MLMSA: Multilabel Multiside-Channel-Information Enabled Deep Learning Attacks on APUF Variants

Yansong Gao, Member, IEEE, Jianrong Yao, Lihui Pang, Wei Yang, Anmin Fu, Said F. Al-Sarawi, Senior Member, IEEE, and Derek Abbott, Fellow, IEEE

Abstract—To improve the modeling resilience of silicon strong physical unclonable functions (PUFs), in particular, the APUFs that yield a very large number of challenge-response pairs (CRPs), a number of composited APUF variants, such as XOR-APUF, interpose-PUF (IPUF), feed-forward APUF (FF-APUF), and OAX-APUF, have been devised. When examining their security in terms of modeling resilience, utilizing multiple information sources, such as power side channel information (SCI) or/and reliability SCI, given a challenge is under-explored, which poses a challenge to their supposed modeling resilience in practice. Building upon multilabel/head deep learning (DL) model architecture, this work proposes multilabel multiside-channel-information-enabled DL attacks (MLMSAs) to thoroughly evaluate the modeling resilience of aforementioned APUF variants. Despite its simplicity, MLMSA can successfully break large-scaled APUF variants, which has not previously been achieved. More precisely, the MLMSA breaks 128-stage 30-XOR-APUF, (9, 9) and (2, 18)-IPUFs, and (2, 2, 30)-OAX-APUF when CRPs, power SCI, and reliability SCI are concurrently used. It breaks 128-stage 12-XOR-APUF and (2, 2, 9)-OAX-APUF even when only the easy-to-obtain reliability SCI and CRPs are exploited. The 128-stage six-loop FF-APUF and one-loop 20-XOR-FF-APUF can be broken by simultaneously using reliability SCI and CRPs. All these attacks are normally completed within an hour with a standard personal computer. Therefore, MLMSA is a useful technique for evaluating other existing or any emerging strong PUF designs.

Index Terms—Multilabel/output model, multilabel classification, multiside-channel information, physical unclonable function (PUF).

I. INTRODUCTION

PHYSICAL unclonable functions (PUFs) provide hardware instance-specific outputs (known as responses) to queried inputs (known as challenges), thus, challenge-response pairs (CRPs) generally function as “fingerprints” of hardware devices [1], [2], [3]. PUFs can be categorized based on the number of yielded CRPs into weak and strong PUFs [1], [2]. Weak PUFs have a limited number of CRPs which must be protected so that its primary application is volatile key provi[4], [5]. On the other hand, strong PUFs offer a very large number of CRPs, which can be used in many security applications ranging from identification, lightweight authentications to oblivious transfer [2]. Among strong PUFs, the arbiter PUF (APUF) [6], [7], [8] is the most studied design due to its compactness and compatibility with silicon fabrication processes. However, the APUF is vulnerable to modeling attacks due to its linear structure. To increase the complexity of modeling attacks, various nonlinearity injection techniques have been used to construct APUF variants including the representative l-XOR-APUF, (x, y)-IIPUF [9], feed-forward APUF (FF-APUF) [10], and (x, y, z)-OAX-APUF [11]. These APUF variants are resilient to modeling attacks to a large extent given that their scale is increased (i.e., 128-stage or/and a large number of underlying APUFs composited) [12] when accessing high-performance computing platform, e.g., server with a cluster of GPUs and resourceful memory is unavailable. State-of-the-Art of Modeling Attacks: Majority of modeling attacks exploit CRPs only to train the model. By using deep learning (DL) (i.e., multiple layer perception network), purely CRP-based modeling attacks can break 128-stage 7-XOR-APUF, 64-stage (11, 11)-IPUF, 128-stage FF-APUF with 5 loops, 64-stage 6-XOR-FF-APUF with five loops [13]. In addition, it has been shown that the side-channel information (SCI), including unreliability [14], [15], power or timing [16], and photonic emission [17], can be utilized to model the APUF or its variants. However, merely relying on SCI is insufficient to break APUF variants once it is properly scaled, and acquisition of some SCIs, e.g., photonic emission and timing require costly peripheral equipment.

To date, little efforts have been paid to hybrid modeling attacks on strong PUFs, in particular, APUF variants. The two-hybrid attacks that are multiclass/single-label multi-SCI attack (SLMSA) [18], and gradient-based reliability hybrid attack (GRA) [19] use not only CRP but also SCIs concurrently to break the APUF variants at a larger scale.

Manuscript received 19 July 2022; revised 6 November 2022; accepted 10 January 2023. Date of publication 12 January 2023; date of current version 22 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002167, Grant 61901209, and Grant 61802186; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200461; and in part by the Natural Science Foundation of Hunan Providence under Grant BK20200461; and in part by the Natural Science Foundation of China under Grant 62002167, Grant 61901209, and also with Data61, CSIRO, Canberra, ACT 2601, Australia (e-mail: yansong.gao@njust.edu.cn).

Jianrong Yao, Wei Yang, and Anmin Fu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with Data61, CSIRO, Canberra, ACT 2601, Australia (e-mail: jianrong.yao.111@njust.edu.cn; yangwei@njust.edu.cn; fuan@njust.edu.cn).

Lihui Pang is with the School of Electrical Engineering, University of South China, Hengyang 421000, China, and also with the Security Engineering Laboratory, Sungkyunkwan University, Suwon 16419, South Korea (e-mail: sunshine.plh@hotmail.com).

Said F. Al-Sarawi and Derek Abbott are with the School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: saidalsarawi@adelaide.edu.au; derek.abbott@ adelaide.edu.au).

Digital Object Identifier 10.1109/TCAD.2023.3236563
However, there are still limitations in these attempts. More specifically, the SLMSA has a large dimension as its trained model output dimension is a multiplication per SCIs (SCI, including the binary response information). In addition, this work mainly examines the efficiency of CRP as well as power SCI hybrid attacks, but efficacy when easy-to-obtain unreliability SCI is available has not been explored by [18]. The reason is that Liu et al. [18] recognized the dimension of the reliability SCI could be much higher, potentially resulting in a dimensional curse (detailed in Section VI-D). For the GRA specifically devised to attack iPUFs, it requires the differential mathematical model of underlying APUsFs, which is nontrivial to adopt without in-depth knowledge of the model given the APUF variant.

**Our Contributions:** The primary contributions and results of this work are summarized as follows. Significantly, all reported results are achieved with a common personal computer and modeling attacks are completed within an hour even for large-scaled strong APUF variants.

1. We are the first to introduce multi-label/head classification to facilitate multi-SCI DL modeling attack, coined as multilabel multiside-channel-information-enabled DL attacks (MLMSAs) that eliminates the curse of dimensionality in the SLMSA. Specifically, the MLMSA model output dimension is now equal to the dimension summation per SCI rather than a dimension multiplication per SCI in the SLMSA. In contrast to SLMSA that requires mapping from the predicted label to the response, MLMSA directly outputs the response.

2. We have successfully attacked 128-stage 10-XOR-APUF, (2, 2, 8)-OAX-APUF and (5, 5)-iPUF with the MLMSA by simultaneously using the response and easy-to-obtain reliability SCI. Notably, 128-stage 12-XOR-APUF, (2, 2, 9)-OAX-APUF are also breakable statistically, that is, among five repetitions, one attempt succeeds in our experiments. For these attacks, the training size is no more than 600 000 and training completes within an hour. In contrast to GRA, the MLMSA does not require a mathematical model of underlying PUFs. As a comparison, the purely CRP-based DL modeling attacks can break 128-stage 7-XOR-APUF but with a significantly increased training size of 30M [13].

3. We have advanced the breakable APUF variants to an even larger scale, albeit the concise design of the proposed MLMSA. By simultaneously exploiting multiple SCIs, including response, power, and reliability, the MLMSA successfully breaks 30-XOR-APUF, (2, 2, 30)-OAX-APUF, (9, 9)- and (2, 18)-iPUFs, all with 128-stage underlying APUFs.

4. Based on silicon measurements, we have further affirmed the merits of leveraging additional easy-to-obtain reliability information to attack XOR-APUFs compared to the setting of merely using response information. In particular, the response and reliability-based DL can successfully attack 128-stage 10-XOR-APUF with 1.5M challenges corresponding response and reliability pairs, whereas merely response-based DL can only attack a 6-XOR-APUF with the same 128-stage and training set size.

II. BACKGROUND

This section provides the necessary background on APUF and its representative variants that this study examines.

A. Arbiter-Based PUF

The APUF exploits manufacturing variability that results in random interconnect and transistor gate time delays [6]. This structure is simple, compact, and capable of yielding a large CRP space. In contrast to the optical PUF that lacks a mathematical model [20], the APUF has a linear additive structure, leading to vulnerability to modeling attacks. In the modeling attack, an attacker utilizes observed CRPs to build a mathematical PUF model that can accurately predict responses for unseen challenges [15], [21], [22], [23].

**Linear Additive Delay Model:** A linear additive delay model of APUsFs is formulated as [10]

\[
\Delta = w^T \Phi
\]

where \( w \) is the weight vector that characterizes the time delay segments in the APUF, and \( \Phi \) is the parity (or feature) vector that can be generally understood as a transformation of the challenge. The dimension of both \( w \) and \( \Phi \) is \( n + 1 \) given an \( n \)-stage APUF, where

\[
\Phi[n] = 1, \quad \Phi[i] = \prod_{j=i}^{n-1} (1 - 2c[j]), \quad i = 0, \ldots, n - 1.
\]

The response of an \( n \)-stage APUF is determined by the delay difference \( \Delta \) between the top path and bottom path of the APUF. This delay difference is the sum of the delay differences of each individual \( n \) stages. The delay difference of each stage depends on the corresponding challenge [15]. Based on (1), the response \( r \) of the challenge \( c \) is modeled as

\[
r = \begin{cases} 
1, & \text{if } \Delta < 0 \\
0, & \text{otherwise}.
\end{cases}
\]
exponentially increase the modeling attack complexity when only CRP is used. However, the l-XOR-APUF unreliability increases when l is increasing, which negatively restricts the large l usage to some extent. In addition, the large l of a l-XOR-APUF is still ineffective against reliability-based modeling attacks—it uses reliability information of the CRP—since the complexity of such attack is only linearly increased as a function of l.

C. OAX-APUF

As is shown in Fig. 2, the OAX-APUF [11] consists of OR, AND, and XOR blocks. The x APUFs’ responses are OR-ed to get r_{or}, y APUFs are belong to AND block, in which the responses are AND-ed. The XOR block contains z APUFs, whose responses are XOR-ed to gain r_{xor}. The responses of three blocks are XOR-ed to obtain the final response r. According to [11], the OAX-APUF has higher reliability than XOR-APUF, while OAX-APUF can defeat covariance matrix adaptation evolution strategy (CMA-ES)-based reliability attacks and demonstrate comparable modeling resilience to logistics regression attack compared to (x+y+z)-XOR-APUF. However, it has relatively lower resilience than (x+y+z)-XOR-APUF when the DL attack is applied [11].

D. iPUF

The iPUF contains two layers of XOR-APUFs [9], [18]. As shown in Fig. 3, the response of x-XOR-APUF is inserted into the i-th position of the challenge to obtain the challenge of (n+1) bits. The new (n+1)-bit challenge is input into y-XOR-APUF to get the final response r. In theory and experiment, the iPUF has been demonstrated to have desired resistance to LR and reliability-based CMA-ES attacks [9], [18]. According to [24], the security of the (x, y)-iPUF against modeling resilience is similar to a ([x/2] + y)-XOR-APUF when the logistic regression (LR)-based divide-and-conquer attack is applied.

E. Feed Forward Arbiter PUF

The feed-forward arbiter-PUF (FF-APUF) [10] adds one or more intermediate arbiters within a basic APUF, and the output response of the intermediate arbiter replaces one or multiple bits of the challenge. This is a typical design of obfuscating the APUF challenge bit(s). The structure of an FF-APUF with one loop is depicted in Fig. 4 [12]. This FF-APUF can be incorporated with xor or OAX operations when multiple underlying FF-APUFs are used.

III. RELATED WORKS

Modeling attacks on strong PUFs normally rely on machine learning (ML) techniques. The ML attacks against strong PUFs can be divided into three categories according to the type of training data used: 1) CRP-based ML attacks; 2) SCI-based attacks; and 3) SCI hybrid attacks, which use CRPs; SCI; and CRPs along with SCI(s) as training data, respectively.

A. CRP-Based ML Attacks

LR, support vector machine (SVM), and evolution strategies were utilized by Rührmair et al. [21] to model XOR-APUF, FF-APUF, and LSPUF in 2010. There are a number of improvements to increase the attacking accuracy [25]. It is always suggested to increase the scale of the APUF variants, in particular, the XOR-APUF to increase the modeling resilience against those CRP-based modeling attacks. In order to increase the complexity of these ML attacks, more APUF variants building upon various forms of recompositions have been proposed, such as MPUF [26], iPUF [9] and OAX-APUF [11].

According to [9], LR is the most efficient attack against l-XOR-APUF, but it cannot be used to attack the iPUF directly. Wisiol et al. [13] made some improvements to the LR attack and reported that the improved LR attack can break 64-bit 8-XOR-APUF with an accuracy of 96.4% by using up to 150M CRPs (i.e., training time is 391 min using 4 threads). The LR-based divide-and-conquer attack [24] (LDA) was proposed to attack the iPUF, which can successfully break 64-stage (1, 7)-iPUF with an accuracy of 97% [18]. As reported by Liu et al. [18], the LDA attack can successfully break 128-stage (6, 6)-iPUF. According to Wisiol et al. [13], the MLP-based divide-and-conquer attack is able to attack 64-stage (11, 11)-iPUF with 650M CRPs.

More recently, DL has been shown to be a simple and effective way to attack strong PUFs without knowing the underlying mathematical strong PUF model.
Alkatheriri and Zhuang [27] showed that a one-hidden layer MLP attack can successfully model FF-APUF with six loops in 2017. In 2018, Aseeri et al. [28] proposed a three-hidden layer MLP attack, which can successfully model 128-bit 7-XOR-APUF with 40M CRPs according to Wisiol et al. [13]. Santikellur et al. [29] proposed DL attacks on XOR-APUF, MPUF, and iPUF in 2019, which can break 128-stage (4, 4)-iPUF, (128, 5)-rMPUF, and 5-XOR-APUF. It has been also shown that the (64, 6)-rMPUF and (32, 7)-rMPUF are breakable [30]. Note that compared to iPUF, XOR-APUF, and OAX-APUF, it requires greatly increased APUFs and many MUXs (e.g., 2-to-1 MUXs are first decomposed into many 2-to-1 MUXs for implementation). The implementation of a further scaled (64, 7)-rMPUF and (32, 8)-rMPUF requires at least 255 64-stage APUFs and 511 32-stage APUFs, respectively, which high area overhead renders its practicality to a large extent. Mursi et al. [31] proposed a three-hidden layer MLP attack, which mainly focuses on XOR-APUF. According to Wisiol et al. [13], this three-hidden layer MLP [31] can successfully break 128-stage 7-XOR-APUF with 30M fully reliable CRPs.

B. SCI-Based Attacks

SCI-based attacks can be divided into pure side-channel analysis (SCA) attacks and SCA-based ML attacks. The pure SCA attacks can be conducted alone to attack a single APUF, such as reliability-based analysis [14] and the photonic emission attack [17]. The SCA-based ML attacks mainly utilize reliability and power SCIs.

The reliability-based ML attack establishes the reliability model of APUF that exploits the relationship between the response reliability and internal parameters [15]. The measured reliability data and challenges are provided to, e.g., the CMA-ES model, as training data to learn the internal parameters of, e.g., XOR-APUFs. The fault injection can be utilized to accelerate the reliability SCI collection [18], [32]. The reliability-based CMA-ES attack [15] can successfully break XOR-APUF and LSPUF. The CMA-ES attack is based on the assumption that the unreliability contribution per APUF of the l-XOR-APUF is equal so that the CMA-ES can converge to any of l APUFs in an equal chance when the attack repeats. Therefore, the complexity of breaking the l-XOR-APUF is linear in l. The OAX-APUF [11] and iPUF [9] breaks such an assumption, thus, can defeat the CMA-ES-based reliability modeling attacks.

Different power-based ML attacks leverage differing methods for analyzing power leakages, e.g., simple power analysis (SPA) and correlation power analysis (CPA) [18]. Becker and Kumar [33] proposed a CPA-based CMA-ES attack that uses power correlation coefficients as the fitness function to model controlled PUFs and LSPUFs. A SPA-based LR attack was proposed by Rührmair et al. [16], which adopts a gradient-based algorithm similar to LR to learn the power side-channel model of XOR-APUF. However, because the relationship between other APUF variants’ power and response is difficult to deduce, so fewer power-based ML attacks are used to model other APUF variants.

Though there are a number of SCI sources that can be used to attack strong PUFs, the reliability SCI is the most easily obtainable one. To collect power SCI, physical access to the PUF device and some expertise are required. The photonic emission collection is costly and usually requires proficient expertise.

C. SCI Hybrid Attacks

The above two types of ML-based attacks with a CRP-based attack or SCI-based attack only use CRP or SCI, and, thus, knowledge gained by both is not utilized. The other type of SCI hybrid ML attack considers using them concurrently to be more efficient. There are two recent studies on SCI hybrid attacks, exhibiting greatly improved attack efficacy. One is the GRA [19] and the other is the multi-class/SLMSA [18].

The GRA [19] was mainly devised to attack (x, y)-iPUF. In essence, it combines the CMA-ES reliability attack and LR CRP attack together. For the CMA-ES reliability attack term, it learns multiple APUFs concurrently (i.e., regular CMA-ES learn an APUF per run [15]) and enforces that each APUF is dissimilar to others to prevent the APUF converging to those easiest-to-learn APUFs through reliability SCI. The GRA attack requires careful constraints to each attack term, which potentially requires manual settings in practical upon trials. Note that the GRA attack is less effective on (x, y)-iPUF with x > 1. Therefore, a multiple pass attack similar to the iPUF splitting attack [24] has to be adopted. In this context, the y-APUF is first learned, then the x-APUF are learned sequentially. In addition, the GRA requires to construct a differential model for the iPUF, which is nontrivial for adoption as it requires an in-depth understanding of the underlying PUFs under attack.

The multiclass classification-based side-channel hybrid attack (SLMSA) proposed by Liu et al. [18] is the state-of-the-art to attack XOR-APUF and iPUF, which avoids the underlying PUF mathematical models by using DL techniques. The SLMSA combines response and power SCI to validate its efficiency. To transform the hybrid information into multiple classes, where there is only one true value and the rest are false values, we use one-hot vector encoding also referred to as single-label classification, so the SLMSA has to first fuse CRP information with SCI to construct the so-called challenge-synthetic-feature pairs (CSFs) via the feature crossing method of Liu et al. [18]. More specifically, feature crossing uses the Cartesian product of the response r and side-channel information p to form a single label. Therefore, the number of categories of the multiclass classification is substantially increased when the dimension per p and the number of p increases due to the usage of the Cartesian product. This could incur dimension curse as recognized by Liu et al. [18]. For example, the response has two categories, and the l-XOR-APUF has (l + 1) power SCI categories so that the dimension is 2 × (l + 1). In this context, Liu et al. take the CRP and the power SCI into consideration. The reliability SCI is not used. Nonetheless, the SLMSA has been shown to successfully break 128-stage 16-XOR-APUF and (2, 16)-iPUF with 600 000 training CSFs (response and power SCI).
However, this study does not validate the efficacy of the reliability SCI that is the most easily obtainable SCI. In addition, the prediction of the response is not immediately available, which is recovered through a remapping according to the CSP process.

IV. MULTILABEL MULTI-SCI-BASED DEEP LEARNING ATTACKS

We propose a multilabel DL-based attack to efficiently and effectively take advantage of multiple SCIs, coined as MLMSA. First, its output dimension is merely a summation per used SCI. Second, it can directly predict the response of the learned strong PUF without additional remapping. Third, it can allow flexible weight tuning per SCIs and response to gain improved attack accuracy—we, thus, can break 128-stage 12-XOR-APUFs by using response and the easy-to-obtain reliability SCI that is not considered in [18].

A. Multilabel Model for MLMSA

Liu et al. used a single-label model [18] to exploit multiple information, e.g., CRP and power SCI. In the single-label model, a given input instance can only belong to one of more than two classes. This results in the inconvenient CSP synthesis, where the dimension of the output is greatly increased, especially, when multiple SCIs are concurrently exploited—the curse of dimensionality recognized by Liu et al. [18]. We note that the single-label model can be circumvented via the multilabel model. In the multilabel classification, there is no constraint on how many of the classes the instance can be assigned to. For instance, an analogy is that a movie can have multiple classes of comedy, romance, and action in the multilabel output.

The multilabel DL model [34], [35] is also usually referred to as multihead/output DL model. Different from the single-label/head DL model, the output layer of the multihead model has multiple outputs or heads, which each head corresponds to a label (i.e., response or power SCI or reliability SCI). Supposing there are k heads/outputs in the multilabel model, then the loss of the model can be expressed as

$$L = \sum_{i=1}^{k} \lambda_i L_i$$  \hspace{1cm} (4)

where $L_i$ means the loss of the $i$th head, $\lambda_i$ means the weight or the regularized factor of the $L_i$, which can be flexibly tuned to gain optimal performance.

B. MLMSA

As shown in Fig. 5, the multihead model can use CRP and a number of SCIs, thus, enhancing the model for learning the underlying PUFs better by leveraging more useful information sources. Because not only the response but also other SCIs observed for given challenges are all simultaneously used to train the model, providing more meaningful information to model the internal parameters of the underlying strong PUF.

This work focuses on power or/and reliability SCIs. The power consumed by the, e.g., 1-XOR-APUF is linearly proportional to the number of responses being “1” in 1 APUFs. More specifically, reliability SCI is obtained by computing the number of responses of “1”s from $m$ repeated measurements given the same challenge queries. Let us consider ten repeated measurements of a given challenge as an example: the number of responses with “1”s obtained from ten repeated measurements has a value ranging from 0 to 10, and, thus, there are 11 possible values. If the reliability SCI is divided into 11 categories, each integer number stands for one category. The proposed MLMSA attack has three stages:

Pretreatment Stage: Collecting CRPs and the exploited SCI(s). Note that the label of a given SCI needs to be converted to a one-hot vector.

Training Stage: Using multihead model to train the targeted strong PUF model. The input is a challenge. One head predicts the response, and the other head(s) predict(s) the rest SCI(s), respectively, of the given challenge. The difference between the predictions and the ground-truth labels are used to optimize the multihead model, according to (4).

Prediction Stage: Once the multihead model is trained, the response given an unseen challenge can be directly predicted by the response head/output.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

According to the dynamic power analysis of PUFs, the amount of drawn charge is linearly proportional to the number of latches exhibiting a value of “1”s [18]. For PUF designs that employ more than one APUFs in parallel, by measuring the amount of current drawn from the supply voltage during any latch transition, the cumulative number of APUFs that respond with “1”s can be determined [18], [36]. In other words, the power consumption is linear with the number of “1” responses produced by APUFs. This has been validated by the consistency between physical measurements and simulations [16], [33], [37]. Following [18], we adopt the counted “1”s as the power SCI. As for reliability information, we apply the same challenge repeatedly many times and classify the reliability information according to the number of responses “1”s obtained by repeated measurements. More specifically, if there are ten repeated measurements, the number of categories of reliability SCI is 11 (i.e., from 0 to 10).

Following [9], [11], [21], [23], [24], we use MATLAB to numerically simulate CRPs, power SCI and reliability SCI...
required by the following experiments—silicon measurement validations of reliability SCI are detailed in Section VI-C. Each APUF is a 128-stage—majority of previous studies using 64-stage APFUs. For the response and power SCI, most experiments use noise-free simulation, in which \( \mu = 0, \sigma = 1 \) are used to generate the weights corresponding to the APFUs as in (1). In this context, we collect training/testing CRPs. The unreliability is produced by injecting the Gaussian noise into the above-generated weights by setting, \( \mu_{\text{noise}} = 0 \) and \( \sigma_{\text{noise}} = 0.05 \) to get the noisy weights per repeated same challenge query. The unreliability of APFUs ranges from 0.05 to 0.08 after noise injection. The reliability SCI consequentially can be collected. For the power SCI, we count the number of “1”s in simulated APFUs as the SCI.

The training set, validation set, and test set are divided according to the ratio of 4:1:1. It should be noted that if CRPs used for testing and CRPs used for training collected under different conditions (e.g., enrolled at 25 °C but regenerated at 50 °C), testing accuracy is expected to be degraded.

For FF-APUF, we have only considered the combination of response and reliability, the responses of the training set and validation set are obtained by majority voting, and the response of the testing set is noise-free. For the XOR-APUF, iPUF, and OAX-APUF, we have considered the hybrid of response, power, and reliability.

The number of hidden layers of the multihead model as exemplified in Fig. 5 in each experiment is 3 or 4, and the activation function is ReLU. Liu et al. [18] used 2 or 3 hidden layers, which breaks 16-XOR-APUF. Our reproduced results of Liu et al. successfully attack 30-XOR-APUF, which can be potentially attributed to the adopted DL architecture with more hidden layers that are optimal values after hyperparameter tuning. For the response head, the loss function uses binary_crossentropy, while the loss function of other heads uses categorical_crossentropy. The Adam optimizer is used for all experiments. The settings of different head loss weights of MLMSA are summarized in Table I. All experiments are completed using a common personal computer with an Intel Core i5-6200U CPU, and 12-GB memory.

### B. Modeling Attacks and Results Analysis

The XOR-APUF, OAX-APUF, iPUF, and FF-APUF are used to validate the effectiveness of the proposed MLMSA. Note that the SLMSA is the most efficient attack using not only CRPs but also SCI (in particular, power SCI). We compare the results with SLMSA by reproducing it under the same experimental settings for fair comparisons. Table II summarizes the main results of the MLMSA and SLMSA attacks on three strong APUF variants. Notably, Multiclass A and Multiclass B belong to the SLMSA attacks [18], where different SCIs are used.

1) **Multiclass A**: The CSPs are formed with power SCI and CRPs.

2) **Multiclass B**: The CSPs are formed with reliability SCI and CRPs.

Note that Multiclass B is not considered in [18] for experimental evaluations, which we explore, for the first time, for comparison purposes.

1) **l-XOR-APUF**: The loss weight settings of multiclass model are described in Table I. As for the Two-Head A model (response head and rower head), the response loss weight is 10, the power loss weight is 2. As for the Two-Head B model (response head and reliability head), the response loss weight is 1 when \( l = 10; 0.8 \) in other cases. In Three-Head, the reliability loss weight is 1 when \( l = 29, 30; 2 \) in other cases.

   **Multiclass A**: For \( l \)-XOR-APUF, the reliability head loss weight is 1.8 when \( l \geq 16 \); 0.8 when \( l < 16 \). The training size is 300 000 when \( l \leq 12 \); 600 000 when \( l \geq 16 \). The training size is 300 000 when \( l \leq 12 \); 600 000 when \( l \geq 16 \), and \( l \) is small.

   **Multiclass B**: When using only response and reliability SCI, the two attacks can reliably break 10-XOR-APUF with accuracy more than 95%—later we show 12-XOR-APUF is statistically breakable under multiple repeated attacks. The larger \( l \), the harder to minimize reliability loss during the training optimization. The Three-Head attack using response, power, and reliability SCI exhibits an improvement over the two-head model that uses response and power SCI only when \( l \) is small. This means the power SCI is more efficient than reliability SCI for attacks. The accuracy of Two-Head A, Three-Head, and Multiclass A are similar. This further indicates the dominant
TABLE II
COMPARISONS OF MLMSA WITH SLMSA (I.E., MULTICLASS) [18] AND DL WITH PURE CRP ATTACK [29] ON COMPLICATED STRONG PUFs

| L-XOR-APUF | Training CRPs | Two-Head A | Two-Head B | Multi-Class A [18] | Multi-Class B [18] | DL [29] |
|------------|---------------|------------|------------|--------------------|--------------------|--------|
| 5-XOR-APUF | 300,000 (600,000 / 655,360) | 96.77% | 97.45% | 98.35% | 98.51% | 98.27% | 97.82% |
| 6-XOR-APUF | 300,000 (600,000 / 1,200,000) | 96.84% | 97.38% | 98.09% | 98.34% | 97.54% | 94.98% |
| 10-XOR-APUF | 300,000 (600,000 / 1,200,000) | 95.48% | 97.81% | 97.11% | 96.14% | 96.35% | / |
| 39-XOR-APUF | 600,000 | 91.51% | 89.77% | 89.13% | / | / | / |
| (1, 2, 3)-OAX-APUF | 300,000 (600,000) | 97.46% | 97.49% | 97.00% | 95.36% | 97.13% | 97.75% |
| (2, 2)-OAX-APUF | 300,000 (600,000) | 97.76% | 97.82% | 97.37% | 97.95% | 98.81% | 90.65% |
| (3, 3)-OAX-APUF | 300,000 (600,000) | 96.17% | 95.00% | 96.15% | 99.30% | 99.84% | / |
| (4, 4)-OAX-APUF | 600,000 | 97.21% | 98.52% | 98.97% | / | / | / |
| (5)-APUF | 400,000 (1,200,000) | 94.64% | 97.59% | 97.02% | 95.73% | 96.16% |
| (6, 6)-APUF | 600,000 | 95.79% | 96.33% | 96.13% | / | / | / |
| (7, 7)-APUF | 400,000 | 95.29% | 95.72% | 95.47% | / | / | / |
| (8)-APUF | 600,000 | 95.11% | 95.33% | 95.82% | / | / | / |
| (9, 9)-APUF | 600,000 | 95.25% | 96.59% | 97.03% | / | / | / |

* Two-Head A of MLMSA uses response and power SCI, Two-Head B of MLMSA uses response and reliability SCI, Three-Head of MLMSA uses response, power SCI, and reliability SCI. Multi-Class A of SLMSA uses response and power SCI, Multi-Class B of SLMSA uses response and reliability SCI.

** For Two-Head A, Three-Head and Multi-Class A, the training size is 300,000 or 600,000. Take L-XOR-APUF as an example, the size of 300,000 is used when the attacks use power SCI 600,000 when attacks use reliability SCI. The 655,000 is the number given by [29].

Fig. 6. Comparisons of MLMSA (i.e., multihead) and SLMSA (i.e., multiclass) attacks using CRP, or/and power or/and reliability SCI on L-XOR-APUFs. The x-axis stands for the L of XOR-APUFs.

Fig. 7. Comparisons of MLMSA (i.e., multihead) and SLMSA (i.e., multiclass) attacks on (x, y)-iPUFs. The x-axis stands for the (x, y) of iPUFs.

...by existing works include [18]. As for Two-Head B by using response and reliability SCI, both the MLMSA and SLMSA can break (5, 5)-iPUF with an accuracy more than 95%.

3) \((x, y, z)\)-OAX-APUF: For the \((x, y, z)\)-OAX-APUF, we fix \(x = 2\), \(y = 2\), and change the setting of \(z\). The loss weight settings of multihed attacks are detailed in Table I. As for the Two-Head A model (response head and power head), the response head loss weight is 10, the power head loss weight is 2. As for the Two-Head B model (response head and reliability head), the response head loss weight is 1, the reliability head loss weight is 0.8 when \(z \leq 7\); 1.8 when \(z = 8\), respectively. As for Three-Head, the response head loss weight is 10, the power head loss weight is 2 and the reliability head loss weight is 2. The training size is 300,000 when \(z \leq 10\); 600,000 when \(z \geq 12\), respectively, if the attacks use power SCI (Two-Head A, Three-Head, and Multiclass A). As for Two-Head B, the training size is 600,000.

As shown in Fig. 8, the performance of MLMSA and SLMSA are similar though the MLMSA is simpler. More specifically, when power SCI is leveraged, both MLMSA and SLMSA can break (2, 2, 30)-OAX-APUF with an accuracy of about 88%. Two-Head B with reliability SCI can reliably break (2, 2, 8)-OAX-APUF with an accuracy of more than 95%—later in Section VI-B we show that (2, 2, 9)-OAX-APUF is statistically breakable.

These experiments further validate the security of the OAX-APUF. Compared with the L-XOR-APUF, the \((x, y, z)\)-OAX-APUF with \(l = x + y + z\) is slightly easier to be...
modelled in front of DL-based attacks, because the OR and AND are easier to be approximated than the XOR operation by DL. Despite the OAX-APUF defeats CMA-based reliability modeling attacks and improves the modeling resilience to the LR-based modeling attacks with only CRPs for training [11].

4) FF-APUF: For FF-APUF, we compare 1) the Two-Head model of MLMSA with multiclass of SLMSA attack and 2) the pure CRP-based DL attack [27]. Note for the first two types of attacks, only the reliability SCI is utilized.

In this experiment, the number of the hidden layer is set to be 2 for Two-Head B and Multiclass B. The training size is 30,000 when the loop number is less than 4; 600,000 when the loop number is 4, 5, and 6. The weight of response head loss is 10, and the weight of reliability head loss is 2. There are three reliability SCI settings: ten times of repeated measurement with 11 classes; 19 times measurements with 4 classes (e.g., 0–4 are one class, 5–9 are one class); and 19 measurements with 20 classes. The challenge feature vector extraction method is consistent with [27].

As results detailed in Table III, the multiclass of MLMSA and multiclass of SLMSA attack can successfully model FF-APUF which has six loops with an accuracy of about 90%. Both attacks that are hybrid attacks that exhibit a better accuracy than the purely CRP-based DL modeling attack [27]. As for the repeated times of reliability SCI, when the response is measured repeatedly for 19 times and the results are divided into 20 categories (i.e., more repeated times and fine-grained class), the response accuracy obtained by the two attacks is the highest. This indicates the higher the fine-grained reliability SCI, the better.

MLMSA can be used to break XOR-FF-APUF, where the FF-APUFs are further XOR-ed. As shown in Fig. 9, when using response and power SCI, MLMSA can successfully break 10-XOR-FF-APUF when the training size is 300,000. By increasing the training size, larger XOR-FF-APUF (i.e., 20-XOR-FF-APUF) is also breakable. Note, here, all FF-APUFs have one loop.

VI. DISCUSSION

A. Different Loss Weights

We take Two-Head B of MLMSA attack to further explore the impact of different head loss weight settings on performance of the MLMSA. In this experiment, the response head loss weight is set to be 1, while the reliability head loss weight ranges between 0.5 and 2.0. The attacked strong PUFs are 10-XOR-APUF, (2, 2, 8)-OAX-APUF, and (5, 5)-iPUF. As shown in Fig. 10, when the reliability head loss weight is small, the chance of response accuracy greater than 90% tends to be small—each weight setting per strong PUF runs one model.
time. The other observation is the attacking accuracy stability, for 10-XOR-APUF and (5, 5)-iPUF, when the reliability head loss weight is greater than 1.5, the response prediction accuracy is high and stably maintained, e.g., above 90% of 10-XOR-APUF.

There are two general implications. First, a slightly higher reliability head loss weight is necessary to enforce its contribution. Otherwise, if its weight is too small, the Two-Head B attack degrades to CRP-only-based DL attacks, which is less effective exhibited by the lower chance of breaking large-scale APUF variants, e.g., 128-stage 10-XOR-APUF. Second, a properly set higher head loss can make the attacking accuracy remain stably high with smaller variance, which can be observed by the accuracy of 10-XOR-APUF and (5, 5)-iPUF.

According to our observations on different loss curves during the training in all aforementioned experiments, the power output head converges the fastest, followed by the response output head, and finally the reliability output head. Though we always adopt a fixed head loss in all our experiments throughout the training process, it is expected that dynamically tuning these loss weights may achieve improved attack effect, e.g., better accuracy or faster convergence for the total loss.

B. Reliability Hybrid Attack

The proposed MLMSA and reproduced SLMMSA attack [18] using the reliability SCI obtained from 11 repeated measurements can reliably model 10-XOR-APUF with an accuracy of 96%. As for (2, 2, 8)-OAX-APUF, the accuracy of the two attacks both reliably achieve 95%. In fact, when these two attacks run for multiple times, there is a certain probability that the response accuracy of even larger scaled 11, 12-XOR-APUF, (2, 2, 9)-OAX-APUF can reach more than 90% (i.e., being successfully broken). To be more precise, we have run five times, and there is one attempt reaching about 94%.

As shown in Table II, Two-Head B of MLMSA and Multiclass B of SLMMSA can successfully break 10-XOR-APUF, (2, 2, 8)-OAX-APUF and (5, 5)-iPUF when the reliability SCI assists the model training. In comparison, when only CRPs are used, the DL attack [29] can only break 5-XOR-APUF, (2, 2, 3)-OAX-APUF and (4, 4)-iPUF, which are inferior to the hybrid attacks. Note to be fair, these attacks by us have used the same MLP structure except for the output layer.

According to Liu et al. [18], the GRA-based hybrid attack using reliability SCI [19] can successfully crack a 128-stage 6-XOR-APUF. But it cannot crack a 12-XOR-APUF. GRA can break (6, 6)-iPUF, which cannot be broken by the two attacks we have attempted. The GRA exhibits improved performance over iPUF due to its knowledge of the differential model for the iPUF. In addition, the GRA relies on a multiple pass attack when the \( x > 1 \) in the \( (x, y) \)-iPUF. In other words, the \( y \)-APUF are first learned, then the \( x \)-APUF are learned sequentially by fixing the learned \( y \)-APUF.

In summary, by using the easily obtainable reliability SCI to assist the hybrid modeling attacks, larger-scaled strong APUF variants can be successfully broken, which cannot be achieved by using the CRP-only-based DL attacks.

C. Silicon Measurement Validations

Following [38] and [39], we use the public ROPUF dataset HOST2018 [40] to synthesize APUF, coined as RO-APUF. The key of this method is to use the reciprocal of four RO frequencies as the four-time delays of each stage of the RO-APUF. For example, when \( x = 1 \) in the \( (x, y) \)-iPUF. In other words, the \( y \)-APUF are first learned, then the \( x \)-APUF are learned sequentially by fixing the learned \( y \)-APUF.

1) Obtain the reciprocal of RO frequencies to serve as the path segment delays of APUF: the reciprocal of four RO frequencies as the four-time delays of each stage of the RO-APUF (see illustration in Fig. 11). To synthesize a 128-stage RO-APUF, 512 RO frequencies are utilized, which mainly have the following three steps.

2) The delay_cross^i = t^i_{14} − t^i_{23} and delay_uncross^i = t^i_{13} − t^i_{24} are computed to represent the cross path delay difference and uncross path delay difference of the i_{th}
stage. The \( w[i] \) is obtained through (5)

\[
\begin{align*}
    w[0] &= \left( \text{delay\_uncross}^0 - \text{delay\_cross}^0 \right) / 2 \\
    w[128] &= \left( \text{delay\_uncross}^{127} + \text{delay\_cross}^{127} \right) / 2 \\
    w[i] &= \left( \text{delay\_uncross}^{i-1} + \text{delay\_cross}^{i-1} \right) / 2 \\
         &\quad + \text{delay\_uncross}^{i} - \text{delay\_cross}^{i} / 2
\end{align*}
\]

\[i = 1, 2, \ldots, 127.\] (5)

3) Compute the response of a given challenge according to (1), (2), and (3).

The HOST2018 ROPUF dataset provides raw data from 217 Xilinx Artix-7 XC7A35T FPGAs, each containing a total of 6,592 ROs, comprising six different routing paths with 550 to 1696 instances per type [40]. Each RO frequency is evaluated 100 times at 5°C, 15°C, 25°C, 35°C, 45°C, and 55°C. These repetitive frequency measurements are used for reliability SCI for the RO-APUF. However, this dataset does not have power consumption measurements—the silicon-measured power SCI of the RO-APUF is unavailable. Therefore, the experiments below consider only the response and reliability SCI as inputs of the MLMSA when compared with its SLMSA counterpart—noting SLMSA in [18] is not evaluated against reliability SCI.

After synthesis, RO-APUF, XOR-RO-APUF, and OAX-RO-APUF are used to validate the efficiency of the proposed MLMSA attack with silicon measurements. The reference response is measured at 25°C. The reliability SCI is measured ten times at 55°C.

For the numerical simulated CRPs and the corresponding reliability SCI obtained through MATLAB simulator, the training size is 600 000. Both MLMSA and SLMSA can successfully break 10-XOR-APUF and (2, 2, 8)-OAX-RO-APUF. However, the same training size and head loss weight settings are not directly applicable to XOR-RO-APUF and OAX-RO-APUF with silicon measurements.

In order to achieve the same attack effect as the MATLAB numerical simulation, the training size and head loss weight settings are adjusted in some cases. When modeling the l-XOR-RO-APUF, the training size is adjusted to 600 000 when \( l \leq 7 \); 1200 000 when \( l = 8, 9 \); and 1500 000 when \( l = 10 \). As for \((x = 2, y = 2, z)\)-OAX-RO-APUF, the training size is adjusted to 600 000 when \( x + y + z \leq 9 \); 1200 000 when \( x + y + z = 10, 11 \); 1500 000 when \( x + y + z = 12 \). The response weight of all Two-Head B attacks on XOR-RO-APUF and OAX-RO-APUF are 1. For l-XOR-RO-APUF, the reliability loss weight is 0.8 when \( l = 5, 6, 8, 9; 1.8 \) when \( l = 7 \); and 1 when \( l = 10 \). The reliability loss weight for the OAX-RO-APUF is always 0.8. While Two-Head B and Multiclass B can model 10-XOR-APUF with an accuracy of about 95%. While Multiclass B of SLMSA can not break (2, 2, 8)-OAX-RO-APUF, Two-Head B of MLMSA can successfully break it with an accuracy of 97.35%.

Generally, the required number of training CRPs and corresponding reliability SCI silicon measurement attacks is larger than the number of numerical simulated strong PUFs. The potential reason is that the unreliability of RO-APUF is lower than that from numerical simulation. The bit error rate or unreliability of RO-APUF is about 3%–5%, while the unreliability of the APUF upon numerical simulation is about 5%–8%. More precisely, as we repeatedly measure the same response ten times to gain 11 categories for the reliability SCI, it indicates that the unreliable responses (categories of 0 and 10 are from those reliable responses, rest 1 to 9 categories are unreliable responses) in the RO-APUF is less than that in numerical simulation. Therefore, the contribution from the reliability SCI is reduced, which requires a larger training size. Fig. 12 manifests this conjecture. We can see that the unreliable response categories of (1–9) of the silicon measurement-based strong PUFs are much less than that from numerical simulations. Nonetheless, MLMSA and SLMSA can still successfully model XOR-RO-APUF and OAX-RO-APUF once the number of unreliable responses (i.e., those in categories 1–9) increases.

D. Curse of Dimensionality

We now compare SLMSA with MLMSA when the large-scaled l-XOR-APUFs are attacked. In this context, when using all three response, reliability, and power SCI features, the dimensionality is high. The results are detailed in Table IV and visualized in Fig. 13.

As we can see, MLMSA always outperforms SLMSA when the dimensionality increases in terms of not only accuracy but also training time. The training time reduction is significant. When attacking l-XOR-APUF, the total dimensionality of SLMSA is \( 2 \times cn \times (l + 1) \) and MLMSA is \( 2 + cn + (l + 1) \). For instance, when attacking 20-XOR-APUF using response, reliability SCI and power SCI, the dimensionality of SLMSA becomes 462 (i.e., \( 2 \times 21 \times 11 = 462 \)) and 840 (i.e., \( 2 \times 21 \times 20 = 840 \)) when \((m = 10, cn = 11)\) and \((m = 19, cn = 20)\), respectively. While for MLMSA, the summed dimensionality of its three output

---

**TABLE IV**

| Attack | (m, cn) | Num. Classes (Input Dimension) | Response Acc | Time |
|--------|---------|--------------------------------|--------------|------|
| MLMSA  | 1       | 21                             | 93.9%        | 7 min 3 s |
|        | 2       | 20                             | 98.12%       | 7 min 38 s |
|        | 3       | 19                             | 98.12%       | 7 min 6 s  |
|        | 4       | 18                             | 98.12%       | 7 min 17 s |
|        | 5       | 17                             | 98.12%       | 7 min 28 s |
|        | 6       | 16                             | 98.12%       | 7 min 40 s |
|        | 7       | 15                             | 98.12%       | 7 min 52 s |
|        | 8       | 14                             | 98.12%       | 7 min 64 s |
|        | 9       | 13                             | 98.12%       | 7 min 76 s |
|        | 10      | 12                             | 98.12%       | 7 min 88 s |
|        | 11      | 11                             | 98.12%       | 7 min 100 s |
|        | 12      | 10                             | 98.12%       | 7 min 12 s  |
|        | 13      | 9                              | 98.12%       | 7 min 24 s  |
|        | 14      | 8                              | 98.12%       | 7 min 36 s  |
|        | 15      | 7                              | 98.12%       | 7 min 48 s  |
|        | 16      | 6                              | 98.12%       | 7 min 60 s  |
|        | 17      | 5                              | 98.12%       | 7 min 72 s  |
|        | 18      | 4                              | 98.12%       | 7 min 84 s  |
|        | 19      | 3                              | 98.12%       | 7 min 96 s  |
|        | 20      | 2                              | 98.12%       | 7 min 108 s |

***The Num. Classes (output dimensionality) for Three-head MLMSA means the number of classes of (response, power, reliability).***
heads is $2 + 21 + 11 = 34$ and $2 + 21 + 20 = 44$, respectively, which is an order of magnitude smaller than that of SLMSA. The increased output dimensionality results in a larger network, which requires longer training time. In addition, when the number of classes of a single head greatly increases in SLMSA, the error increases—the classification hardness (slightly) goes up [41].

E. Lightweight Cryptographic Module Incorporation

The general take-away of this study is that by barely increasing the scale of a strong PUF, in particular, the APUF variants, is still challenging when it is confronted with rapidly evolving DL techniques, especially, by combining response information and multiple SCIs. Therefore, the practical solution of using strong PUFs appears to incorporate lightweight cryptographic modules, such as the Lockdown-PUF [42], TREVERSE constructions [38], and RSO-APUF [43], to protect the CRP interface. The overhead caused by lightweight cryptographic modules can be lower or comparable to the overhead incurred by increasing the APUF variants to a large scale, e.g., 128-stage 30-XOR-APUF being breakable.

F. MLP Insensitivity to Nonuniformity

Aghaie and Moradi [44] demonstrated that implementation defects resulted in bias that hardens certain ML-based modeling attacks (i.e., LR attacks) on complex PUF architectures (i.e., iPUF). Nonetheless, not any learning algorithm is sensitive to the APUF’s nonuniformity when it is used to modeling the APUF variants. As recognized by Aghaie and Moradi [44], the ANN (in essence, the MLP is used) is insensitive to the nonuniformity based on their validations, which has been used to replace LR to mitigate the nonuniformity degradation on the attacking accuracy. In MLMSA, we have utilized a similar MLP learning algorithm.

We have further evaluated the MLP’s attacking accuracy when the underlying PUFs’ uniformity varies from 0.5 to 0.9 to check whether the MLP attack is indeed sensitive to the PUF uniformity. The MLP structure, hyperparameter settings, and training CRP size in the experiments follow [29]. We note that the reported accuracy [29] is hardly achieved until we replace the activation function $\text{ReLU}$ used in [29] with $\text{Tanh}$. We note that the $l$-XOR-APUF with $l = 3, 4, 5$ and $(3, 3)$-iPUF, $(4, 4)$-iPUF have been used by [44], which we evaluate as well. As detailed in Fig. 14, when the uniformity increases, the accuracy almost does not decline but is even higher in some cases. Our results align with that of Aghaie and Moradi [44], that is the nonuniformity has negligible influence on increasing the APUF variants’ modeling resilience to MLP. Therefore, MLMSA evaluations building upon the MLP model architecture hold their validity.

G. Minimal Training Size

It is preferable to find the minimal number of training CRPs to comprehend the robustness of PUFs from a theoretical (i.e., PAC Learning [45]) or empirical (i.e., Intrinsic ID [46]) point of view to demonstrate the attack efficiency from another perspective. For ML-based APUF modeling attacks, it is non-trivial to gain a theoretical bound of the minimal training size. This is, especially, the case for MLMSA that leverages multiple features given the same challenge. We have aimed at empirically finding the minimal training size with the same MLP trainable size and reliability granularity. The results are shown in Tables V and VI against $l$-XOR-APUF and $(x, y)$-iPUF, respectively. Note that when the PUF scale goes up, the minimal training size does gradually increase, see the visualization in Fig. 15 with a much small slope. This minimal training size increase is slight (i.e., very small slope) compared to that reported by Becker [15] that used a reliability-based attack incorporating a divide-and-conquer strategy to break underlying APUFs within $l$-XOR-APUF one-by-one—the CMA-ES algorithm is leveraged. In this case, the attacking complexity (i.e., attacking time and required number of CRPs) is also (approximately) linear with $l$. We note that the CMA-ES will find each APUF’s numerical weights that have a linear relationship with the real-time-delay segments of the APUF. In other words, it needs to recover the mathematical model per APUF.
TABLE V
MINIMAL TRAINING SIZE OF MLMSA TO ATTACK  
(\(l\)-XOR-APUF—RESPONSE, POWER, AND RELIABILITY (\(m = 10\) AND \(cn = 11\)) FEATURES ARE USED. THE MLP HAS FOUR HIDDEN LAYERS  
FOR \(l = 20\) AND THREE HIDDEN LAYERS FOR THE REST)

| \(l\) | Training Size | Response Weight | Power Weight | Reliability Weight | Best Acc | Success Rate |
|------|---------------|-----------------|--------------|-------------------|----------|--------------|
| 5    | 50,000        | 2               | 10           | 2                 | 92.1%    | 10/10        |
| 6    | 40,000        | 2               | 10           | 2                 | 93.24%   | 10/10        |
| 7    | 30,000        | 2               | 10           | 2                 | 93.79%   | 10/10        |
| 8    | 15,000        | 2               | 10           | 2                 | 94.01%   | 10/10        |
| 9    | 10,000        | 2               | 10           | 2                 | 94.16%   | 10/10        |
| 10   | 5,000         | 2               | 10           | 2                 | 94.16%   | 10/10        |

Moreover, when the reliability SCI is used, the minimal training size can be reduced if the granularity (i.e., \(m\) and \(cn\)) of the reliability increases—see experimental results in Table VIII for attacking \(l\)-XOR-APUF and Table IX for attacking \((x, y, z)\)-OAX-APUF. For example, in Table VIII, when the reliability granularity is low (\(m = 10\) and \(cn = 11\)), the 7-XOR-APUF and 8-XOR-APUF cannot be broken. When \(m = 20\) and \(cn = 21\), the 7-XOR-APUF is breakable with the same training size but increased reliability granularity. When \(m = 50\) and \(cn = 51\), the 8-XOR-APUF is further broken with the same training size but a further increased reliability granularity.

H. Limitations

It has been found that the modeling attack results of simulated CRPs can (reasonably) reflect the APUF variants modeling resilience based on silicon CRPs [22]. Using simulated CRPs following the well-established linear additive delay model [10], [14], [22], [23], [26], [47], [48], [49] to evaluate the modeling resilience is a common and acceptable means, despite other parameters (i.e., uniformity and uniqueness) sometimes exhibiting some inconsistency when the physical implementation is not properly tuned (i.e., some paths are extremely asymmetric, resulting in severe bias).

Silicon measurement-based validations are preferable and have been conducted in other modeling resilience studies [9], [16], [22]. It is worth validating the MLMSA and SLMSA with silicon measurements, especially, for the power SCI in future work. The absence of silicon fabrication-based validation is a limitation of current study.

To remedy this limitation to some extent, following [38], [39], we have evaluated the MLMSA efficacy through silicon measurement via the synthesized RO-APUFs—the response and reliability SCI are emulated in this context. The characteristics are the same for the numerical simulation-based evaluations.

We further note that the implemented MLP model’s performance is dependent on its trainable parameters. That is, by properly tuning the MLP trainable size, the performance (i.e., the attacking accuracy) can be improved even given the same training size—see experimental results in Table VII. For example, when the 5-XOR-APUF is attacked, the minimal training size is reduced from 80,000 to 45,000 when more hidden layers are leveraged in the MLP—higher number of trainable parameters.
The weights or regularization factors of response and other SCIs (i.e., power and reliability) in the MLMSA are empirically determined, which have shown sufficient attacking performance. It is interesting to incorporate adaptive weights learning during the training to automate their optimization in future work.

VII. CONCLUSION

This work proposes the MLMSA attack that constructively leverages multithread DL to concurrently exploit useful multichannel information to attack strong PUFs, particularly, APUF variants. With this simple and efficient MLMSA attack, we have successfully attacked a 128-stage 30-XOR-APUF, a (9, 9)- and (2, 18)-iPUF, and a (2, 2, 30)-OAX-APUF when we have successfully attacked a 128-stage 30-XOR-APUF, a (9, 9)- and (2, 18)-iPUF, and a (2, 2, 30)-OAX-APUF. All these large-scaled strong APUF variants have not been achieved by state-of-the-art attacks. We conclude that MLMSA can serve as an efficient technique for examining other existing or emerging strong PUF’s modeling resilience due to its simplicity, efficacy, and the avoidance of underlying mathematical models.

REFERENCES

[1] C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, “Physical unclonable functions and applications: A tutorial,” Proc. IEEE, vol. 102, no. 8, pp. 1126–1141, Aug. 2014.
[2] Y. Gao, S. F. Al-Sarawi, and D. Abbott, “Physical unclonable functions,” Nat. Electron., vol. 3, no. 2, pp. 81–91, 2020.
[3] W. Liu et al., “XOR-based low-cost reconfigurable PUFs for IoT security,” ACM Trans. Embedded Comput. Syst., vol. 18, no. 3, pp. 1–21, 2019.
[4] Y. Gao, Y. Su, L. Xu, and D. C. Ranasinghe, “Lightweight (reverse) fuzzy extractor with multiple reference PUF responses,” IEEE Trans. Inf. Forensics Security, vol. 14, pp. 1109–1123, 2019.
[5] Y. Gao, Y. Su, S. Nepal, and D. C. Ranasinghe, “NoisFre: Noise-tolerant memory fingerprints from commodity devices for security functions,” 2021, arXiv:2109.02942.
[6] B. Gassend, D. Clarke, M. Van Dijk, and S. Devadas, “Silicon physical random functions,” in Proc. 9th ACM Conf. Comput. Commun. Security, 2002, pp. 148–160.
[7] G. E. Suh and S. Devadas, “Physical unclonable functions for device authentication and secret key generation,” in Proc. 44th ACM/IEEE Design Autom. Conf., 2007, pp. 9–14.
[8] S. S. Zalivaka, A. A. Ivanuik, and C.-H. Chang, “Reliable and modeling attack resistant authentication of arbiter PUF in FPGA implementation with trinary quadruple response,” IEEE Trans. Inf. Forensics Security, vol. 14, pp. 1109–1123, 2019.
[9] P. H. Nguyen, D. P. Sahoo, C. Jin, K. Mahmood, U. Rührmair, and M. van Dijk, “The interpose PUF: Secure PUF design against state-of-the-art machine learning attacks,” IACR Trans. Cryptogr. Hardw. Embedded Syst., vol. 2019, no. 4, pp. 243–290, 2019.
[10] D. Lim, J. W. Lee, B. Gassend, G. E. Suh, M. Van Dijk, and S. Devadas, “Extracting secret keys from integrated circuits,” IEEE Trans. Very Large Scale Integ. (VLSI) Syst., vol. 13, no. 10, pp. 1200–1205, Oct. 2005.
[11] J. Yao et al., “Design and evaluate recomposed QR-AND-XOR-PUF,” IEEE Trans. Emerg. Topics Comput., vol. 10, no. 2, pp. 662–677, Apr.–Jun. 2022.
[12] Y. Gao et al., “Systematically evaluation of challenge obfuscated APUs,” 2022, arXiv:2203.13316.
[13] N. Wissol, K. T. Mursi, I. P. Seifert, and Y. Zhuang, “Neural-network modeling attacks on XOR arbiter PUFs revisited,” IACR Cryptol. ePrint Arch., Bellevue, WA, USA, Rep. 555/2021, 2021.
[39] M. Gao, K. Lai, and G. Qu, “A highly flexible ring oscillator PUF,” in Proc. 51st Annu. Design Autom. Conf., 2014, pp. 1–6, doi: 10.1145/2593069.2593072.

[40] R. Hessellbarth, F. Wilde, C. Gu, and N. Hanley, “Large scale RO PUF analysis over slice type, evaluation time and temperature on 28nm Xilinx FPGAs,” in Proc. IEEE Int. Symp. Hardw. Oriented Security Trust (HOST), 2018, pp. 126–133.

[41] K. Fukunaga and T. E. Flick, “Classification error for a very large number of classes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-6, no. 6, pp. 779–788, Nov. 1984.

[42] M.-D. Yu, M. Hiller, J. Delvaux, R. Sowell, S. Devadas, and R. Hesselbarth, F. Wilde, C. Gu, and N. Hanley, “Large scale RO PUFs against machine learning attacks,” IEEE Trans. Multi-Scale Comput. Syst., vol. 2, no. 3, pp. 146–159, Jul.–Sep. 2016.

[43] J. Zhang and C. Shen, “Set-based obfuscation for strong PUFs against machine learning attacks,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 68, no. 1, pp. 288–300, Jan. 2021.

[44] A. Aghaie and A. Moradi, “Inconsistency of simulation and practice in delay-based strong PUFs,” IACR Trans. Cryptograph. Hardw. Embedded Syst., vol. 2021, no. 3, pp. 520–551, 2021.

[45] F. Ganji, S. Tajik, and J.-P. Seifert, “PAC learning of arbiter PUFs,” J. Cryptogr. Eng., vol. 6, no. 3, pp. 249–258, 2016.

[46] P. Santikellur and R. S. Chakraborty, “Correlation integral-based intrinsic dimension: A deep-learning-assisted empirical metric to estimate the robustness of physically unclonable functions to modeling attacks,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 41, no. 10, pp. 3216–3227, Oct. 2022.

[47] J. Shi, Y. Lu, and J. Zhang, “Approximation attacks on strong PUFs,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 39, no. 10, pp. 2138–2151, Oct. 2020.

[48] N. Wisiol, G. T. Becker, M. Margraf, T. A. Soroceanu, J. Tobisch, and B. Zengin, “Breaking the lightweight secure PUF: Understanding the relation of input transformations and machine learning resistance,” in Proc. Int. Conf. Smart Card Res. Adv. Appl., 2019, pp. 40–54.

[49] N. Wisiol, B. Thapaliya, K. T. Mursi, J.-P. Seifert, and Y. Zhuang, “Neural network modeling attacks on Arbiter-PUF-based designs,” IEEE Trans. Inf. Forensics Security, vol. 17, pp. 2719–2731, 2022.

Yansong Gao (Member, IEEE) received the Ph.D. degree from The University of Adelaide, Adelaide, SA, Australia, in 2017. Then, he was with Data61, CSIRO, Sydney, NSW, Australia, from 2017 to 2020. Currently, he is with the Nanjing University of Science and Technology, Nanjing, China, as an Associate Professor. His current research interests include side-channel attacks and countermeasures.

Jianrong Yao received the bachelor’s degree from the Beijing Jiaotong University, Beijing, China, in 2020. He is currently pursuing the master’s degree with the Nanjing University of Science and Technology, Nanjing, China. His current research interest includes side-channel attacks and countermeasures.

Lihui Pang received the bachelor’s degree from the Henan University of Science and Technology, Luoyang, China, in 2008, the M.Sc. degree from Xihua University, Chengdu, China, in 2011, and the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. She is with the University of South China, Hengyang, China, as a Lecturer, and as a Research Director with Sungkyunkwan University, Suwon, South Korea. Her current research interests are physical unclonable functions design and applications, signal separation, and recognition.

Wei Yang received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in January 2018. He is currently a Lecturer with the Nanjing University of Science and Technology, Nanjing, China. His main research interests include side-channel attacks and countermeasures.

Ammín Fu received the Ph.D. degree in information security from Xidian University, Xi’an, China, in 2011. From 2017 to 2018, he was a Visiting Research Fellow with the University of Wollongong (Wollongong Campus), Wollongong, NSW, Australia. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His research interest includes IoT security, cloud computing security, and privacy preserving.

Said F. Al-Sarawi (Senior Member, IEEE) received the general certificate in marine radio communication in 1987, the B.Eng. degree (First-Class Hons.) in marine electronics and communication from the Arab Academy for Science and Technology, Alexandria, Egypt, in 1990, and the Ph.D. degree in mixed analog and digital circuit design techniques for smart wireless systems with special commendation in electrical and electronic engineering from The University of Adelaide, Adelaide, SA, Australia, in 2003, where he also received the Graduate Certificate in education (higher education) in 2006. He is currently the Director of the Centre for Biomedical Engineering and a Founding Member of the Education Research Group of Adelaide with The University of Adelaide. His research interests include design techniques for mixed signal systems in complementary metal–oxide–semiconductor and optoelectronic technologies for high-performance radio transceivers, low-power and low-voltage radio-frequency identification systems, data converters, mixed-signal design, and microelectromechanical systems for biomedical applications. His current educational research is focused on innovative teaching techniques for engineering education, research skill development, and factors affecting students evaluations of courses in different disciplines.

Derek Abbott (Fellow, IEEE) was born in South Kensington, U.K. He received the B.Sc. degree (Hons.) in physics from Loughborough University, Loughborough, U.K., in 1982, and the Ph.D. degree in electrical and electronic engineering from The University of Adelaide, Adelaide, SA, Australia, in 1997, under the supervision of K. Eshraghian and B. R. Davis. His research interests include the areas of multidisciplinary physics and electronic engineering applied to complex systems. His research programs span a number of areas of stochastics, game theory, photonics, energy policy, biomedical engineering, and computational neuroscience. Dr. Abbott received a number of awards, including the South Australian Tall Poppy Award for Science, in 2004; an Australian Research Council Future Fellowship, in 2012; the David Dewhurst Medal, in 2015; the Barry Inglis Medal, in 2018; and the M. A. Sargent Medal for Eminence in Engineering, in 2019. He has served as an Editor and/or a Guest Editor for a number of journals, including the IEEE Journal of Solid-State Circuits, Journal of Optics B, Chaos, Fluctuation and Noise Letters, Proceedings of the IEEE, and the IEEE Photonics Journal. He has served on the board of Proceedings of the IEEE, and is currently on the Editorial Boards of Scientific Reports (Nature), Royal Society OS, Frontiers in Physics, PNAS Nexus, and IEEE Access. He serves on the IEEE Publication Services and Products Board and the current Editor-in-Chief of IEEE ACCESS. He is a Fellow of the Institute of Physics, London, U.K., and an Honorary Fellow of Engineers Australia.