning systems in our interconnected world.

II. Taxonomy of Hardware Support for Trustworthy ML

In this section, we offer a concise overview of our taxonomy, which focuses on hardware support for trustworthy ML. This taxonomy uniquely integrates threat models and hardware support categories, forming a comprehensive framework. Figure 1 shows this taxonomy in detail, systematically organizing various defense mechanisms into specific categories. This taxonomy is designed to offer a clear and organized framework, enabling the readers to easily understand and differentiate between the various defense mechanisms within these categories.

A. Threat Model-based Taxonomy

The threats that ML models face can be categorized into to main categories: security threats and privacy threats.

*1) **ML Security**:* Machine Learning (ML) systems are susceptible to security threats, with adversarial sample attacks being a significant challenge. These attacks involve manipulating input data to mislead the model, potentially causing misclassification in critical applications like autonomous cars, finance, disease diagnosis, and malware detection. Another threat is poisoning attacks, compromising the model's integrity by introducing malicious data during training.

*2) **ML Privacy**:* ML poses a significant privacy threat due to its reliance on personal and sensitive data. Concerns arise from the potential unauthorized access to individual information, as ML models learn (and inadvertently reveal) sensitive patterns during training. ML privacy encompasses *model privacy* and *data privacy*, with membership inference attacks representing a substantial threat to data privacy. These attacks aim to determine if a specific data point was part of the training dataset, compromising the confidentiality of sensitive information by exploiting the model's responses.

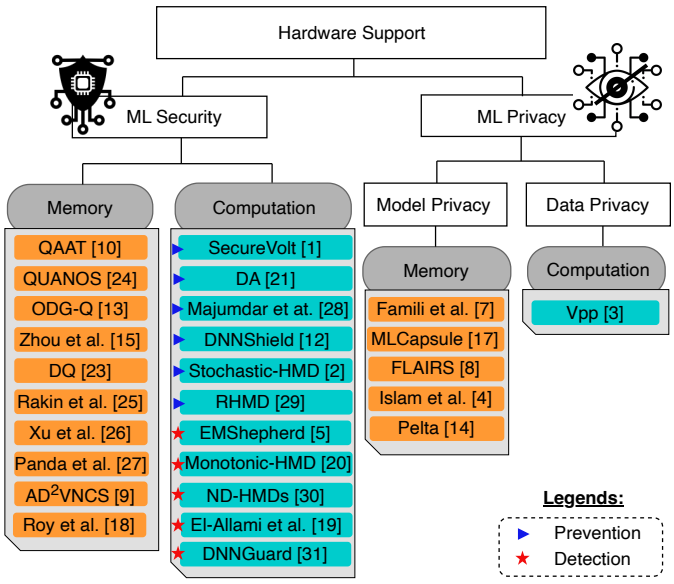

Fig. 1. Taxonomy of Hardware Support for Trustworthy ML. Defenses are grouped under each support category. 'Asterisk' and 'triangle' indicate the defenses that perform attack *detection* and attack *prevention*, respectively.

B. Hardware Support-based Taxonomy

Hardware-based defenses for ML security and privacy are found to leverage two types of supports: Memory and Computation. We elaborate these hardware supports as follows.

*1) **Memory**:* We observe various memory supports that contributed to defense design.

(i) Quantization: Quantization emerges as a robust defense strategy, simultaneously fortifying machine learning models against adversarial sample attacks and enhancing privacy against membership inference attacks. By offloading the computational burden of quantization to specialized hardware, this approach ensures that reduced precision, a hallmark of quantization, is implemented efficiently [15]. By reducing the precision of input itself or model (*e.g.*,, weight, bias, activation), quantization introduces robustness against adversarial sample attacks by diminishing the model's sensitivity to subtle input manipulations [10]. The noise introduced during quantization acts as a deterrent, disrupting the optimization processes used by attackers. Simultaneously, the reduced precision and abstraction make it more challenging for adversaries to discern specific training data points, bolstering the model's privacy against membership inference attacks [23].

(ii) Non-Volatile Memory (NVM) crossbar non-ideality: There exists efforts to investigate memory supports aimed at enhancing security against adversarial sample attacks. Consider Resistive Random-Access Memory (RRAM) crossbar, for instance. RRAM demonstrates specific variations and imperfections in its resistive states, introducing an element of unpredictability to the stored data. These dynamic inherent non-idealities pose challenges for adversaries attempting to precisely manipulate or reverse-engineer specific memory states, thus hindering the generation of potential adversarial

samples [18].

(iii) Enclave: By providing isolated and trusted execution environment (TEE) within the hardware, secure enclaves (*e.g.,* Intel SGX, ARM TrustZone, AMD SEV) shield sensitive computations from external interference. This protection is pivotal in thwarting adversarial attempts to manipulate model behavior through subtle input perturbations [14]. Besides, secure enclaves contribute significantly to privacy preservation by preventing unauthorized access to critical model information, and thus, it becomes exceedingly challenging for adversaries to glean information about the training data or model parameters [17].

*2) **Computation**:* Under this support category, defenses are found to utilize approximate/stochastic computing through various hardware components: CPU, GPU, FPGA, ASIC, etc. Such computing is observed by *undervolting* of those hardware or introducing *randomized hardware* components. Randomized hardware proves to be a formidable ally in securing machine learning models against adversarial sample attacks and bolstering privacy against membership inference attacks. The introduction of randomness at the hardware level adds an unpredictable element to the computations, making it significantly more challenging for adversaries to craft effective adversarial samples. This inherent variability moves the decision boundary of models unpredictably, disrupting the precision required for adversarial manipulation [1]. Simultaneously, randomized hardware contributes to enhanced privacy by introducing uncertainty in the model's responses, thwarting attempts to infer membership information. The unpredictability injected at the hardware level creates a dynamic defense mechanism, reinforcing both security and privacy aspects [3].

III. Hardware-support for ML Security

This section provides an overview of hardware-based defenses against security attacks. This section primarily focuses on hardware-based defenses against adversarial attacks in ML systems, as there are currently no published hardware-based solutions for countering poisoning attacks, which presents a promising direction for future research. The security defenses are categorized based on the hardware support utilized as a building block for the defense; *memory* and *computational* support categories.

A. Memory-based Defenses

We specifically present the defenses that use *quantization* and other *memory support*. To facilitate a comparative analysis, we consolidate the quantization-based defenses in Table I and elaborate each specific defense as below.

QAAT [10]: Authors introduced Quantization-Aware Adversarial Training (QAAT) for executing 8-bit DNN models on FPGA while maintaining robustness against adversarial examples (AEs). Implemented in the Vitis-AI development environment, the 8-bit parameter QAAT model successfully ran on an FPGA. Evaluation on MNIST and CIFAR-10 tasks demonstrated that QAAT achieved comparable robustness to the 32-bit precision Adversarial Training (AT) model, while a 32-bit precision AT model, applied after clean image training, significantly decreased AE robustness.

QUANOS [24]: QUANOS is a framework for layer-specific hybrid quantization based on Adversarial Noise Sensitivity (ANS). By leveraging the novel noise stability metric ANS for DNNs, QUANOS determines optimal bit-width per layer to enhance adversarial robustness and energy efficiency with minimal accuracy loss. Evaluation on precision-scalable Multiply and Accumulate (MAC) hardware architectures demonstrates QUANOS outperforming homogenously quantized 8-bit precision baselines in terms of adversarial robustness using CIFAR10 and CIFAR100 datasets.

ODG-Q [13]: ODG-Q is a novel method recasting robust quantization as an online domain generalization problem. Utilizing quantization-aware training with a uniform quantizer and XNOR-net for various bit configurations (1,2,4,8 bits) and 1-bit quantization, ODG-Q covers both weights and activation. Demonstrating superior performance, it generates diverse adversarial data at low training cost.

Zhou et al. [15]: This defense introduces a cloud-native service that generates and distributes adversarially robust compressed models for edge deployment using a novel post-training quantization approach (using 6, 7 bits for weight and activation). Experimental results show that, despite vulnerability to universal adversarial perturbation (UAP), post-training quantization on synthesized, adversarially-trained models effectively counters these perturbations.

DQ [23]: Defensive Quantization (DQ) optimizes deep learning model efficiency and robustness by quantizing activation using 1-5 bits. Empirical studies reveal that standard quantization is more vulnerable to adversarial attacks due to an error amplification effect. DQ addresses this by controlling the Lipschitz constant during quantization, defending against adversarial examples and achieving superior robustness compared to full-precision models.

Rakin et al. [25]: This study focuses on using activation quantization as an effective defense against adversarial attacks. This work introduces Dynamic Quantized Activation (DQA), treating activation quantization function thresholds as tuning parameters during adversarial training. Adaptive adjustment of these thresholds plays a crucial role in improving network robustness.

Xu et al. [26]: This work proposes to detect adversarial images by squeezing the input image. For this, the image is processed with color depth bit reduction (5 bits) and smoothed by a 2×2 median filter. The central concept involves comparing the model's predictions on the original input with those on the compressed input during testing. If the low resolution image is classified differently as the original image, then this image is detected as adversarial.

Panda et al. [27]: Discretization significantly boosts DLN robustness against adversarial attacks, reducing pixel levels from 256 values (8-bit) to 4 values (2-bit). Binary neural networks (BNNs) and related variants are intrinsically more robust than full precision counterparts. Combining input discretization with BNNs enhances robustness, eliminating

TABLE I
PROMINENT HARDWARE QUANTIZATION-BASED DEFENSES AND THEIR CHARACTERISTICS.

Defenses	Tested Attacks	Used H/W	# of quantization bits	What is quantized?	When applied?	ML Models	Dataset
QAAT [10]	PGD	FPGA	8 bits	Weight	Train, Test	3-layer MLP, VGG-11	MNIST, CIFAR-10
ODG-Q [13]	FGSM, PGD, BIM, TPGD	GPU	1,2,4,8 bits	Weight, Activation	Train	XNOR-Net, ResNet-18	MNIST, CIFAR-10, ImageNet
Zhou et al. [15]	Universal Adversarial Perturbation	Edge (Jetson Nano)	6,7 bits	Weight, Activation	Test	MobileNet, ResNet, WideResNet	CIFAR-10, CIFAR-100, SVHN
QUANOS [24]	FGSM, PGD	FPGA	8,16 bits	MAC operation	Train, Test	VGG-19, ResNet-18	CIFAR-10, CIFAR-100
DQ [23]	FGSM, PGD	CPU	4 bits	Activation	Train, Test	Wide ResNet, VGG-16	CIFAR-10, SVHN
Rakin et al. [25]	FGSM, PGD, CW	CPU	1-3 bits	Activation (Tanh)	Train, Test	LeNet-5, ResNet-18	MINIST, CIFAR-10
Xu et al. [26]	FGSM, BIM, DeepFool, JSMA, CW	GPU	1-5 bits	Input	Test	7-layer CNN, DenseNet, MobileNet	MNIST, CIFAR-10, and ImageNet
Panda et al. [27]	R-FGSM, PGD, CW	GPU	2,4 bits	Input, Model	Train, Test	BNN, XNOR	MNIST, CIFAR-10, CIFAR-100, and ImageNet
Kowalski et al. [11]	Membership Inference Attack	CPU	4,16 bits	Weight, Activation	Test	ResNet-18	Fashion MNIST, CIFAR-10

the need for adversarial training under certain perturbation magnitudes.

AD^2VNCS **[9]:** Memristive crossbar-based neuromorphic computing systems (NCS) have shown outstanding performance in accelerating neural networks. This paper delves into the robustness of deep neural networks (DNNs) against adversarial sample attacks using Resistive Random Access Memory (RRAM) approximation. This investigation introduces innovative training strategies, including DFS (Deep neural network Feature importance Sampling) and BFS (Bayesian neural network Feature importance Sampling), to simultaneously address device variation and adversarial sample attacks on DNNs.

Roy et al. [18]: Non-Volatile Memory NVM crossbars, characterized by their analog nature, provide efficient Matrix Vector Multiplication (MVM) but introduce approximations. The study explores the consequences of these approximations in adversarial conditions, revealing that the non-ideal behavior of analog computing diminishes the efficacy of adversarial attacks in both Black-Box and White-Box scenarios.

B. Computation-based Defenses

SecureVolt [1], [16]: These works explore inference computation while voltage over-scaling (VOS), a method of reducing the supply voltage without altering the frequency, as a lightweight defense against adversarial attacks. Implementing a moving-target defense for DNNs using the stochastic timing violations induced by VOS, the experiments demonstrate its effectiveness in countering various attack methods in image classification without necessitating software/hardware modifications, while also leading to reduced power consumption.

DA [21]: DA uniquely utilizes an approximate 32-bit floating-point multiplier, called Ax-FPM, designed with aggressive approximation of full adders (FAs) to inject computational noise. This noise, making the decision boundary of DA randomly different from the initial model, is a distinctive feature as it incurs no overhead and naturally arises from a simpler and faster AC implementation. DA serves as a seamless replacement for hardware without specific training requirements or alterations to the CNN architecture or parameters.

Majumdar et at. [28]: This paper introduces a novel method for enhancing image classifier robustness using controlled undervolting of FPGA to induce compute errors during the inference process. These errors disrupt adversarial inputs, enabling correction or detection of adversarial attacks. Evaluation on FPGA design and software simulation demonstrates average detection rates of 77% and 90% for 10 tested attacks on two widely used DNNs.

DNNShield [12]: The work reveals that existing approximate computing approaches, while effective for various inputs, fall short against stronger, high-confidence adversarial attacks. In response, it introduces DNNSHIELD, a hardware-accelerated defense that adjusts response strength based on adversarial input confidence. DNNSHIELD employs dynamic and random sparsification through hardware to efficiently approximate inference with fine-grained control over error. Adversarial inputs are detected by analyzing the output distribution characteristics of sparsified inference compared to a dense reference.

Stochastic-HMDs [2]: Stochastic-HMDs, a defense in the malware detection domain, utilizes approximate computing to enhance hardware malware detectors' (HMDs) resilience against evasive malware attacks. By introducing controlled undervolting to induce stochastic timing violations during

detection, these detectors offer effective defense, especially against reverse-engineering and transferability.

RHMD [29], Opt-RHMD [22]: Resilient Hardware Malware Detectors (RHMDs) comprise multiple base detectors and dynamically switch between detectors in a stochastic manner, increasing the difficulty of reverse engineering. RHMDs exhibit resilience against both reverse engineering and evasion, providing a robust defense against evasive malware with minimal additional complexity. However, the random switching employed by RHMDs may potentially decrease overall detection accuracy. Opt-RHMDs introduce an optimized switching strategy, formulated through a Bayesian Stackelberg game, aiming to maximize accuracy in detecting both evasive and non-evasive malware.

EMShepherd [5]: EMShepherd is a framework leveraging electromagnetic (EM) emanations during model inference. By capturing and processing EM traces, the work detects adversarial attacks using only benign samples for training. EMShepherd, functioning in an air-gapped environment, effectively identifies various adversarial attacks on a widely used FPGA deep learning accelerator.

Monotonic-HMDs [20]: MonotonicHMDs, utilizing only the malicious features, provide a robust defense against adversarial evasion attacks on Hardware Malware Detectors (HMDs). This approach ensures that adding benign features to malware won't evade detection, and incorporating malicious features increases the probability of malware detection.

ND-HMDs [30]: Non-differentiable Hardware Malware Detectors (ND-HMDs) defend against transient execution attacks by using gradient-free classifiers like Decision Trees and Random Forests, reducing the effectiveness of evasion through obfuscated Hardware Performance Counter (HPC) traces. ND-HMDs successfully resist gradient-based and sleep-based attacks while maintaining high detection accuracy on non-evasive transient execution attacks, offering a robust defense against hardware malware threats.

El-Allami et at. [19]: The investigation focuses on the robustness of SNNs to adversarial attacks, considering different values for neuron firing voltage thresholds and time window boundaries. This work conducts a comprehensive analysis of SNN security, examining various adversarial attacks in a strong white-box setting with different noise budgets and variable spiking parameters.

DNNGuard [31]: This paper presents DNNGuard, an efficient elastic DNN accelerator that integrates original DNN networks and a detection algorithm for adversary sample attacks into a single chip. The architecture enhances data transfer efficiency and information protection, featuring an extended AI instruction set for neural network synchronization and task scheduling. Implemented on RISC-V and NVDLA, DNNGuard effectively validates the legitimacy of input samples, as shown in experimental results.

IV. Hardware-support for ML Privacy

This section summarizes hardware-based defenses for ML privacy, particularly against model privacy and data privacy. Interestingly, defenses that protect the model privacy are memory based and the only defense against data-privacy is computation-based.

A. Memory-based Defenses

All of the following defenses protect the model privacy.
Famili et al. [7]: This paper explores the impact of quantization on privacy leakage and introduces a novel quantization method designed to enhance a neural network's resistance against Membership Inference Attacks (MIA). Unlike conventional quantization methods focusing on compression or speed, this proposed framework prioritizes defense against MIA. Evaluation on various benchmark datasets and model architectures demonstrates improved metrics, including precision, recall, and F1-score, compared to full bitwidth models.

MLCapsule [17]: MLCapsule offers a secure offline deployment solution for Machine Learning as a Service (MLaaS) through SGX-enclave protection. This approach enables local execution of the machine learning model on the user's client, ensuring data privacy as it remains on the client side. MLCapsule provides service providers with control and security equivalent to server-side execution, safeguarding against direct model access and offering defenses against advanced attacks like model stealing, reverse engineering, and membership inference.

FLAIRS [8]: This research explores leveraging FPGA-based TEE computation to enhance the security of Federated Learning (FL) against backdoor and inference attacks. By utilizing FPGA-based enclaves, the approach addresses inference attacks during FL aggregation, employing an advanced backdoor-aware aggregation algorithm on the FPGA to counter backdoor threats.

Islam et al. [4]: The paper introduces T-Slices, a framework for securely incorporating memory-intensive DL models into ARM TrustZone-based embedded devices with limited trusted memory. It ensures protected inference of pre-trained models, leveraging TrustZone's security features to safeguard data and model parameters. T-Slices effectively defends against both black-box and white-box membership inference attacks by requiring decryption within the secure TrustZone memory, thwarting access to training data and model information.

Pelta [14]: Pelta is a defense against gradient-based evasion attacks in Federated Learning, utilizing hardware obfuscation to hide critical in-memory values near the input during inference. This work relies on hardware-enabled trusted execution environments (TEEs) in Arm TrustZone, offering privacy and integrity guarantees for Vision Transformer (ViT) models. However, TrustZone's limited enclave memory, particularly for models exceeding 500 MB, requires Pelta to be a light, partial obfuscation of the model.

B. Computation-based Defenses

So far, there is only one recent hardware-based defense that protect the data privacy of the ML models.

V_{PP} **[3]:** Privacy Preserving Volt (V_{PP}) introduces a lightweight inference-time defense method using undervolting for privacy-preserving machine learning. VPP maintains

protected models' utility without re-training, obscures the outcome of membership inference attacks (MIAs) by introducing computational randomness through undervolting FPGA during inference, and achieves a compelling utility/privacy tradeoff, surpassing prior defenses.

V. Concluding Remarks

This comprehensive review paper presents a systematic taxonomy that bridges the gap between threat models in Machine Learning and hardware support, paving the way for more trustworthy ML systems. This taxonomy not only simplifies the complexity of various defense strategies but also provides a structured approach for researchers and practitioners to navigate and select appropriate hardware-based solutions for enhancing the security and privacy of ML systems.

Acknowledgment

This work has been supported in part by the National Science Foundation grant CCF-2212427.

References

[1] M. S. Islam, I. Alouani, and K. N. Khasawneh, "Securevolt: Enhancing deep neural networks security via undervolting," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2023.

[2] M. S. Islam, I. Alouani, and K. N Khasawneh, "Stochastic-hmds: Adversarial-resilient hardware malware detectors via undervolting," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2023, pp. 1–6.

[3] M. S. Islam, B. Omidi, I. Alouani, and K. N. Khasawneh, "Vpp: Privacy preserving machine learning via undervolting," in *2023 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*. IEEE, 2023, pp. 315–325.

[4] M. S. Islam, M. Zamani, C. H. Kim, L. Khan, and K. W. Hamlen, "Confidential execution of deep learning inference at the untrusted edge with arm trustzone," in *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, 2023, pp. 153–164.

[5] R. Ding, C. Gongye, S. Wang, A. A. Ding, and Y. Fei, "Emshepherd: Detecting adversarial samples via side-channel leakage," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, 2023, pp. 300–313.

[6] S. Queyrut, V. Schiavoni, and P. Felber, "Mitigating adversarial attacks in federated learning with trusted execution environments," in *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2023, pp. 626–637.

[7] 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.

[8] H. Li, P. Rieger, S. Zeitouni, S. Picek, and A.-R. Sadeghi, "Flairs: Fpga-accelerated inference-resistant & secure federated learning," in *2023 33rd International Conference on Field-Programmable Logic and Applications (FPL)*. IEEE, 2023, pp. 271–276.

[9] Y. Bi, Q. Xu, H. Geng, S. Chen, and Y. Kang, "Ad2vncs: Adversarial defense and device variation-tolerance in memristive crossbar-based neuromorphic computing systems," *ACM Transactions on Design Automation of Electronic Systems*, 2023.

[10] Y. Fukuda, K. Yoshida, and T. Fujino, "Evaluation of model quantization method on vitis-ai for mitigating adversarial examples," *IEEE Access*, 2023.

[11] C. Kowalski, A. Famili, and Y. Lao, "Towards model quantization on the resilience against membership inference attacks," in *ICIP*. IEEE, 2022, pp. 3646–3650.

[12] M. H. Samavatian, S. Majumdar, K. Barber, and R. Teodorescu, "Dnnshield: Dynamic randomized model sparsification, a defense against adversarial machine learning," *arXiv preprint arXiv:2208.00498*, 2022.

[13] C. Tao and N. Wong, "Odg-q: Robust quantization via online domain generalization," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1822–1828.

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

[15] X. Zhou, R. Canady, Y. Li, S. Bao, Y. Barve, D. Balasubramanian, and A. Gokhale, "Guarding against universal adversarial perturbations in data-driven cloud/edge services," in *2022 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2022, pp. 233–244.

[16] M. S. Islam, I. Alouani, and K. N. Khasawneh, "Lower voltage for higher security: Using voltage overscaling to secure deep neural networks," in *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE, 2021, pp. 1–9.

[17] L. Hanzlik, Y. Zhang, K. Grosse, A. Salem, M. Augustin, M. Backes, and M. Fritz, "Mlcapsule: Guarded offline deployment of machine learning as a service," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3300–3309.

[18] D. Roy, I. Chakraborty, T. Ibrayev, and K. Roy, "On the intrinsic robustness of nvm crossbars against adversarial attacks," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 565–570.

[19] R. El-Allami, A. Marchisio, M. Shafique, and I. Alouani, "Securing deep spiking neural networks against adversarial attacks through inherent structural parameters," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2021, pp. 774–779.

[20] M. S. Islam, B. Omidi, and K. N. Khasawneh, "Monotonic-hmds: Exploiting monotonic features to defend against evasive malware," in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2021, pp. 97–102.

[21] A. Guesmi, I. Alouani, K. N. Khasawneh, M. Baklouti, T. Frikha, M. Abid, and N. Abu-Ghazaleh, "Defensive approximation: securing cnns using approximate computing," in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2021, pp. 990–1003.

[22] M. S. Islam, K. N. Khasawneh, N. Abu-Ghazaleh, D. Ponomarev, and L. Yu, "Efficient hardware malware detectors that are resilient to adversarial evasion," *IEEE Transactions on Computers*, vol. 71, no. 11, pp. 2872–2887, 2021.

[23] J. Lin, C. Gan, and S. Han, "Defensive quantization: When efficiency meets robustness," *Proceedings of the International Conference on Learning Representations (ICLR) 2019*, 2019.

[24] P. Panda, "Quanos: adversarial noise sensitivity driven hybrid quantization of neural networks," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020, pp. 187–192.

[25] A. S. Rakin, J. Yi, B. Gong, and D. Fan, "Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions," *arXiv preprint arXiv:1807.06714*, 2018.

[26] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[27] P. Panda, I. Chakraborty, and K. Roy, "Discretization based solutions for secure machine learning against adversarial attacks," *IEEE Access*, vol. 7, pp. 70 157–70 168, 2019.

[28] S. Majumdar, M. H. Samavatian, K. Barber, and R. Teodorescu, "Using undervolting as an on-device defense against adversarial machine learning attacks," in *2021 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*. IEEE, 2021, pp. 158–169.

[29] K. N. Khasawneh, N. Abu-Ghazaleh, D. Ponomarev, and L. Yu, "Rhmd: Evasion-resilient hardware malware detectors," in *Proceedings of the 50th Annual IEEE/ACM international symposium on microarchitecture*, 2017, pp. 315–327.

[30] M. S. Islam, A. P. Kuruvila, K. Basu, and K. N. Khasawneh, "Nd-hmds: Non-differentiable hardware malware detectors against evasive transient execution attacks," in *2020 IEEE 38th International Conference on Computer Design (ICCD)*. IEEE, 2020, pp. 537–544.

[31] X. Wang, R. Hou, B. Zhao, F. Yuan, J. Zhang, D. Meng, and X. Qian, "Dnnguard: An elastic heterogeneous dnn accelerator architecture against adversarial attacks," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 19–34.