Middle East Research Journal of Engineering and Technology
ISSN: 2789-7737 (Print) ISSN: 2958-2059 (Online)
Frequency: Bi-Monthly
DOI: 10.36348/merjet.2022.v02i02.009

KUWAIT SCHOLARS PUBLISHER
Website: http://www.kspublisher.com/
Email: office@kspublisher.com

SecureCNN: Strengthening Distributed Inference Security via Strategic PDF Validation Dataset Optimization

Zaynab B. Bello[1*], Temidayo J. Omotinugbon[2], Mabel Ogonna[3], Ifeanyichukwu J. Umoga[4]
[1]Department of Computer Science, Texas Tech University, Lubbock, Texas, United States
[2]Eller College of Management, University of Arizona, Tucson, Arizona, United States
[3]Rawls College of Business, Texas Tech University, Lubbock, Texas, United States
[4]Eller College of Management, University of Arizona, Tucson, Arizona, United States

Abstract: Hardware Accelerator-Based (HAB) CNN Inference is rapidly growing in recognition. In many scenarios, to achieve short-time-to-market, HAB CNN inference can be outsourced to third parties (3P) for design and deployment. These 3P may be malicious and hence embed harmful circuitry in the deployed hardware design. Recently, approaches for embedding harmful circuitry targeted at collaborative inference have been proposed. These approaches make use of statistical analysis on validation dataset (VD) for the design of stealthy attacks on the hardware design of CNNs. In this paper, we propose three approaches to obscure relevant information regarding the VD that may either achieve the detection or mitigation of embedded attacks proposed in these approaches. These three approaches include Gaussian Distribution Shifting Approach (GDSA), Gaussian Distribution Compression Approach (GDCA) and Gaussian Distribution Expansion Approach (GDEA). These approaches are tested on LeNet CNN infected with attack proposed in [1] implemented on Xilinx PYNQ-Z1.

Keywords: Convolutional Neural Network, Security, Hardware, Optimization.

Research Paper

***Corresponding Author:**
Zaynab B. Bello
Department of Computer Science, Texas Tech University, Lubbock, Texas, United States

How to cite this paper:
Zaynab B. Bello, Temidayo J. Omotinugbon, Mabel Ogonna, Ifeanyichukwu J. Umoga (2022). SecureCNN: Strengthening Distributed Inference Security via Strategic PDF Validation Dataset Optimization. *Middle East Res J. Eng. Technol, 2*(2): 54-59.

Article History:
| Submit: 23.10.2022 |
| Accepted: 28.11.2022 |
| Published: 30.12.2022 |

I. INTRODUCTION

Hardware Accelerator-Based (HAB) Convolutional Neural Network (CNN) inference has become very popular in recent times [2] to achieve real-time image classification, object recognition [3], [4] and so on. To achieve short time-to-market and access to state-of-the-art techniques, the mapping and deployment of CNN models on Hardware accelerators are often outsourced to untrusted third parties (3P). The 3P-IP designers may provide soft IPs, firm IPs, or hard IPs to the project owner. Due to the untrusted nature of 3P-IP designers, hardware intrinsic security of the design may be compromised as they may embed malicious circuitry, which is very difficult to detect, especially if the 3P-IP provides hard IPs such as bitstream files or GDSII files. Several approaches of embedding hardware intrinsic attacks in deployed CNN inference architecture have been studied in the literature. Clements et. al [5] proposes an attack that generates perturbations whose addition to output feature maps of one or more targeted layers of the CNN model in runtime can cause misclassification resulting in reduced accuracy. Liu et al. in [6] introduce a neural network-based attack with a focus on the generation of samples of the input image from a CNN model for the trigger and payload design with the primary aim of causing misclassification. These aforementioned approaches require manipulation of the CNN parameters (weights and biases) to achieve their aim of misclassification, which can be detected by model integrity test on the hardware design. These approaches also assume that the attacker has full knowledge of the CNN architecture.

CNNs have the inherent attribute of being computationally heavy and memory intensive, making their deployment of resource- constrained devices challenging [7]. This has led to the adoption of partitioned CNN models, which employ horizontal and vertical collaboration [8–11]. Partitioned CNN models can be considered a defense technique against attacks because the layers of the CNN can be distributed to different 3P-IP designers, no one 3P designer has access to the full CNN architecture, making some of the attack techniques mentioned above ineffective [12]. Recent approaches as seen in [1, 12] and [13] propose attacks targeted at hardware CNN model inference employing edge offloading of partitioned CNN models where the attacker has only access to partial CNN architecture (some layers) of the CNN. To design a stealthy attack in the aforementioned approaches, the attacker requires a validation dataset (VD) for the verification of the implementation correctness of the hardware design. The

VD gives the untrusted 3P designer insights into the behavior of the layers and is hence used to design a stealthy attack. *These challenges lead to the research question of how can the project owner (defender) makes the VD secure against potential attacks from malicious 3P-IP designers.*

In this paper, we demonstrate a mitigation/detection technique focused on attacks that require VD to deploy CNNs on hardware accelerators and vulnerabilities that they may introduce. The proposed methodology informs how the VD for intermediate CNN layers can be modified based on its probability distribution function (ModPDF)- that prevents the attackers at the untrusted 3PIP from exploiting the VD to achieve stealthy hardware attacks. This paper studies three different ModPDF techniques, shifting, compression, and expansion (detailed in Section III). The results show that with the ModPDF approaches, the hardware attacks will either become detectable or will lead to no triggering of hardware attacks whatsoever. This defense approach has no hardware overhead.

II. THREAT MODEL

This work focuses on gray-box attacks where the attacker has only partial knowledge of the CNN architecture and no knowledge of the training or testing dataset. We assume that the third-party IP designer (3PIP) requires a VD [14] for the verification of the implementation correctness of the CNN hardware design. The VD consists of input feature maps and corresponding output feature maps of the partitioned CNN. We also focus on scenarios where edge offloading and collaborative inference is employed in which attackers usually design partitioned CNN design without the first and final layers (as this is considered as more secure [12]). Finally, we also assume the attacker provides the CNN hardware design as a bitstream file to the defender (the project owner).

III. PROPOSED METHODOLOGY TO MODIFY PDF (MODPDF) OF VALIDATION DATASET TO SECURE CNN

Some of the recent hardware attacks on distributed CNNs [1], [12] and [13] make use of 3-σ analysis on the histogram plots of Gaussian Distributed feature maps generated in response to VD to obtain rarely occurring Range of Values (RoV) that can be used for the trigger design. In this work, we propose the modification of the probability density functions (ModPDFs) of these VDs that are provided to the untrusted third parties to design one part of the partitioned CNN, as shown in Fig.1. This work proposes three ModPDF approaches for the original validation dataset (OVD), namely: Gaussian Distribution Shifting Approach (GDSA), Gaussian Distribution Compression Approach (GDCA), and Gaussian Distribution Expansion Approach (GDEA). These approaches are further discussed in subsequent subsections. These approaches are used to generate a complementary validation dataset (CVD) as shown in Fig. 1 in the design phase. These CVDs and the partitioned CNN model are provided to the untrusted 3P-IP designers for the hardware design. In the testing phase (indicated on the right-hand side of Fig. 1), the hardware design IP is tested with the OVD by the system integrator, which will either lead to the detection or mitigation of any embedded malicious circuitry in the hardware design (details provided in subsections).

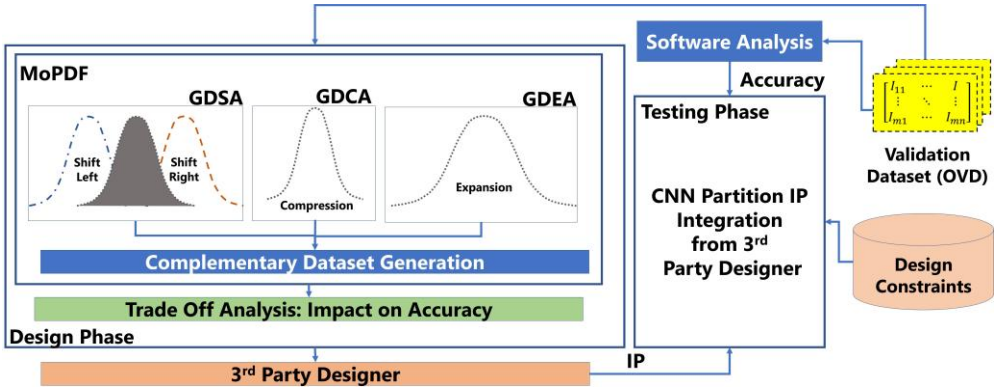

Fig. 1: Conceptual depiction of the overall methodology showing the 3 approaches proposed in this work to achieve the obscurity of relevant information in the VD. The diagram shows the Gaussian distribution Shifting Approach (GDSA), which requires the left or right shift of the Original Gaussian distribution by increasing or decreasing the mean of the original dataset. The diagram shows Gaussian distribution Compressing and Expansion Approaches which requires the compression and expansion of the Original Gaussian distribution by increasing or decreasing the standard deviation of the original dataset.

A. Gaussian Distribution Shifting Approach (GDSA)

The Gaussian Distribution Shifting Approach (GDSA) approach considers a scenario where the Original validation dataset (OVD) is altered to create a CVD whose mean is centered around the 3-σ region of the OVD, i.e., it is shifted either to the left or right as shown in Fig. 2. For empirical analysis, we consider LeNet CNN [15], and for the sake of argument, it is assumed that it is are partitioned from conv2-ip1, as shown at the top of the Fig. 2, indicated as CNN and

CNN partition. OVD is obtained by using 100 images from MNIST dataset [16] and obtained its feature maps at the output of pool1 layer. The characteristics of the Gaussian PDF of the OVD is obtained and labeled as G (see Fig. 2b). The GDSA provides defense by shifting the Gaussian PDF (G) of OVD to either left or right as illustrated conceptually in Fig. 1 and with empirical values in Figs. 2a and 2c. The new data set, CVD, has Gaussian PDF of G'_S (when it is shifted left, Fig. 2a), and G''_S (when shifted right, Fig. 2c). Hence, from Fig. 2, If an attacker makes use of values beyond the 3-σ regions from G_S' or G_S'' in Fig. 2) for the trigger design, it may lead to frequent triggering and hence detection. This is because the selected values beyond the 3-σ of G_S' or G_S'' lie in the frequently occurring regions of the original Gaussian PDF (G). Hence if the attacker makes use of values that lie in regions R2, R3 of G'_S, or G''_S, it will lead to detection. If the attacker makes use of values that lie in regions R1, R4 of G'_S or G''_S, it will lead to low or no triggering of the attack and hence mitigation of the attack

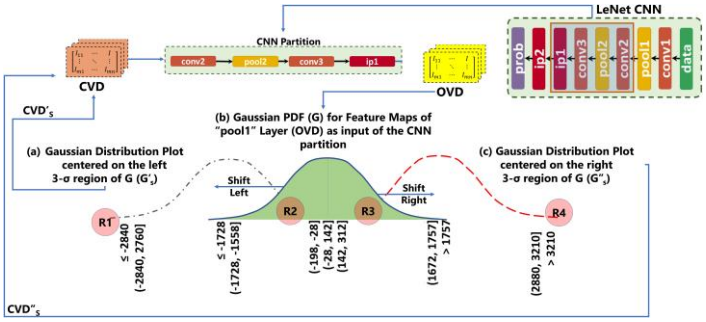

Fig. 2: 3-σ statistical exploratory analysis of output feature map of *pool* layer serving as OVD. We observe different Gaussian PDFs. Fig. 2b (green) is showing original PDF (*G*) of the *pool* feature map. Fig. 2a is illustrating the Gaussian PDF whose mean is shifted to the left (the mean of Fig. 2a is the negative 3-σ value of the original Gaussian PDF shown in the middle). Similarly, in Fig. 2c, the Gaussian PDF is shifted to the right (mean of Fig. 2c is the positive 3-σ value of Fig. 2b).

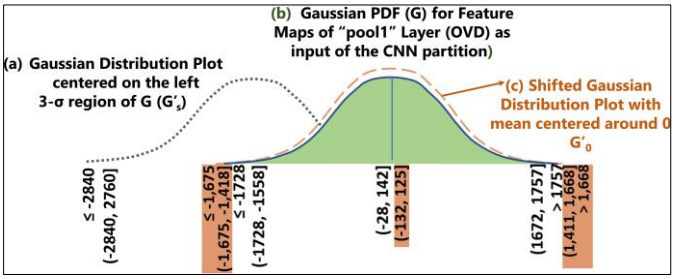

Fig. 3: Depiction of drawback to the GDSA where an estimate of the original low occurring range of values can be obtained by generating random values whose mean is centered around zero while maintaining the same standard deviation. Fig. 3b shows 3-σ values in the range similar to the original dataset in Fig. 3b.

Limitations of GDSA: As an example, we empirically observed that the mean of the feature maps of all the layers of LeNet CNN ranges from -100 to 100 (sometimes very close to zero) shown in the fourth column of Table I. Next, by analyzing the shifted CV D, CV Ds′, as an example, we realize that its Gaussian PDF, G'_S, can raise a suspicion alarm for the malicious user if it has the understanding of mean behavior of the feature maps of similar CNN. This is because the attacker can observe that G'_S or G''_S have means that are deviated far to the left and the right from the observed close to zero mean for G, respectively. In such a case, the attacker can easily generate a new dataset centered around meanwhile intelligently utilizing the known values of σ and RoV to generate a new Gaussian PDF, very similar to G. We empirically proved this analysis as shown as the modified graph G'_O in Fig. 3. This new PDF, G'_O, can neutralize all the defense strategies that are provided by the GDSA. To counter this neutralizing technique from the attacker, we propose two more techniques (detailed in the following two subsections), via compression and expansion of the Gaussian PDF, which does not require the re-centering of (G).

B. Gaussian Distribution Compression Approach (GDCA)

In Gaussian Distribution Compression Approach (GDCA), the standard deviation (σ) of the Gaussian PDF (G) of OVD is collected and reduced to obtain a reduced deviation which results in the compressed Gaussian PDF (G'_C), thereby modifying the PDF of the original distribution (G) as shown in Fig. 4. The compressed Gaussian PDF (G'_C) is used to generate a complementary dataset. This is provided to the third-party designer as shown in Fig. 1. From Fig. 4, it can be observed that rarely occurring RoVs that lie beyond the 3-σ values of the compressed Gaussian PDF (G'_C) falls in the region of values of high frequency in the original Gaussian distribution G. Hence, GDCA will detect attacks or malicious circuits embedded in the

CNN hardware design, which requires rarely occurring RoVs obtained from the validation dataset.

C. Gaussian Distribution Expansion Approach (GDEA)

We also propose a Gaussian Distribution Expansion Approach (GDEA), like GDCA, the standard deviation (σ) of the OVD is collected and increased to achieve the expansion and modification of the PDF of the original Gaussian distribution (G) to obtain a complementary Gaussian PDF (G'_E) as shown in Fig. 5. The expanded Gaussian PDF (G'_E) can be used to generate CVD. From Fig. 5, the rarely occurring RoVs that lie beyond the 3-σ values of the expanded Gaussian distribution (G'_E) falls in the region non-existent in original Gaussian distribution G. Hence, with GDEA, any attack embedded using this approach has a very low likelihood of triggering, resulting in the mitigation of likely attacks.

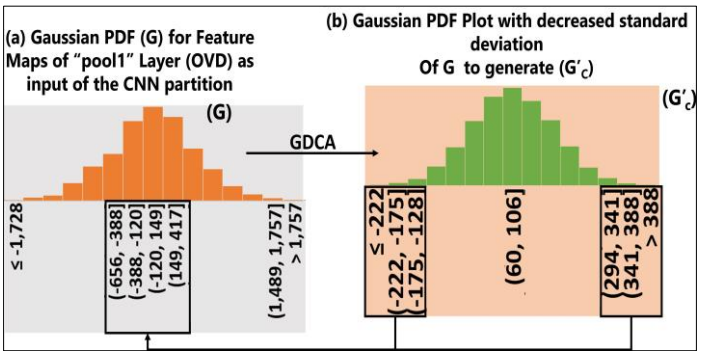

Fig. 4: Gaussian distribution Compression showing that low occurring values beyond the 3-σ values in Fig. 4b lies in the range of values with the probability of high occurrence in Fig. 4a. Hence if the low occurring values of Fig. 4b are used to design an attack, it will have high triggering probability lead to detection.

TABLE I: Results showing the comparison between selected RoVs, X,σ, number of triggering (from the attacker's perspective) and number of wrong predictions (from the defender's perspective) of the original dataset (OVD) in comparison with the usage of CVD using our proposed defense approaches (GDSA, GDCA, GDEA) on LeNet CNN subjected to SoWaF [1] and FeSHI [12] attacks.

Layer	Original Dataset (OVD)					GDSA (CVD'_S)				
				Attack					Attack	
	RoV	$\overline{X}$—σ	% of Trig. (Attacker's View)	SoWaF % of Wrong Pred. (Defender's View)	FeSHI % of Wrong Pred. (Defence View)	RoV	$\overline{X}$—σ	% of Trig. (Attack View)	SoWaF % of Wrong Pred. (Defender's View)	FeSHI % of Wrong Pred. (Defender's View)
*pool*1	-	54.7‖153.3	-	-	-	-	680‖153	-	-	
*conv*2	-1.73e3, -1.71e3	-177.0‖402.2	11	9	15	71, 145	-1865.7‖991.3	15	74	39
*pool*2	-7.1e2, -7.05e2	78.24‖358.0	12	9	17	212, 332	-1687.1‖980.9	6	73	94
*conv*3	1151, 1205	-5.79‖56.4	4	4	21	-1099, -839	333.6‖1716.6	9	48	53
*fc*1	1780, 1790	84.2‖561	8	8	18	1326, 1581	-3026‖3281.2	18	64	50
	GDCA (CVD'_C)					GDEA (CVD'_E)				
*pool*1	-	54.7‖20	-	-		-	54.7‖500	-	-	
*conv*2	-326, -320	-148.7‖81.2	10	74	39	1985, 2030	-33.8‖607.0	8	0	0
*pool*2	-280, -271	-123.2‖76.3	16	61	79	1932. 1990	544.2‖484.4	17	0	0
*conv*3	-123, -105	244.8‖120.7	12	67	76	-2797, -2687	-1103.2‖696.4	21	0	0
*fc*1	-706, -666	-222.3‖231.5	7	53	51	3790, 4010	1061.1‖1300.9	14	0	0

IV. Experiment Setup, Results, and Discussion

A. Experimental Setup

The mapped CNN IP is designed using Xilinx's Vivado and Vivado HLS 2018.3 and to generate an IP. Vivado is used to integrate the generated IP with AXI-interconnects and ZYNQ processor. The proposed methodologies are implemented on LeNet (Fig. 2) trained on the MNIST dataset. In this work, we propose four different scenarios, where the statistical attributes of the output feature maps in response to the VD (OVD or CVD) of each layer (from *conv*2 to *ip*1) of the partial architecture are collected and evaluated. To prove the robustness of our defense approach, experiments were conducted to test each approach against both SoWaF [1], and FeSHI [12] attacks. We implement both attacks on the partial architecture and determine the number of triggering (from the attacker's perspective) and the number of wrong predictions (from the defender's perspective) using the proposed approaches.

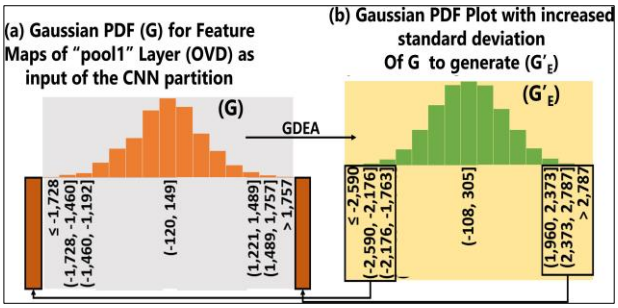

Fig. 5: Gaussian distribution Expansion showing that low occurring values beyond the 3-σ values in Fig. 5b lies out of bounds of the range of values in Fig. 5a. Hence if the low occurring values of Fig. 5b are used to design an attack, it will have very low or no triggering probability lead to mitigation of the attack.

B. Discussions

Table I shows statistical results and a comparison of the three defense approaches and how they perform detection or mitigation of these attacks. From Table I, a scenario where the attacker has access to the OVD (of 100 data instances) from *pool*1 layer for the trigger design. Table I shows the result in terms of the number of triggering evident to the attacker in comparison with the number of wrong predictions experienced by the defender when the hardware design of the CNN partition is merged with the full CNN architecture and tested with the MNIST image. From Table I the percentage number of triggering from the attacker is comparable to the number of the wrong prediction experienced by the defender for both attacks (as shown in rows 5-8 and columns 5-6 of Table I). For example, in *conv*2 layer, the percentage of trigger occurrence is 11% with 100 data instances in the VD, which translates to 9% of the wrong prediction when tested on 1000 images for SoWaF attack. In the GDSA ModPDF approach, the mean of the VD provided to the attacker is centered around 680 compared to 54.7 of the OVD. The RoVs selected for the attacks lie in the region (R3) from Fig. 2. From the result, it can be seen that the number of wrong predictions visible to the defender is at least 5 times more than the rate of triggering of the attack visible to the attacker for SoWaF attack. For example, is in the case of *conv*2 layer, the attack percentage of attack triggering is 15% on 100 instances of the VD, but the rate of wrong prediction evident to the defender is 74% when tested on 1000 input images leading to detection of the attacks. In the GDCA ModPDF approach, the generated ($CV\,D'$) is obtained by decreasing the standard deviation (σ) of the aggregated OVD. From Table I, it is evident that the rate of triggering of the CNN is stealthy from the attacker's point of view but leads to at least 53% rate of obtaining wrong prediction (in *ip*1 layer) at the defender's end during testing for SoWaF attack. This leads to the detection of the attacks across all the layers of the partitioned CNN mode. Hence GDCA ModPDF approach leads to the detection of attacks. In the GDEA ModPDF approach, the generated CVD ($CV\,D'$) is obtained by increasing the standard deviation (σ) of the aggregated values of the OVD. This produces an expanded Gaussian PDF. From Table I, across all the layers of the partitioned CNN model, the number of wrong predictions is 0 when the IP design is tested with the VD of 1000 input images, Hence the SoWaF attack is never triggered. Hence GDEA ModPDF approach leads to the mitigation of attacks. Similar behavior can also be observed from Table I across all layers for the FeSHI attack.

V. Conclusion

This work proposes defense approaches targeted at hardware attacks focused on collaborative and distributed inference. These hardware attacks require a validation dataset (VD) for the design of the trigger by the attacker (usually a third-party IP designer). Our proposed defense, ModPDF, ensures secure CNN inference by modifying the probability density function (PDF) of the VD to generate a complementary validation dataset given to the third party. We propose three modification approaches, namely Gaussian Distribution Shifting Approach (GDSA), Gaussian Distribution Compression Approach (GDCA), and Gaussian Distribution Expansion Approach (GDCA). From the results, these approaches successfully detected or mitigated the embedded attacks.

References

1. T. A. Odetola and S. R. Hasan, "Sowaf: Shuffling of weights and feature maps: A novel hardware intrinsic attack (hia) on convolutional neural network (cnn)," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.
2. K. Abdelouahab, M. Pelcat, J. Serot, and F. Berry, "Accelerating cnn inference on fpgas: A survey," *arXiv preprint arXiv:1806.01683*, 2018.
3. A. Aboah, "A vision-based system for traffic anomaly detection using deep learning and decision trees," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4207–4212.
4. M. Shoman, A. Aboah, and Y. Adu-Gyamfi, "Deep learning framework for predicting bus delays on multiple routes using heterogenous datasets," *Journal of Big Data Analytics in Transportation*, vol. 2, no. 3, pp. 275–290, 2020.
5. J. Clements and Y. Lao, "Hardware trojan attacks on neural networks," *arXiv preprint arXiv:1806.05768*,

2018.
6. Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," 2017.
7. S. Dey, A. Mukherjee, A. Pal, and P. Balamuralidhar, "Partitioning of cnn models for execution on fog devices," in *Proceedings of the 1st ACM International Workshop on Smart Cities and Fog Computing*, 2018, pp. 19–24.
8. R. Hadidi, J. Cao, M. Woodward, M. S. Ryoo, and H. Kim, "Distributed perception by collaborative robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3709–3716, 2018.
9. Hu and B. Krishnamachari, "Fast and accurate streaming cnn inference via communication compression on the edge," in *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 2020, pp. 157–163.
10. J. Mao, X. Chen, K. W. Nixon, C. Krieger, and Y. Chen, "Modnn: Local distributed mobile computing system for deep neural network," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. IEEE, 2017, pp. 1396–1401.
11. J. Mao, Z. Yang, W. Wen, C. Wu, L. Song, K. W. Nixon, X. Chen, H. Li, and Y. Chen, "Mednn: A distributed mobile system with enhanced partition and deployment for large-scale dnns," in *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2017, pp. 751–756.
12. T. Odetola, F. Khalid, T. Sandefur, H. Mohammed, and S. R. Hasan, "Feshi: Feature map based stealthy hardware intrinsic attack," *IEEE Access, 2021 (in press) DOI: 10.1109/ACCESS.2021.3104520*, 2021.
13. A. Adeyemo, F. Khalid, T. Odetola, and S. R. Hasan, "Security analysis of capsule network inference using horizontal collaboration," in *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2021, pp. 1074–1077.
14. T. A. Odetola, K. M. Groves, and S. R. Hasan, "2l-3w: 2-level 3-way hardware-software co-verification for the mapping of deep learning architecture (dla) onto fpga boards," *arXiv preprint arXiv:1911.05944*, 2019.
15. Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.
16. G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.