Side-Channel Extraction of Dataflow AI Accelerator Hardware Parameters

Guillaume Lomet*, Rubén Salvador†, Brice Colombier‡, Vincent Grosso‡, Olivier Sentieys*, Cédric Killian‡
* Univ Rennes, Inria, IRISA, Rennes, France
† CentraleSupélec, Inria, CNRS, IRISA, Rennes, France
‡ Université Jean Monnet Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516 F-42023, SAINT-ETIENNE, France

***Abstract*—Dataflow neural network accelerators efficiently process AI tasks on FPGAs, with deployment simplified by ready-to-use frameworks and pre-trained models. However, this convenience makes them vulnerable to malicious actors seeking to reverse engineer valuable Intellectual Property (IP) through Side-Channel Attacks (SCA). This paper proposes a methodology to recover the hardware configuration of dataflow accelerators generated with the FINN framework. Through unsupervised dimensionality reduction, we reduce the computational overhead compared to the state-of-the-art, enabling lightweight classifiers to recover both folding and quantization parameters. We demonstrate an attack phase requiring only 337 ms to recover the hardware parameters with an accuracy of more than 95% and 421 ms to fully recover these parameters with an averaging of 4 traces for a FINN-based accelerator running a CNN, both using a random forest classifier on side-channel traces, even with the accelerator dataflow fully loaded. This approach offers a more realistic attack scenario than existing methods, and compared to SoA attacks based on *tsfresh*, our method requires 940× and 110× less time for preparation and attack phases, respectively, and gives better results even without averaging traces.**

I. Introduction

In recent years, artificial intelligence (AI) powered by deep neural networks (DNN) has become pervasive in various fields, including autonomous systems, image processing, the medical domain, or generative AI. However, these algorithms require substantial computational power and large datasets, driving the development of more powerful and efficient solutions beyond classical computer hardware (HW). One solution is using FPGAs, which can create reconfigurable accelerators optimized for specific tasks and requirements without the need to design a custom chip for each solution.

In several use cases, such as video stream processing or large-scale image batch processing, dataflow accelerators have emerged as one of the most efficient solutions. FPGAs have also become more accessible through remote instances with shared HW and frameworks for end-to-end DNN design, reducing costs and providing ready-to-use development environments. However, DNN accelerators are valuable targets for attacks because of the high value of their intellectual property (IP) and the significant cost associated with creating and training models. Furthermore, remote FPGA instances like cloud servers or embedded systems are vulnerable to physical attacks, which can bypass traditional security measures within these devices [1].

Attacking DNNs using side-channel attacks (SCA) follows a sequence of steps, where each stage requires specific knowledge depending on its position in the attack flow. To penetrate the black box of a DNN accelerator, the first crucial element to understand is the HW architecture, which is the root of all the SCA strategies, from model recovery to stealing inputs. In FPGAs, this architecture is configurable, and users can tune parameters such as the parallelism or the arithmetic used. Although these parameters need to be captured to mount a successful weight recovery through SCAs, most attack models assume the attacker knows the HW architecture.

This work **proposes a methodology to remotely reverse-engineer elements of the HW architecture in a dataflow accelerator implemented on FPGAs.** This is achieved by comparing the traces captured from the accelerator to a dataset of patterns corresponding to traces from accelerators with known architectures. The attack uses unsupervised dimensionality reduction of the traces, and then the transformed data are classified by each HW parameter. Our contributions include:

- To the best of our knowledge, we are the first to **recover multiple HW parameters simultaneously from SCA traces**: the folding, which is the parallelization of data (SIMD) and processing elements (PE), and the quantization used in the accelerator. For this study, we only focus on the **folding of the PE and the quantization**, and with an averaging of 4 traces with random inputs to reduce noise, we can fully recover these parameters.
- Using an **unsupervised dimensionality reduction** technique like principal component analysis (PCA), **we lower the computational cost**, enabling the use of more classifiers for extracting the HW parameters and improving the performance of the parameter recovery.
- Compared to the state-of-the-art (SoA), **our methodology enables a more realistic attack model** and can be used as a **first step to learning the HW architecture**, which was an assumption made by previous attacks.

This paper is organized as follows: Section II discusses the necessary background, including SCA on FPGA NN accelerators. Section III presents the attacker model, while Section IV explains the experimental methodology. Results are presented in Section V before concluding in Section VI.

II. Background

A. DNN accelerators on FPGA

Two framework types are typically used to accelerate DNN computation on FPGA. First, sequential accelerators compute specific operations like tensor multiplications and activations layer-by-layer, similar to CPUs/GPUs (e.g., Vitis AI, TVM-VTA [2], Gemmini [3]). Second, dataflow accelerators process layers in parallel as soon as inputs are available (e.g., FINN [4]). Dataflow accelerators have proven efficient for tasks like large batch or data stream inference [5], often outperforming GPUs or specialized ASICs [6]. However, they require more custom work during DNN model creation, increasing their IP value and thus interest to be recovered through, e.g., SCA.

B. Remote sensors in FPGA for SCA

To deploy FPGA DNN accelerators, remote solutions like in datacenters [7] or embedded systems optimize FPGA usage by sharing resources, reducing costs, and standardizing tools. Remote SCAs are possible via embedded sensors [8], simulating power analysis. While tools protect against some sensors like ring oscillators, others using Time-to-Digital Converters (TDC) [9] or arithmetic unit misuse [10] remain hard to mitigate. Though these sensors are challenging to calibrate remotely, recent studies implement automatic calibration systems, helping attackers get the right sensitivity using TDCs [11], [12].

C. SCA for DNN reverse engineering on FPGA

Valuable data and IP, like the architecture and trained weights from DNN models, can be extracted via SCA [13]. Sensitive inputs, e.g., medical data, can also be recovered using knowledge of the trained model [14]. However, FPGA's reconfigurable HW makes SCAs strongly dependent on knowing the accelerator's architecture. We identify four categories of data and IP recovery attacks through SCA on FPGA, ranging from closest to black-box to closest to white-box scenarios.

1) Hardware architecture: Attacks on the HW architecture include the framework type used [15], [16], data and processing parallelism [17], and the arithmetic type used (number representation, precision). To the best of our knowledge, there is still no work in the last category, which is one of the focuses of this work. Unlike other categories, attacks on the HW architecture are based on reverse engineering using SCA (SCARE) [18] on the circuit, rather than retrieving secrets.

2) Model architecture: These attacks target elements of the neural network (NN) model: number of layers, type of layers, model structure, and layer configuration [15], [19].

3) Model parameters: Several works focus on SCA to extract model parameters, including weights and normalization [16], [19]–[21].

4) Inputs: The last category considers retrieving inputs used at inference time [22] [23].

To get closer to a black-box scenario, attackers need to know the HW architecture of the accelerator, as parallelism can disturb model recovery due to increased noise [17]. Some attacks are linked to the arithmetic used [14], or exploit information from the implementation tool [16]. The closest work [10] to our approach to extracting HW parameters from dataflow accelerators also targets FINN [4]. It extracts the number of parallel PEs in a matrix-vector-accumulation unit to aid in recovering neurons in fully connected layers, using a profiling attack based on power traces and a machine learning classifier. Despite their key role in model parameter extraction, arithmetic parameters have yet to be known during the SCA. This attack [10] also requires computationally heavy statistical analysis, which will be difficult to generalize to other HW parameters. Similarly, EM-based SCA reverse-engineering was used to recover the structure and operation scheduling (input partitioning, channel/pixel-level parallelism, and spatial mapping of input/output channels and HW) of the AMD-Xilinx DPU encrypted DNN commercial accelerator [16].

Conversely, our attack is the first to recover the arithmetic type (number representation, precision) [cat. 1] and to relax the assumption of knowing the HW architecture [cat. 2, 3].

D. Machine Learning based SCA

Machine learning has been shown to improve SCA attacks, e.g., using convolutional neural networks (CNN) for profiling attacks without trace realignment [24]. Another approach reduces trace dimensions by transforming them into smaller vectors for profiling. Early works on cryptographic accelerators have used RNN-based auto-encoders [25], [26] to reduce trace dimensionality, improving noise elimination by preserving key features. However, to recover DNN elements in FPGA accelerators, the primary method used is feature selection [17], [27] using *tsfresh* [28]. While this method is effective and relies on deterministic algorithms, (1) its reliability depends on fixed statistical elements, which can sometimes limit the number of relevant features [29]; and (2) it takes up to three orders of magnitude more time compared to our approach using PCA and lightweight classifiers, as we show in this work.

III. Attacker Model

We focus on a DNN accelerator embedded in an FPGA with remote sensors, aiming to recover key parameters such as the folding strategy and quantization levels used in the victim's model through SCA. Table I summarizes the attacker's knowledge, capabilities, constraints, and results.

The attacker is assumed to have knowledge of the framework used, as well as all the configurable accelerator parameters. Additionally, the attacker can deploy a clock-synchronized sensor on the same FPGA, allowing precise monitoring of the accelerator's activity through power side channels. This synchronized setup provides the attacker with a detailed view of the accelerator's internal operations.

Our method stands out from SoA approaches [17] in several key ways, offering a model that requires less effort from the attacker, making it easier to implement and thus more effective. In our case, the data is processed in batches, meaning that when we capture power traces, the accelerator's dataflow pipeline is already fully loaded. In this scenario, the noise and complexity increase, making it harder to interpret traces directly. Additionally, since we do not have access to the input data, we cannot capture repeated traces of the same data for

TABLE I
COMPARISON OF ATTACKER MODELS IN [17] AND THIS WORK.

Attack Parameter		Our method	[17]
Attacker knowledge	Framework used	Yes	Yes
	Possible configurations of the victim	Yes	Yes
	Victim model	Yes	Yes
Constraints	**CPU activity impact traces**	**Yes**	**No**
	Capture traces in loaded dataflow	**Yes**	**No**
Attacker capabilities	Can implement synchronized sensors	Yes	Yes
	Attacker has access to inputs/outputs	No	No
	Uses batches of inputs	**Yes**	**No**
Results	Recovers the folding	Yes	Yes
	Recovers the quantization	**Yes**	**No**

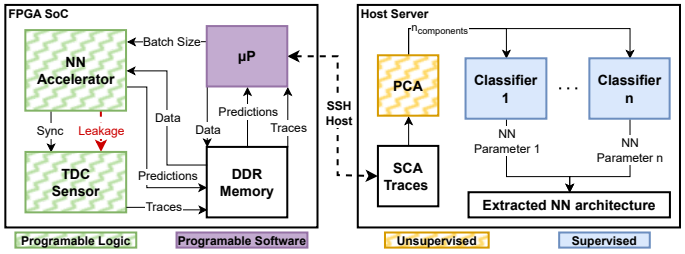


Fig. 1. Overview of the attack system.

averaging but only traces with different inputs; however, we show how this still helps to reduce noise. Lastly, the CPU in the SoC FPGA remains active during our attack running Linux OS and a Jupyter server, introducing additional noise into the traces. This makes our approach more realistic, as it better reflects typical deployments on remote FPGAs where the CPU is not isolated, leading to non ideal and noisier conditions.

Despite these challenges, our method remains effective. We successfully recover the folding and quantization parameters by capturing traces in this noisier, fully loaded dataflow state and using a data-independent approach. This highlights the robustness of our attack in extracting several key model details, even in more complex and realistic FPGA environments.

IV. METHODOLOGY

Our proposed attack methodology targets an implementation of a DNN accelerated on an FPGA, as illustrated in Fig.1, and requires the use of a remote sensor located on the same chip as the accelerator and synchronized with the latter. Through leakage, the sensor will capture the power activity resulting from the calculations of the DNN accelerator.

The proposed method consists of three phases as outlined in Algorithm 1: i) trace acquisition, ii) offline attack preparation, and iii) online attack. Traces captured during trace acquisition by a remote sensor are stored in the FPGA and then transferred via SSH for analysis. The attack involves dimensionality reduction of the traces via PCA and classification to identify the accelerator parameters, with one classifier used for each parameter type. The method is detailed in the following.

A. Trace collection

The acquisition phase consists of two stages: first, creating the labeled database DB for attack preparation, which requires the ability to choose the accelerator parameters; second, performing the attack to recover these parameters without prior knowledge of the configuration. In both phases, the acquired traces are preprocessed as follows:

1) Windowing traces: Traces must contain at least all the samples corresponding to the the slowest stage of the dataflow DNN accelerator. This stage timing can be determined, for instance, through the FINN accelerator generation reports. During dataset construction, we set the minimum number of samples $N_{\text{s\_min}}$ to capture all relevant data from the slowest accelerator considered in our dataset, i.e., the one with less parallelism. This is computed using the slowest stage latency T_{dataflow} and the sensor frequency F_{s} from $N_{\text{s\_min}} = T_{\text{dataflow}} \times F_{\text{s}}$.

2) Removing dataflow loading: Initial samples are skipped to ensure the dataflow is fully loaded. In a remote context, we consider that when triggering a blind measurement, it is unlikely that this would start at the exact same time as the accelerator with an empty dataflow. During the dataflow's loading phase, there is less noise, which would facilitate SCAs. Thus, in our case, our attack remains effective even in a more complex scenario where each stage of the dataflow is active simultaneously and loaded with different data. This is computed using the latency of one complete inference T_{inf} of the slowest accelerator with $N_{\text{load}} = F_{\text{s}} \times T_{\text{inf}}$. The two previous steps are referred to as *Trim* the trace in Algorithm 1.

3) Averaging traces: Averaging traces by a factor n_{average} helps reduce noise, such as that from the CPU on SoC FPGAs. Since our attack model does not allow for input control, averaging is done between traces with different random inputs from the same victim. This parameter must be determined during the offline phase, as explained below.

4) Normalizing traces: To remove the artifacts due to the automatic sensor calibration, we remove the mean of each trace:

$$x_{\text{normalized}} = x_i - \overline{X}, \quad \forall i \in \{1, 2, \ldots, n\}$$

where $x_{\text{normalized}}$ represents the normalized value of the trace, x_i is the i-th time sample in the trace and $\overline{X}$ is the mean of all samples in the trace

B. Features extraction and classification

The core idea of our attack methodology consists of two parts, as detailed in Algorithm 1: i) Dimensionality reduction of traces using feature extraction with PCA, ii) Classification of the extracted features using the dataset labels.

The PCA transformation is an unsupervised process that reduces the dimensionality of the traces while amplifying the variability between traces corresponding to different DNN parameter configurations. PCA computation, which transforms the trace samples into n_{comp} (number of components), requires a fitting step performed during the preparation phase. It is described next.

1) Offline phase (preparation): The PCA is computed *(fit)* on a part of the database $\text{DB}_{\text{training}}$ and produces *(transform)* a dimension reduction from $N_{\text{s\_min}}$ samples to n_{comp} PCA components, which must be selected according to the classifier performance. Regarding the selected classifier type, we tune it through a grid search to find the best combination of n_{comp} and classifier parameters. For each combination, we measure the accuracy of the prediction thanks to DB. Finally, we perform

a validation of the best combination as follows using a PCA reduction (transform) on DB_{test} followed by a classification. We keep the combination of the results with the best accuracy.

2) Online phase (attack): To perform the attack, we acquire one or several traces, followed by preprocessing and PCA transformation. The n_{comp} are then passed to the classifier, which outputs the most probable accelerator configuration.

Algorithm 1: Acquisition and Attack Phases

Output: $DB_{training}$ and DB_{test}
Acquisitions
  for *all possible NN accelerator configurations* **do**
    repeat
      capture a side-channel trace;
      label the trace;
    until N_{traces};
  Split DB into $DB_{training}$ and DB_{test};
Offline phase (preparation)
  Input: $DB_{training}$, DB_{test}, n_{comp_max}, $n_{average}$
  Output: Configuration with the best test accuracy
  Common preprocessing (trimming + averaging + PCA)
    Trim the trace to the Normalized Window Size;
    Normalize traces;
    if *averaging is required* **then**
      average over $n_{average}$ traces;
    Apply PCA on $DB_{training}$ and keep the top n_{comp_max} components;
  Classifier tuning
    for $n_{comp} \in [1, n_{comp\_max}]$ **do**
      Keep n_{comp} components;
      for *selected classifier configurations (grid search)* **do**
        Tune the classifier on $DB_{training}$ with the selected configuration;
        Save $(n_{comp}, \text{config\_classifier}, \text{accuracy})$;
  Classifier validation
    Apply PCA on DB_{test};
    for *classifier configurations with best accuracy* **do**
      Classify DB_{test};
Online phase (attack)
  Input: config_classifier, n_{comp}, $n_{average}$
  Output: Most probable NN accelerator configuration
  for *each parameter of the NN accelerator* **do**
    Capture a new side-channel trace (or more);
    Apply preprocessing with n_{comp} and $n_{average}$;
    Perform classification with config_classifier;

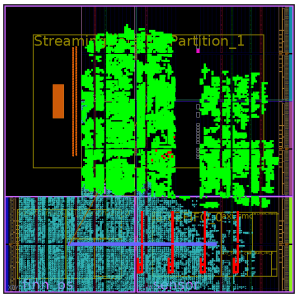
V. Experimental Results

A. Experimental Setup

We used a PYNQ-Z2 FPGA board with a Xilinx Z7020 SoC containing an Artix-7 FPGA and a dual-core Arm Cortex-A9 CPU, both sharing 512 MB of DDR3 RAM and running Ubuntu OS with a Jupyter Notebook server. The considered DNN model is a 5-layer CNN trained on the MNIST dataset (28x28-pixel monochromatic images) to recognize handwritten digits, as detailed in Fig. 2(a). The associated victim accelerator is based on FINN (version 0.9) with the following HW configuration: no DSP, same folding and quantization size for all layers, and all parameters stored on embedded cache (LUTRAM and BRAM) at synthesis. Only the Direct Memory Access (DMA) used to push data in the input and to pull the data from the output of the accelerators communicates with the DRAM and the CPU. The designed accelerator parameters are: folding

ayer type		Data shape
Image	Input-0	[28, 28]
CONV-1	Conv-1	[16, 24, 24]
	ReLu-2	[16, 24, 24]
	Pool-3	[16, 12, 12]
CONV-2	Conv-4	[16, 8, 8]
	ReLu-5	[16, 8, 8]
	Pool-6	[16, 4, 4]
CONV-3	Conv-7	[16, 2, 2]
	ReLu-8	[16, 2, 2]
	Pool-9	[16, 1, 1]
FC-1	Lin-10	[16]
	ReLu-11	[16]
FC-2	Lin-12	[10]

(a)



(b)

Fig. 2. Overview of the experimental setup. (a) CNN victim model, (b) FPGA implementation of FINN victim accelerator: accelerator in green, TDC sensors in red, communication components in light blue.

1x, 2x, 4x, 8x and quantization 4-bits, 6-bits for a total of 8 different accelerator versions.

For the sensors, we use a modified version of the TDC from SCABox [12] with a length of 128 latches. Both the accelerator and TDC sensors use a clock frequency of 100 MHz. To capture the TDC sensor output, we added a DMA module to store the captured traces in the external DRAM memory. The DMA ensures synchronized capture of traces with the accelerator when it starts. Fig. 2(b) shows the complete system floorplan.

B. Data Acquisition

Following the methodology from Section IV, through trace observations and FINN reports, we estimate the minimal length for the sample window around 129k samples. Also, as explained in Section IV-A, we trim the dataflow loading phase from the trace, corresponding to $129\text{k} \times 2 = 258\text{k}$ samples.

All the traces are collected on the FPGA and then sent through SSH after each acquisition. For our experiments, $N_{\text{traces}} = 800$ traces are collected per accelerator configuration to have enough data for our dataset. The final dataset of traces includes a total of N_{total} traces with

$$N_{\text{total}} = N_{\text{traces}} \times N_{\text{folding}} \times N_{\text{quantization}}$$

where N_{folding} and $N_{\text{quantization}}$ are the number of possible folding and quantization, respectively. In our case study $N_{\text{folding}} = 4$, $N_{\text{quantization}} = 2$, hence a total of 6400 traces for DB. We use 5120 traces for $DB_{training}$ and 1280 traces DB_{test}, balanced within all the configurations.

Fig. 3 shows an example trace with 6-bit quantization and folding factors of 1 and 8. To improve visibility in the figure, we reduced noise by averaging 800 traces. However, in the attack phase, averaging is limited to 1 to 4 traces in our case study. Without this averaging step the difference between the cases is not visible at first sight. These traces reveal a repeating pattern, with the pattern frequency proportional to the folding factor. This clearly indicates that the DNN accelerator's power consumption leaks information through side channels.

Trace capture is triggered when the accelerator is launched, so the initial samples capture the NN's activity with an empty dataflow. This loading phase, labelled as *Dataflow Loading* in Fig. 3, has a duration inversely proportional to the folding factor and corresponds to the time taken for a single inference on the same data. As detailed in Section IV, remote SCAs are

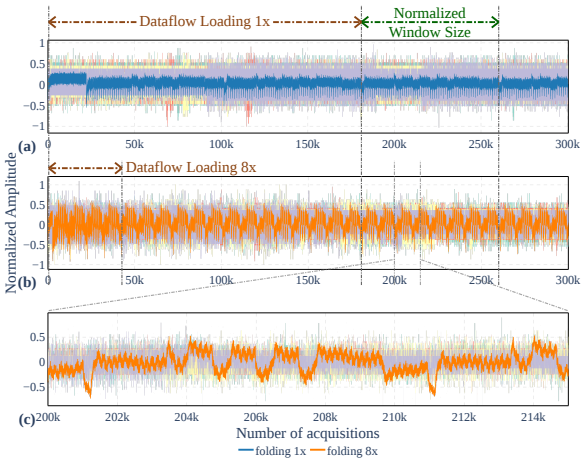


Fig. 3. Average of 800 traces for 6-bit quantization and folding 1× (a) and 8× (b) and (c). Light-coloured traces represent raw traces without averaging.

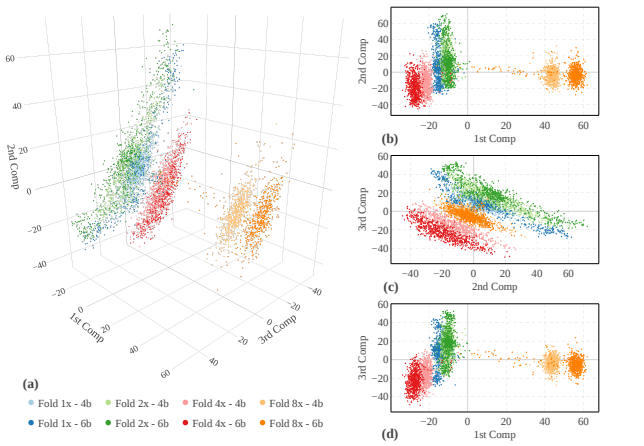


Fig. 4. First three PCA components. (a) 3D view; projection of components (b) 1 and 2, (c) 2 and 3, (d) 1 and 3. Each point represents a trace.

unlikely to capture this event, so the dataflow loading regions are excluded from the acquired samples in the experiments.

Fig. 3 highlights the *Normalized Window Size*, corresponding to the duration of the slowest stage of the dataflow in the slowest accelerator configuration (here 1× folding, no parallelism). During this period, other dataflow stages continue operating with different data, allowing us to gain insights into the activity of the entire NN accelerator. Our method focuses solely on this duration, which is much shorter than the full inference on the same data. This can be seen in Fig. 3 by comparing the Dataflow Loading duration (time for full inference) to the Normalized Size Window (time when all dataflow stages execute simultaneously but on different data). During an attack, as the exact folding value is unknown, we consistently capture the Normalized Size Window. As a side effect, for the same number of acquired samples, we capture more pattern repetitions compared to the lowest folding configuration.

C. Dataset Preprocessing

We apply preprocessing as described in Section IV, discarding the dataflow loading phase and retaining only the samples from the Normalized Size Window, followed by normalization. Next, we fit the PCA using the preprocessed training dataset, employing *scikit-learn* PCA function with default parameters, except for n_{comp}. After fitting the PCA, we transform ($\text{DB}_{\text{training}}$) and ($\text{DB}_{\text{test}}$). This PCA transformation reduces the trace dimensionality while amplifying the variance between traces corresponding to different NN parameter configurations.

We plotted the PCA transformation results of the $\text{DB}_{\text{training}}$ in Fig. 4. A 3D representation (Fig. 4(a)) is provided alongside 2D projections (Fig. 4(b–d)) for clearer point projections. Each point represents a 129k-sample trace transformed into 3 values, known as components. While more components can be used (see Section V-D), their graphical representation becomes more complex. Points are coloured per accelerator configuration, using equivalent light and dark colours for 4-bit and 6-bit quantization, respectively. In Fig. 4, traces from the same accelerator configuration are clustered together. Dark clusters, separated by folding values, show PCA's ability to detect significant variability among traces for each folding. This distinct clustering will enable the association of a new trace from a blind attack with a cluster through a classifier to recover the corresponding folding. However, this separation is less pronounced for lower folding values, likely due to greater amplitude variation in traces at higher folding. Nevertheless, these clusters remain separable by a classifier.

To compare with SoA folding extraction [17], we adapt our methodology to use *tsfresh* instead of PCA. We follow a similar method, with feature extraction on the full dataset, and a selection of the top 10 features from our traces, as in [17].

D. Classifier Exploration

After dimension reduction with PCA, our attack is compatible with any supervised classifier. The effectiveness of our method is emphasized by the simplicity of the classifiers we can use. The PCA preprocessing step efficiently transforms the data, separating distinct traces while grouping similar ones. We explore two computationally simple classifiers: k-Nearest neighbors (k-NN) and Random Forest (RF). Table II shows the computation time for each main step of the methodology for both classifiers during the preparation and attack phases, excluding acquisition time. The computations rely on a server with a dual AMD EPYC 7443 24-core CPU at 4 GHz and 512 GB DDR4 3200 MT/s, running CentOS 7. DB preparation took 15.7 seconds for the k-NN and 17.7 seconds for the RF classifiers when using the PCA, whereas using *tsfresh* took around 4.62 hours, which represents more than 940× compared to the PCA. The time to run the attack on a single trace was 337ms for k-NN and 421ms for RF using the PCA, where it takes more than 110s to attack one trace with *tsfresh*. This demonstrates that our methodology is less computationally intensive, thanks to the efficiency of PCA dimensionality reduction. This also opens the possibility to improve the results, e.g., using an average factor of 4 with a small cost in time, as also shown in Table II.

In the following, we continue to apply our methodology, as presented in Section IV. Both classifiers will be fed with $\text{DB}_{\text{training}}$ for tuning and validated with DB_{test}, through a grid search process. As mentioned in Section IV, we use multiple

TABLE II
TIME REQUIRED PER STEP OF OUR ATTACK METHODOLOGY.

		DB Build	Norm.	Extract.	Folding	Quant.	Total
Preparation on DB [s]	K-NN (tsfresh)	7.5	1.7	16650	0.85	0.95	16661
	K-NN (PCA)	7.5	1.7	6.5	0.85	0.85	17.4
	K-NN (PCA-A4)	7.5	1.7	100	0.85	0.85	110.9
	RF (tsfresh)	7.5	1.7	16650	0.95	0.95	16661.1
	RF (PCA)	7.5	1.7	8.5	1.0	1.0	19.7
	RF (PCA-A4)	7.5	1.7	80	1.0	1.0	91.2
1 trace attack [s]	K-NN (tsfresh)	-	0.003	110	0.25	0.25	110.5
	K-NN (PCA)	-	0.003	0.28	0.027	0.027	0.337
	K-NN (PCA-A4)	-	0.003	0.64	0.03	0.03	0.703
	RF (tsfresh)	-	0.003	110	0.75	0.75	111.5
	RF (PCA)	-	0.003	0.29	0.06	0.07	0.421
	RF (PCA-A4)	-	0.003	0.66	0.06	0.07	0.79

lightweight classifiers to recover parameters. One classifier recovers the folding configuration, while another handles quantization. This approach offers better accuracy than a single, larger classifier trying to distinguish all classes simultaneously.

1) k-nearest neighbors (k-NN): The first classification approach employs k-NN classifiers, which are deterministic and computationally lightweight. Classification requires a preprocessed new trace, a labeled DB for comparison, and a k number of neighbors for similarity-based classification. Only k and n_{comp} need to be explored during the grid search. We limit grid search exploration to $n_{\text{comp\_max}} = 50$ and explore k from 3 to 19. Based on the grid search results, we set the attack parameters to n_{comp} = 12 and k = 13.

We evaluate the final accuracy of our attack using DB_{test}, first applying the PCA and then using the classifiers. The results are shown in Table III, which illustrates the classification accuracy for folding and quantization across different values of n_{average}. For folding, our classifier achieves 91.84% accuracy without averaging and 96.78% accuracy with averaging. This demonstrates the effectiveness of PCA and shows that the NN accelerator leaks sufficient information, even with a limited number of samples relative to the Normalized Window Size and a simple k-NN classifier. Importantly, our method is data-independent, as evidenced by the impact of averaging independent traces, i.e., traces obtained using different inputs.

For better understanding, Table III shows how the classifier determines the correct implementation parameters from a trace. For instance, when Folding 1× is combined with 4 bits of quantization, the k-NN dedicated to quantization outputs the correct result 65.9% of the time. This indicates that classification is more challenging for quantization with Folding 1× and 2×. This can be explained by the fact that the clusters for these folding values are less distinct in the PCA space, as mentioned in Section V-C and visible in Fig 4. However, with the PCA, the recovery is globally better than using *tsfresh* with 90.91%, both having some difficulties in some cases. To improve these results, a stronger classifier could be used such as RF.

2) Random Forest: Unlike k-NN, which may struggle with closely clustered data points, RF should be more effective by leveraging an ensemble of decision trees to capture complex decision boundaries, enhancing prediction accuracy, even when the data clusters are not visually separable. We limit grid search exploration with $n_{\text{comp\_max}} = 50$. The explored parameters are $n_{\text{estimators}}$ in $\{100, 200, 400\}$, and $min_{\text{samples\_split}}$ in $\{2, 5, 10\}$. The best accuracy is found for the parameters: $n_{\text{comp}} = 40$, $n_{\text{estimators}} = 400$ and $min_{\text{samples\_split}} = 5$. Table III shows significant improvements over k-NN, achieving 99.7% accuracy for folding and over 93.7% for quantization without any averaging, and 99.96% for both elements when using an averaging of 4 traces (PCA-A4). Regarding the detail of parameter recovery with respect to the implemented DNN, Table III highlights that even for Folding 1× and 4-bit quantization, the classification accuracy exceeds 96%, confirming that RF can efficiently identify closely clustered data points while maintaining limited computation time.

TABLE III
CLASSIFIER ACCURACY VS. HW PARAMETERS AND LABEL.

Accelerator parameters	Quant.	4 bits				6 bits				**Avg**
	Fold.	1x	2x	4x	8x	1x	2x	4x	8x	
K-NN (tsfresh) accuracy [%]	Quant.	100	84.1	99.3	**100**	94.1	99.2	88.3	97.6	90.91
	Fold.	91.6	70.2	96.7	99.4	80	96.7	75.4	81.9	
K-NN (PCA) accuracy [%]	Quant.	65.9	85.7	89.3	98.7	71.9	77.8	99.3	96.9	91.84
	Fold.	93.5	98.7	**100**	**100**	99.3	92.4	**100**	**100**	
K-NN (PCA-A4) accuracy [%]	Quant.	84.6	92.4	100	100	91.7	79.8	100	100	96.78
	Fold.	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	
RF (tsfresh) accuracy [%]	Quant.	**100**	90	**100**	99.4	96.4	**100**	97	98.8	95.72
	Fold.	98	84.7	94.5	98.1	92.3	99.2	86.1	97	
RF (PCA) accuracy [%]	Quant.	82.7	90.3	94.6	96.3	91.5	98.6	98.1	97.5	96.70
	Fold.	**100**	98.2	**100**	**100**	**100**	99.4	**100**	**100**	
RF (PCA-A4) accuracy [%]	Quant.	**100**	**99.4**	**100**	**100**	**100**	**100**	**100**	**100**	**99.96**
	Fold.	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	

VI. CONCLUSION

This paper proposes an SCA attack methodology for FPGA-based DNN accelerators using remote sensors. Our approach leverages PCA dimensionality reduction and demonstrates its efficiency in enabling the use of lightweight classifiers for parameter recovery. Our attack model is more robust and coherent in a remote context compared to SoA solutions and is data-independent. Experiments show the attack phase requires only 400 ms to recover over 95% of folding and quantization parameters of a FINN-based CNN accelerator using an RF classifier. This is achieved by acquiring one trace with a duration equal to the slowest dataflow stage, validating our methodology. Compared to SoA attacks based on *tsfresh*, our method requires 940× and 110× less time, for preparation and attack phases, respectively, and gives better results even without averaging traces. By leveraging this time gain, we could use an average of 4 traces to fully recover all the HW configurations tested within only 800 ms for the attack phase. Lastly, the core of our method relies on PCA, an unsupervised process that allows for the use of unsupervised classifiers to distinguish between different implementations, though recovering unknown parameters may require additional effort.

ACKNOWLEDGMENTS

This work is partially funded by the French *Agence Nationale de la Recherche* (ANR) Young Researchers (JCJC) program, under grant number ANR-21-CE39-0018 (project ATTILA).

REFERENCES

[1] M. Stojilović, K. Rasmussen, F. Regazzoni, M. B. Tahoori, and R. Tessier, "A Visionary Look at the Security of Reconfigurable Cloud Computing," *Proceedings of the IEEE*, pp. 1–24, 2023.

[2] T. Moreau, T. Chen, L. Vega, J. Roesch, E. Yan, L. Zheng, J. Fromm, Z. Jiang, L. Ceze, C. Guestrin, and A. Krishnamurthy, "A Hardware-Software Blueprint for Flexible Deep Learning Specialization," *CoRR*, vol. abs/1807.04188, 2018.

[3] H. Genc, S. Kim, A. Amid, A. Haj-Ali, V. Iyer, P. Prakash, J. Zhao, D. Grubb, H. Liew, H. Mao, A. Ou, C. Schmidt, S. Steffl, J. Wright, I. Stoica, J. Ragan-Kelley, K. Asanovic, B. Nikolic, and Y. S. Shao, "Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration," in *Proceedings of the 58th Annual Design Automation Conference (DAC)*, 2021.

[4] Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "FINN: A Framework for Fast, Scalable Binarized Neural Network Inference," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2016.

[5] S. I. Venieris, A. Kouris, and C.-S. Bouganis, "Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions," Mar. 2018, arXiv:1803.05900 [cs].

[6] F. Hamanaka, T. Odan, K. Kise, and T. V. Chu, "An Exploration of State-of-the-Art Automation Frameworks for FPGA-Based DNN Acceleration," *IEEE Access*, vol. 11, pp. 5701–5713, 2023.

[7] "Instances F1 Amazon EC2." [Online]. Available: https://aws.amazon.com/fr/ec2/instance-types/f1/

[8] O. Glamocanin, L. Coulon, F. Regazzoni, and M. Stojilović, "Are cloud FPGAs really vulnerable to power analysis attacks?" in *Proceedings of the 23rd Conference on Design, Automation and Test in Europe (DATE)*, San Jose, CA, USA, Mar. 2020, pp. 1007–1010.

[9] F. Schellenberg, D. R. Gnad, A. Moradi, and M. B. Tahoori, "An inside job: Remote power analysis attacks on FPGAs," in *2018 Design, Automation & Test in Europe Conference Exhibition (DATE)*, Mar. 2018, pp. 1111–1116, iSSN: 1558-1101.

[10] V. Meyers, D. R. E. Gnad, N. M. Dang, F. Schellenberg, A. Moradi, and M. B. Tahoori, "Stealthy Logic Misuse for Power Analysis Attacks in Multi-Tenant FPGAs (Extended Version)," 2023, report Number: 935.

[11] B. Udugama, D. Jayasinghe, H. Saadat, A. Ignjatovic, and S. Parameswaran, "VITI: A Tiny Self-Calibrating Sensor for Power-Variation Measurement in FPGAs," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 657–678, 2022.

[12] J. Gravellier, J.-M. Dutertre, Y. Teglia, P. Loubet-Moundi, and O. Francis, "Remote Side-Channel Attacks on Heterogeneous SoC," in *18th International Conference on Smart Card Research and Advanced Applications (CARDIS)*, Prague, Czech Republic, Nov. 2019.

[13] L. Batina, S. Bhasin, D. Jap, and S. Picek, "CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 515–532, qID: Q115407770.

[14] S. Maji, U. Banerjee, and A. P. Chandrakasan, "Leaky Nets: Recovering Embedded Neural Network Models and Inputs through Simple Power and Timing Side-Channels – Attacks and Defenses," *IEEE Internet of Things Journal*, 2021.

[15] S. Tian, S. Moini, A. Wolnikowski, D. Holcomb, R. Tessier, and J. Szefer, "Remote Power Attacks on the Versatile Tensor Accelerator in Multi-Tenant FPGAs," in *2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, May 2021, pp. 242–246, iSSN: 2576-2621.

[16] C. Gongye, Y. Luo, X. Xu, and Y. Fei, "Side-Channel-Assisted Reverse-Engineering of Encrypted DNN Hardware Accelerator IP and Attack Surface Exploration," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Oct. 2023, pp. 1–1.

[17] V. Meyers, D. Gnad, and M. Tahoori, "Reverse Engineering Neural Network Folding with Remote FPGA Power Analysis," in *2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. New York City, NY, USA: IEEE, May 2022, pp. 1–10.

[18] C. Clavier, "An Improved SCARE Cryptanalysis Against a Secret A3/A8 GSM Algorithm," in *Information Systems Security*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. McDaniel, and S. K. Gupta, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4812, pp. 143–155, series Title: Lecture Notes in Computer Science. [Online]. Available: https://hal.science/hal-02487036v1/file/2004-049.pdf

[19] V. Yli-Mäyry, A. Ito, N. Homma, S. Bhasin, and D. Jap, "Extraction of Binarized Neural Network Architecture and Secret Parameters Using Side-Channel Information," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2021, pp. 1–5, iSSN: 2158-1525.

[20] H. Weerasena and P. Mishra, "Revealing CNN Architectures via Side-Channel Analysis in Dataflow-based Inference Accelerators," Nov. 2023, arXiv:2311.00579 [cs].

[21] K. Yoshida, M. Shiozaki, S. Okura, T. Kubota, and T. Fujino, "Model Reverse-Engineering Attack against Systolic-Array-Based DNN Accelerator Using Correlation Power Analysis," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E104-A, no. 1, pp. 152–161, Jan. 2021.

[22] L. Wei, B. Luo, Y. Li, Y. Liu, and Q. Xu, "I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators," in *Proceedings of the 34th Annual Computer Security Applications Conference (ACSAC)*. Association for Computing Machinery, Dec. 2018, pp. 393–406.

[23] S. Moini, S. Tian, D. Holcomb, J. Szefer, and R. Tessier, "Remote Power Side-Channel Attacks on BNN Accelerators in FPGAs," in *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Feb. 2021, pp. 1639–1644, iSSN: 1558-1101.

[24] E. Cagli, C. Dumas, and E. Prouff, "Convolutional Neural Networks with Data Augmentation Against Jitter-Based Countermeasures: Profiling Attacks Without Pre-processing," in *Cryptographic Hardware and Embedded Systems – CHES 2017*, 2017, vol. 10529, pp. 45–68.

[25] K. Ramezanpour, P. Ampadu, and W. Diehl, "SCAUL: Power Side-Channel Analysis With Unsupervised Learning," *IEEE Transactions on Computers*, vol. 69, no. 11, pp. 1626–1638, Nov. 2020.

[26] Z. Liu, Z. Wang, and M. Ling, "Side-channel Attack Using Word Embedding and Long Short Term Memories," *Journal of Web Engineering*, Jan. 2022.

[27] Y. Zhang, "Stealing Deep Learning Model Secret through Remote FPGA Side-channel Analysis," Master's thesis, UC Irvine, 2021.

[28] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018.

[29] T. Henderson and B. D. Fulcher, "An Empirical Evaluation of Time-Series Feature Sets," Oct. 2021, arXiv:2110.10914 [cs].