MLMSA: Multi-Label Multi-Side-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 multi-label/head deep learning model architecture, this work proposes Multi-Label Multi-Side-channel-information enabled deep learning Attacks (MLMSA) 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—Physical unclonable function, Multi-side-channel information, Multi-label classification, Multi-head/output model.

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]–[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 provision [4], [5]. On the other hand, strong PUFs offer a very large number

Y. Gao, J. Yao, W. Yang, and A. Fu are with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Y. Gao is also with Data61, CSIRO, Australia. e-mail: {120106222744,yansong.gao;generalyzy;fuam}@njust.edu.cn.

L. Pang† (corresponding author) is with School of Electrical Engineering, University of South China, Hengyang, China, and security engineering laboratory, Sungkyunkwan University, Korea. e-mail: sunshine.plh@hotmail.com

S. Al-Sarawi and D. Abbott are with School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia. e-mail: {said.alsarawi;derek.abbott}@adelaide.edu.au

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]–[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 non-linearity injection techniques have been used to construct APUF variants including the representative l-XOR-APUF, (x, y)-Iterpose PUF (iPUF) [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 large number of underlying APUFs composited) [12] when accessing to 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 5 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 multi-class/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. 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 underling APUFs, which is non-trivial 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.

- We are the first to introduce multi-label/head classification to facilitate multi-SCI DL modeling attack, coined as MLMSA 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 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.
- 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 significantly increased training size of 30M [13].
- We have advanced the breakable APUF variants to a 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.
- 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 corresponded 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 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 mathematical model [20], the APUF has a linear additive structure, leading to vulnerability to modeling attacks. In the
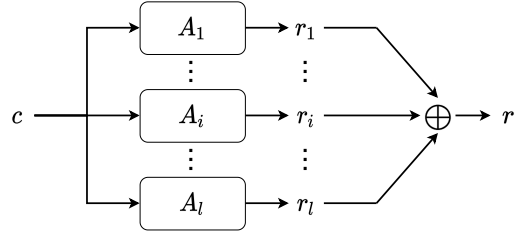

Fig. 1. l-XOR-APUF consists of l APUFs and each of the APUF response, $\{r_1, ..., r_l\}$ is XOR-ed at the end to form a 1-bit response r. All APUFs share the same challenge $\boldsymbol{c}$.

modeling attack, an attacker utilizes observed CRPs to build a mathematical PUF model that can accurately predict responses for unseen challenges [15], [21]–[23].

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

$$\Delta = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}, \tag{1}$$

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

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

The response of an n-stage APUF is determined by the delay difference Δ 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 Eq. 1, the response r of the challenge $\boldsymbol{c}$ is modeled as:

$$r = \begin{cases} 1, & \text{if } \Delta < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

B. XOR-APUF

As shown in Fig. 1, l-XOR-APUF has l underlying APUFs in parallel. Each APUF shares the same challenge and produces a digital response. All l responses are XOR-ed to form the final l-XOR-APUF response. Using a larger l can nearly exponentially increase the modeling attack complexity when *only CRP is used*. However, the l-XOR-APUF unreliaiblity 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
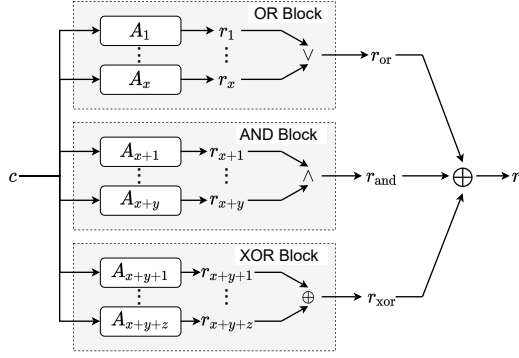
Fig. 2. Overview of (x, y, z)-OAX-PUF, which has three blocks: OR, AND, and XOR blocks.

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 $(\frac{x}{2} + y)$-XOR-APUF when the logistic regression (LR)-based divide-and-conquer attack is applied.
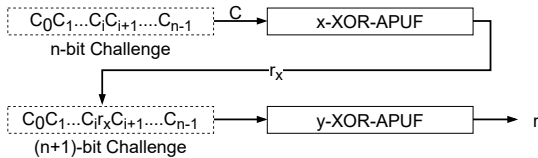

Fig. 3. n-bit (x,y)-iPUF [9].

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 a 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.
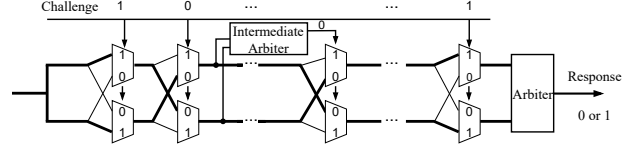

Fig. 4. An exemplified FF-APUF with one loop.

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: CRP-based ML attacks, SCI-based attacks and SCI hybrid attacks, which uses CRPs; SCI; and CRPs along with SCI(s) as training data, respectively.

A. CRP-based ML Attacks

Logistic regression (LR), support vector machine (SVM), and evolution strategies were utilized by Rührmair et $el.$ [21] to model XOR-APUF, FF-APUF and LSPUF in 2010. There are number of improvements to increase the attacking accuracy [25]. It is always suggested 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 $el.$ [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 minutes 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 $el.$ [18], the LDA attack can successfully break 128-stage $(6, 6)$-iPUF. According to Wisiol et $el.$ [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 underling mathematical strong PUF model*. Alkatheiri et $el.$ [27] showed that a 1-hidden layer MLP attack can successfully model FF-APUF with 6 loops in 2017. In 2018, Aseeri et $el.$ [28] proposed a 3-hidden layer MLP attack, which can successfully model 128-bit 7-XOR-APUF with 40M CRPs according to Wisiol et $el.$ [13]. Santikellur et $el.$ [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 compared to iPUF, XOR-APUF and OAX-APUF, it requires greatly increased APUFs and many MUXs (e.g., 2^k-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 el.* [31] proposed a 3-hidden layer MLP attack, which mainly focuses on XOR-APUF. According to Wisiol *et el.* [13], this 3-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 PUF that exploits the relationship between the response reliability and internal parameters [15]. The measured reliability data and challenges are provided to, e.g., 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 *et el.* [33] proposed a CPA-based CMA-ES attack that uses power correlation coefficients as the fitness function to model controlled PUFs and LSPUFs. An SPA-based LR attack was proposed by Rührmair *et el.* [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 gradient-based reliability hybrid attack (GRA) [19] and the other is the multi-class/single-label multi-SCI attack (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 are firstly learned, then the x-APUF are learned sequentially. In addition, the GRA requires to construct a differential model for the iPUF, which is non-trivial for adoption as it requires in-depth understanding of the underlying PUFs under attack.

The multi-class 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 firstly fuse CRP information with SCI to construct the so called challenge-synthetic-feature pairs (CSPs) 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 multi-class classification is substantially increased when the dimension per **p** and number of **p** increases due to 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 $2 \times (l + 1)$. 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 shown to successfully break 128-stage 16-XOR-APUF and $(2, 16)$-iPUF with $600,000$ training CSPs (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. MULTI-LABEL MULTI-SCI BASED DEEP LEARNING ATTACKS

We propose a multi-label DL based attack to efficiently and effectively take advantage of multiple SCIs, coined as MLMSA. Firstly, its output dimension is merely a summation per used SCI. Secondly, its can directly predict the response of

the learned strong PUF without additional remapping. Thirdly, 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. Multi-Label 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 multi-label model. In the multi-label 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 has multiple classes of comedy, romance, and action in the multi-label output.

The multi-label DL model [34], [35] is also usually referred to as multi-head/output DL model. Different from the single-label/head DL model, the output layer of the multi-head 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 multi-head model, then the loss of the model can be expressed as Eq.4:

$$L = \sum_{i=1}^{k} \lambda_i L_i, \tag{4}$$

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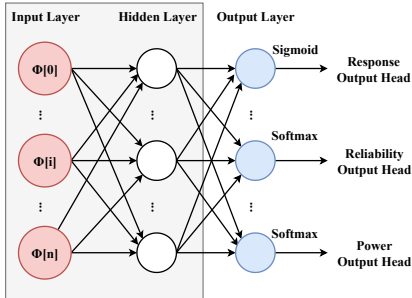

Fig. 5. Exemplified three-head DL architecture with heads of response, reliability SCI, and power SCI. The input layer and hidden layer(s) are shared by multiple heads.

B. MLMSA

As shown in Fig. 5, the multi-head 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 a 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., l-XOR-APUF is linearly proportional to number of responses being '1' in l 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 10 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 one-hot vector.

Training Stage: Using multi-head 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 multi-head model, according to Eq. 4.

Prediction Stage: Once the multi-head 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 10 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 128 stage—majority of previous studies using 64-stage APUF. 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 APUFs as in Eq. 1. In this context, we collect training/testing CRPs. The unreliability is produced by injecting 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 APUFs ranges

TABLE I
MLMSA Multi-Head Loss Weight λ Settings.

PUF	Multi-Head	Response Weight	Power Weight	Reliability Weight
XOR-APUF	Two-Head A	10	2	/
	Two-Head B	1	/	0.8(1.8)
	Three-Head	10	2	2(1)
OAX-APUF	Two-Head A	10	2	/
	Two-Head B	1	/	0.8(1.8)
	Three-Head	10	2	2
iPUF	Two-Head A	10(2)	2(3)	/
	Two-Head B	1	/	0.8
	Three-Head	10(2)	2	2

* 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 and reliability SCIs.

** In Two-Head B, for l-XOR-APUF, the reliability weight is 1.8 when $l = 10$; and 0.8 in other cases. In Three-Head, the reliability weight is 1 when $l = 29, 30$; 2 in other cases.

*** In Two-Head B, for (x, y, z)-OAX-APUF, the reliability head loss weight is 1.8 when $x + y + z = 12$; 0.8 in other cases.

**** In Two-Head A, for $(8, 8)$-iPUF, $(9, 9)$-iPUF, $(2, 16)$-iPUF and $(2, 18)$-iPUF, the response weight is 2, power weight is 3, these two weights are 10 and 2 respectively in other cases. In Three-Head, for $(8, 8)$-iPUF, $(9, 9)$-iPUF, $(2, 16)$-iPUF and $(2, 18)$-iPUF, the response weight is 2. In Three-Head, for other iPUFs, the weight of response is 10 (Three-Head).

***** These head loss weight settings are based on few empirically tuning. There may be better choices, which need to be analyzed on a case by case basis.

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 APUFs 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 training set and validation set are obtained by majority voting, and the response of 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 multi-head model as exemplified in Fig. 5 in each experiment is 3 or 4, and the activation function is ReLU. In [18], Liu *et al.* 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 hyper-parameter 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(R) Core(TM) 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, the Multi-Class A and Multi-Class B belong to the SLMSA attacks [18], where different SCIs are used:

1) **Multi-Class A:** The CSPs are formed with power SCI and CRPs;
2) **Multi-Class B:** The CSPs are formed with reliability SCI and CRPs.

Note that the Multi-Class 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 multi-head 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 Two-Head B model (response head and reliability head), the response weight is 1, the reliability loss weight is 0.8 when $l \leq 9$; 1.8 when $l = 10$. As for Three-Head, the response loss weight is 10, the power loss weight is 2. While the reliability loss weight is 2 when $l \leq 28$; 1 when $l = 29, 30$. The training size is $300, 000$ when $l \leq 12$; $600, 000$ when $l \geq 16$ if the attack uses *power* SCI (e.g., Two-Head A, Three-Head and Multi-Class A that is the SLMSA). As for Two-Head B that only uses the easy-to-obtain reliability SCI, the training size is $600, 000$ for all l settings—the largest l in this case is 10.

Fig. 6 depicts the results of multi-head (i.e., our MLMSA) and multi-class (in particular, the SLMSA) attacks on l-XOR-APUF ($l \leq 30$). When the power SCI is used, both MLMSA and SLMSA classification can attack 128-stage 30-XOR-APUF with accuracy about 90%, which scale has not been achieved in all previous studies. Liu *et el.* [18] only reported the accuracy of 97.8% when modeling 16-XOR-APUF by using Multi-Class A—as mentioned above, one potential reason is that Liu *et el.* [18] used 2 or 3 hidden layers that is less powerful than 3 or 4 hidden layers we adopted in the reproduction.

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 Multi-Class A are similar, This further indicates the dominant contribution of the power SCI compared to reliability SCI. Notably, *when the power SCI is unavailable*, reliability SCI can indeed help to break larger scale strong PUFs that CRP based modeling attacks cannot achieve alone, as specifically validated in Section VI-B.

2) (x, y)-iPUF: The loss weight settings for the (x, y)-iPUF are detailed in Table I. For most of the iPUF configurations with 128-stage x-APUF and 129-stage y-APUF, the response head loss weight is set to 10, the power head loss weight is 2, and the reliability head loss weight is 2. However, these settings are not successful in attacking $(8, 8)$-iPUF and $(2, 16)$-iPUF. It is necessary to tune the corresponding loss weight

TABLE II
COMPARISONS OF MLMSA WITH SLMSA (I.E., MULTI-CLASS) [18] AND DL WITH PURE CRP ATTACK [29] ON COMPLICATED STRONG PUFS.

PUF	Training CRPs	Two-Head A	Three-Head	Multi-Class A [18]	Two-Head B	Multi-Class B [18]	DL2019 [29]
5-XOR-APUF	300,000 (600,000 / 655,000)	96.97%	97.45%	98.25%	98.51%	98.27%	97.61%
6-XOR-APUF	300,000 (600,000 / 1,200,000)	96.98%	97.38%	98.09%	98.24%	97.54%	49.96%
10-XOR-APUF	300,000 (600,000)	95.43%	95.61%	97.12%	96.14%	96.32%	/
30-XOR-APUF	600,000	91.85%	89.77%	92.13%	/	/	/
(2,2,3)-OAX-APUF	300,000 (600,000)	97.49%	97.94%	97.40%	98.08%	97.87%	97.57%
(2,2,4)-OAX-APUF	300,000 (600,000)	97.18%	97.82%	97.37%	97.98%	96.81%	50.08%
(2,2,8)-OAX-APUF	300,000 (600,000)	96.17%	95.00%	96.31%	95.30%	96.84%	/
(2,2,30)-OAX-APUF	600,000	87.27%	84.23%	88.95%	/	/	/
(4,4)-iPUF	300,000 (600,000 / 647,000)	97.05%	97.58%	97.33%	96.82%	96.63%	74.73%
(5,5)-iPUF	600,000 (1,200,000)	96.94%	97.29%	97.09%	95.70%	95.36%	/
(8,8)-iPUF	600,000	95.79%	94.35%	95.71%	/	/	/
(9,9)-iPUF	600,000	95.29%	94.72%	95.47%	/	/	/
(2,16)-iPUF	600,000	92.81%	90.33%	92.82%	/	/	/
(2,18)-iPUF	600,000	89.26%	89.00%	90.49%	/	/	/

[*] 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 5-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].
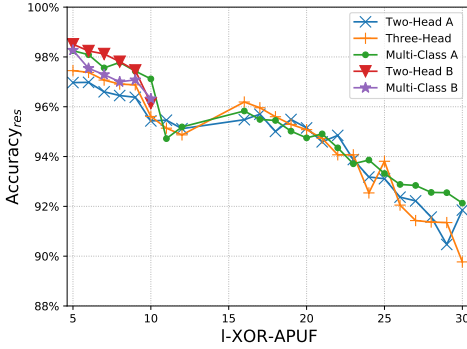

Fig. 6. Comparisons of MLMSA (i.e., multi-head) and SLMSA (i.e., Multi-Class) attacks using CRP, or/and power or/and reliability SCI on l-XOR-APUFs. The x-axis stands for the l of XOR-APUFs.

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

settings for optimization, the response head loss weight is 2, the power head loss weight is 3 when the Two-Head A attack is used; the response head loss weight is 2 when the Three-Head attack is used. As for Two-Head A, Three-Head and Multi-Class A, the training size is $600,000$. While for Two-Head B and Multi-Class B, the training size is $600,000$ for $(4, 4)$-iPUF, $1,200,000$ for $(5, 5)$-iPUF, respectively.

Though MLMSA is simple, it can also break $(2, 16)/(8, 8)$-iPUF with accuracy of $90.33\%/94.35\%$ that is comparable to the SLMSA when response, power SCI and reliability SCI are exploited. When using response, power SCI and reliability SCI, Three-Head of MLMSA can also successfully model $(2, 18)/(9, 9)$-iPUF with accuracy of $89.00\%/94.72\%$, which has not been reached 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 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 multi-head attacks are detailed in Table I. As for 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 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 $z \le 7$; 1.8 when $z = 8$, respectively. As for Three-Head, the response head loss weight is 10, the
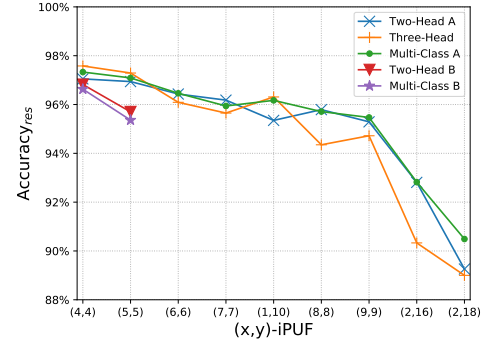
power head loss weight is 2 and the reliability head loss weight is 2. The training size is $300,000$ when $z \le 10$; $600,000$ when $z \ge 12$, respectively, if the attacks use power SCI (Two-Head A, Three-Head, and Multi-Class A). As for Two-Head B, the training size is $600,000$.

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 modeled 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 i) the Two-Head model of MLMSA with multi-class of SLMSA attack and ii) the pure CRP based DL attack [27]. Note for the first two type attacks, only the reliability SCI is utilized.

In this experiment, the number of hidden layer is set to be 2 for Two-Head B and Multi-Class B. The training size is
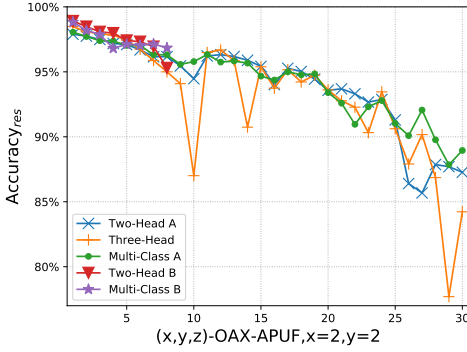
Fig. 8. Comparisons of MLMSA and SLMSA using power and reliability SCIs on OAX-APUFs. The x-axis stands for the z of (x, y, z)-OAX-APUFs.

$30,000$ 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: 10 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 multi-head of MLMSA and multi-class of SLMSA attack can successfully model FF-APUF which has 6 loops with an accuracy of about 90%. Both attacks that are hybrid attacks 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 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.
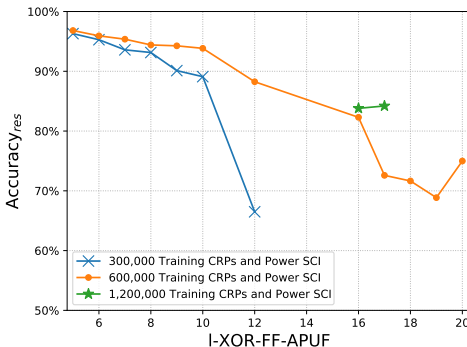

Fig. 9. Comparisons of Two-Head A Attack with different number of training CRPs and power SCI on l-XOR-FF-APUF in which FF-APUFs have 1 loop (63→80). The x-axis stands for the l of XOR-FF-APUFs.

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 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. Firstly, 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. Secondly, 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.
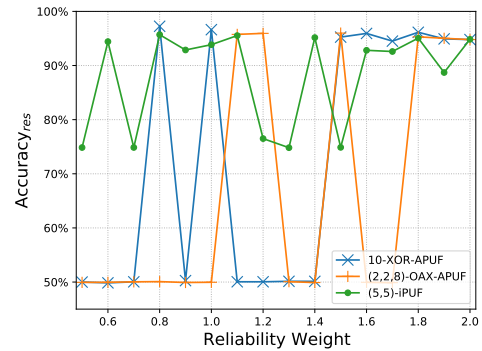
Fig. 10. Response prediction accuracy under different loss settings.

B. Reliability Hybrid Attack

The proposed MLMSA and reproduced SLMSA 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

TABLE III
COMPARISONS OF MLMSA WITH SLMSA (I.E., MULTI-CLASS) [18] AND DL WITH PURE CRP ATTACK [27] ON FF-APUFs WITH VARYING NUMBER OF LOOPS.

loopNum	(m, cn)	Start → End	Two-Head B	Multi-Class B	TowardsFast	Start → End	Two-Head B	Multi-Class B	TowardsFast
1	(19,20)	15→80	99.02%	99.04%	95.87%	63→80	98.94%	99.17%	97.77%
	(19,4)		98.75%	99.03%	95.61%		98.86%	99.14%	96.85%
	(10,11)		98.83%	98.91%	95.25%		98.95%	99.11%	98.06%
2	(19,20)	15→80,85	94.90%	95.00%	95.09%	63→80,85	94.88%	95.11%	95.17%
	(19,4)		94.79%	94.90%	95.11%		94.79%	94.83%	94.99%
	(10,11)		94.82%	94.92%	94.99%		94.83%	94.98%	95.07%
3	(19,20)	15→80,85,90	94.33%	94.31%	91.84%	63→80,85,90	94.34%	94.59%	93.36%
	(19,4)		94.17%	94.24%	91.53%		94.14%	94.39%	93.30%
	(10,11)		94.26%	94.69%	91.83%		94.22%	94.43%	92.88%
4	(19,20)	15→80,85 15→90,95	91.45%	91.53%	91.48%	63→80,85 63→90,95	91.50%	91.51%	91.44%
	(19,4)		91.41%	91.55%	91.40%		91.43%	91.49%	91.59%
	(10,11)		91.43%	91.55%	91.32%		91.29%	91.38%	91.38%
5	(19,20)	15→80,85 15→90,95,100	90.83%	90.73%	88.62%	63→80,85 63→90,95,100	91.12%	91.04%	89.62%
	(19,4)		90.66%	90.80%	88.69%		90.93%	91.05%	89.39%
	(10,11)		91.00%	91.01%	88.21%		91.12%	90.91%	89.49%
6	(19,20)	15→80,85,90 15→95,100,105	91.28%	91.38%	89.69%	63→80,85,90 63→95,100,105	91.22%	91.30%	90.15%
	(19,4)		91.33%	91.17%	89.52%		90.92%	91.21%	89.67%
	(10,11)		91.02%	91.21%	89.69%		91.09%	91.37%	90.02%

* Two-Head B of MLMSA uses response and reliability SCI. Multi-Class B of SLMSA uses response and reliability SCI.
** (m, cn) means reliability SCI is obtained by repeating the measurements for m times, which is divided into cn categories/classes.
*** In the experiments of FF-APUFs, the responses of training set are obtained by a majority vote. While the responses of test set are noise-free.

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 5 times, and there is one attempt reaching about 94%.

As shown in Table II, Two-Head B of MLMSA and Multi-Class B of SLMSA 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 firstly 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], [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 (see illustration in Fig. 12). To synthesize a 128-stage RO-APUF, 512 RO frequencies are utilized, which mainly has the following three steps.

1) Obtain the reciprocal of RO frequencies to serve as the path segment delays of APUF: the reciprocal of four RO frequencies are used as the segment time delays of the i-th stage of APUF: t_{13}^i, t_{14}^i, t_{23}^i, t_{24}^i, as illustrated in Fig. 12.
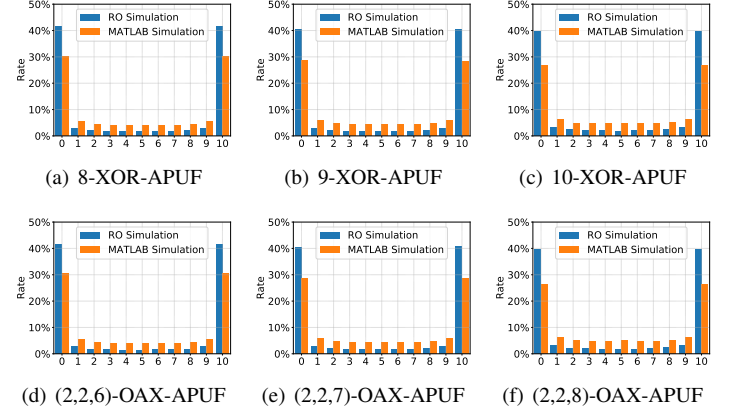

Fig. 11. The response unreliability SCI category statistics of XOR-APUF (a-c) and OAX-APUF (d-f) based on silicon measurement sythesized RO-APUF and MATLAB numerical simulated APUF.
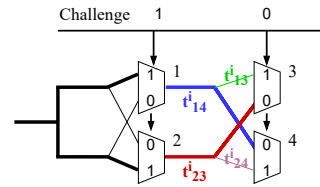

Fig. 12. Example of i-th stage of an APUF with two signal paths (i.e., top and bottom paths).

2) The delay_cross^i = $t_{14}^i - t_{23}^i$ and delay_uncross^i = $t_{13}^i - t_{24}^i$ 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 Eq. 5.

$$w[0] = (\text{delay\_uncross}^0 - \text{delay\_cross}^0)/2,$$
$$w[128] = (\text{delay\_uncross}^{127} + \text{delay\_cross}^{127})/2,$$
$$w[i] = (\text{delay\_uncross}^{i-1} + \text{delay\_cross}^{i-1} \qquad (5)$$
$$+ \text{delay\_uncross}^i - \text{delay\_cross}^i)/2,$$
$$i = 1, 2, ..., 127.$$

3) Compute the response of a given challenge according to

Eq. 1, Eq. 2, and Eq. 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 1,696 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 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 10 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-APUF. However, the same training size and loss head 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 $l \leq 7$; $1,200,000$ when $l = 8, 9$; and $1,500,000$ when $l = 10$. As for $(x = 2, y = 2, z)$-OAX-RO-APUF, the training size is adjusted to $600,000$ when $x + y + z \leq 9$; $1,200,000$ when $x + y + z = 10, 11$; $1,500,000$ 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. Both Two-Head B and Multi-Class B can model 10-XOR-APUF with an accuracy of about 95%. While Multi-Class 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% to 5%, while the unreliability of the APUF upon numerical simulation is about 5% to 8%. More precisely, as we repeatedly measure the same response 10 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. 11 manifests this conjecture. We can see

TABLE IV
Comparison between MLMSA and SLMSA when attacking large-scaled l-XOR-APUFs resulting into high dimensionality by using response, reliability SCI and power SCI.

Attack	l	(m,cn)	Num. Classes (Output Dimens.)	Response Acc	Time
MLMSA	10	(10,11)	(2,11,11)	95.94%	7 min 2 s
	10	(19,20)	(2,11,20)	96.12%	9 min 36 s
	12	(10,11)	(2,13,11)	94.43%	9 min 33 s
	12	(19,20)	(2,13,20)	94.29%	13 min 39 s
	16	(10,11)	(2,17,11)	96.08%	39 min 21 s
	16	(19,20)	(2,17,20)	96.21%	42 min 51 s
	20	(10,11)	(2,21,11)	95.14%	43 min 49 s
	20	(19,20)	(2,21,20)	94.28%	45 min 42 s
SLMSA	10	(10,11)	242	95.26%	27 min 33 s
	10	(19,20)	440	95.66%	39 min 14 s
	12	(10,11)	286	94.77%	32 min 42 s
	12	(19,20)	520	95.08%	55 min 37 s
	16	(10,11)	374	94.92%	89 min 44 s
	16	(19,20)	680	94.55%	2 h 38 min
	20	(10,11)	462	92.12%	2 h 59 min
	20	(19,20)	840	93.58%	4 h 59 min

* (m, cn) means reliability side-channel information is obtained by repeating the measurement for m times, which is divided into cn categories/classes.
** SLMSA means the specific Multi-Class C that uses response, power SCI and reliability SCI to build CSPs.
*** MLMSA and SLMSA have the same number of hidden layers. The number of hidden layer is 3 when $l = 10, 12, 16$. The number of hidden layer is 4 when $l = 20$.
**** The Num.Classes (output dimensionality) for Three-head MLMSA means the number of classes of (response, power, reliability). The Num.Classes for SLMSA Multi-Class C means $2 \times |power| \times |reliability|$.

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 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
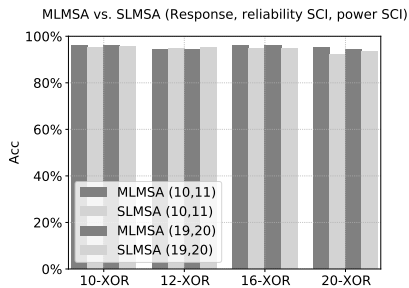
(a) Acc
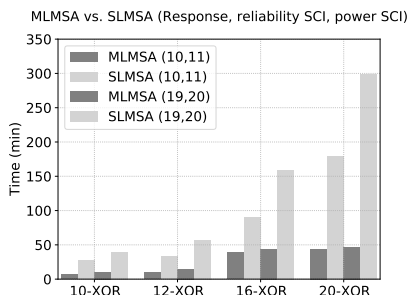
(b) Time

Fig. 13. MLMSA vs. SLMSA when the dimensionality is high (using response, reliability and power features).

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 Non-Uniformity

Aghaie *et al.* [44] demonstrated that implementation defects resulted in bias that hardens certain machine learning 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 non-uniformity when it is used to modeling the APUF variants. As recognized by Aghaie *et al.* [44], the ANN (in essence, the MLP is used) is insensitive to the non-uniformity based on their validations, which has been used to replace LR to mitigate the non-uniformity 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 underling 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 Relu used in [29] with 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
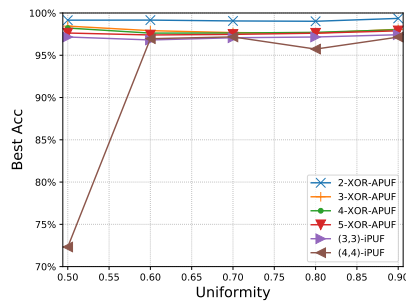

Fig. 14. MLP attacking accuracy with biased l-XOR-APUFs and (x, y)-iPUFs.

higher in some cases. Our results align with that of Aghaie *et al.* [44], that is the non-uniformity has negligible influence on increasing the APUF variants' modeling resilience to MLP. Therefore, MLMSA evaluations building upon the MLP model architecture hold their validity.

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	92.91%	10/10
6	62,000	2	10	2	93.24%	10/10
7	68,000	2	10	2	91.82%	8/10
8	80,000	2	10	2	91.94%	8/10
9	92,000	2	10	2	91.66%	6/10
10	100,000	2	10	2	92.12%	7/10
11	145,000	2	10	2	94.01%	10/10
12	164,000	4	5	2	93.98%	10/10
13	164,000	2	10	2	93.52%	6/10
14	200,000	4	5	2	93.72%	9/10
15	240,000	4	5	2	93.10%	8/10
16	260,000	10	8	2	93.09%	7/10
17	280,000	5	5	2	93.41%	6/10
18	300,000	5	5	2	93.51%	9/10
19	340,000	10	8	2	93.18%	7/10
20	360,000	10	5	2	93.88%	9/10

TABLE VI
MINIMAL TRAINING SIZE OF MLMSA TO ATTACK
(x, y)-IPUF—RESPONSE, POWER, AND RELIABILITY ($m = 10$ AND $cn=11$) FEATURES ARE USED. THE MLP HAS FOUR HIDDEN LAYERS.

(x,y)	Training Size	Response Weight	Reliability Weight	Power Weight	Best Acc	Success Rate
(4,4)	100,000	2	2	10	93.66%	9/10
(5,5)	130,000	2	2	10	94.35%	10/10
(1,10)	160,000	2	2	10	91.61%	7/10
(6,6)	170,000	2	2	10	94.17%	6/10
(7,7)	200,000	2	2	10	93.61%	8/10
(8,8)	250,000	2	2	10	93.55%	10/10
(2,16)	480,000	2	2	10	91.19%	6/10

TABLE VII
MINIMAL TRAINING SIZE GENERALLY DECREASES GIVEN HIGHER TRAINABLE PARAMETERS OF MLP IN THE MLMSA ATTACK AGAINST l-XOR-APUF AND (x, y)-IPUF.

PUF	Training Size	Hidden Layer Num	Response Weight	Power Weight	Reliability Weight	Best Acc	Success Rate
5-XOR-APUF	80,000	2	2	10	2	92.15%	8/10
5-XOR-APUF	50,000	3	2	10	2	92.91%	10/10
5-XOR-APUF	45,000	4	2	10	2	92.99%	9/10
(4,4)-iPUF	110,000	3	2	10	2	94.49%	10/10
(4,4)-iPUF	100,000	4	2	10	2	93.66%	9/10
(4,4)-iPUF	100,000	5	2	10	2	94.28%	9/10
(4,4)-iPUF	90,000	5	2	10	2	93.85%	7/10

TABLE VIII
INCREASE RELIABILITY GRANULARITY HELPS TO IMPROVE THE MLMSA
ATTACK ACCURACY GIVEN THE SAME TRAINING SIZE. THE
l-XOR-RO-APUF IS ATTACKED WITH A $600,000$ TRAINING SIZE
(RESPONSE AND RELIABILITY FEATURE ARE USED). THE MLP HAS
THREE HIDDEN LAYERS.

m	cn	5-XOR-APUF	6-XOR-APUF	7-XOR-APUF	8-XOR-APUF
10	11	98.38%	98.13%	50.25%	49.95%
20	21	98.30%	98.39%	97.72%	50.02%
30	31	98.52%	97.93%	97.75%	50.10%
40	41	98.37%	98.22%	97.99%	50.08%
50	51	98.35%	97.97%	97.98%	97.65%

TABLE IX
INCREASE RELIABILITY GRANULARITY HELPS TO IMPROVE THE MLMSA
ATTACK ACCURACY GIVEN THE SAME TRAINING SIZE. THE
(x, y, z)-OAX-RO-APUF IS ATTACKED WITH A $600,000$ TRAINING SIZE
(RESPONSE AND RELIABILITY FEATURE ARE USED).

m	cn	(2,2,2)	(2,2,3)	(2,2,4)	(2,2,5)	(2,2,6)
10	11	98.38%	98.07%	97.56%	97.19%	49.92%
20	21	98.40%	98.22%	97.75%	97.46%	50.31%
30	31	98.50%	98.19%	97.82%	97.64%	50.33%
40	41	98.50%	98.37%	97.89%	97.47%	51.46%
50	51	98.33%	98.33%	97.89%	97.42%	97.22%

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 machine learning 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 Table V and Table 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 Figure 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 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
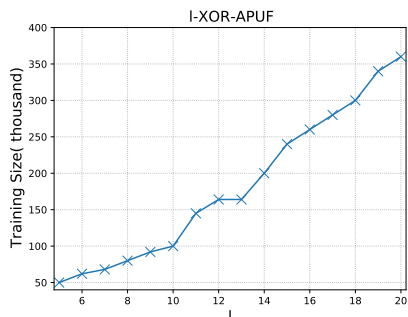
relationship with the real time-delay segments of the APUF. In other words, it needs to recover the mathematical model per APUF.

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.

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]–[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.

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.

Fig. 15. The minimal training size of MLMSA attack on l-XOR-APUFs.

VII. CONCLUSION

This work proposes the MLMSA attack that constructively leverages multi-head DL to concurrently exploit useful multi-

channel 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 CRPs, power SCI and reliability SCI are simultaneously used. With access to only easy-to-obtain reliability SCI and CRPs, the MLMSA can stably break a 128-stage 10-XOR-APUF, $(2, 2, 8)$-OAX-APUF, and 6-loop FF-APUF; and statistically break a 12-XOR-APUF and $(2, 2, 9)$-OAX-APUF. All these large-scaled strong APUF variants have not been achieved by state-of-the-arts 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 model.

REFERENCES

[1] C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, "Physical unclonable functions and applications: A tutorial," *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1126–1141, 2014.

[2] Y. Gao, S. F. Al-Sarawi, and D. Abbott, "Physical unclonable functions," *Nature Electronics*, vol. 3, no. 2, pp. 81–91, 2020.

[3] W. Liu, L. Zhang, Z. Zhang, C. Gu, C. Wang, M. O'neill, and F. Lombardi, "XOR-based low-cost reconfigurable PUFs for IoT security," *ACM Transactions on Embedded Computing Systems (TECS)*, 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 Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1887–1901, 2018.

[5] Y. Gao, Y. Su, S. Nepal, and D. C. Ranasinghe, "NoisFre: Noise-tolerant memory fingerprints from commodity devices for security functions," *arXiv preprint arXiv:2109.02942*, 2021.

[6] B. Gassend, D. Clarke, M. Van Dijk, and S. Devadas, "Silicon physical random functions," in *Proc. of the 9th ACM Conf. on Computer and Communications Security*, 2002, pp. 148–160.

[7] G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in *44th ACM/IEEE Design Automation Conf.*, 2007, pp. 9–14.

[8] S. S. Zalivaka, A. A. Ivaniuk, and C.-H. Chang, "Reliable and modeling attack resistant authentication of arbiter PUF in FPGA implementation with trinary quadruple response," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 1109–1123, 2018.

[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 Transactions on Cryptographic Hardware and Embedded Systems*, 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 Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 10, pp. 1200–1205, 2005.

[11] J. Yao, L. Pang, Y. Su, Z. Zhang, W. Yang, A. Fu, and Y. Gao, "Design and evaluate recomposited OR-AND-XOR-PUF," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 2, pp. 662–677, 2022.

[12] Y. Gao, J. Yao, L. Pang, Z. Zhang, A. Fu, N. Xiong, and H. Kim, "Systematically evaluation of challenge obfuscated APUFs," *arXiv preprint arXiv:2203.15316*, 2022.

[13] N. Wisiol, K. T. Mursi, J.-P. Seifert, and Y. Zhuang, "Neural-Network-Based Modeling Attacks on XOR Arbiter PUFs Revisited." *IACR Cryptol. ePrint Arch.*, vol. 2021, p. 555, 2021.

[14] J. Delvaux and I. Verbauwhede, "Side channel modeling attacks on 65nm arbiter PUFs exploiting CMOS device noise," in *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, 2013, pp. 137–142.

[15] G. T. Becker, "The gap between promise and reality: On the insecurity of XOR arbiter PUFs," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2015, pp. 535–555.

[16] U. Rührmair, X. Xu, J. Sölter, A. Mahmoud, M. Majzoobi, F. Koushanfar, and W. Burleson, "Efficient power and timing side channels for physical unclonable functions," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2014, pp. 476–492.

[17] S. Tajik, E. Dietz, S. Frohmann, H. Dittrich, D. Nedospasov, C. Helfmeier, J.-P. Seifert, C. Boit, and H.-W. Hübers, "Photonic side-channel analysis of arbiter PUFs," *Journal of Cryptology*, vol. 30, no. 2, pp. 550–571, 2017.

[18] W. Liu, R. Wang, X. Qi, L. Jiang, and J. Jing, "Multiclass classification-based side-channel hybrid attacks on strong PUFs," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 924–937, 2022.

[19] J. Tobisch, A. Aghaie, and G. T. Becker, "Combining optimization objectives: New modeling attacks on strong PUFs," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 357–389, 2021.

[20] R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, "Physical one-way functions," *Science*, vol. 297, no. 5589, pp. 2026–2030, 2002.

[21] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling attacks on physical unclonable functions," in *Proceedings of the 17th ACM Conf. on Computer and communications security*, 2010, pp. 237–249.

[22] U. Rührmair, J. Sölter, F. Sehnke, X. Xu, A. Mahmoud, V. Stoyanova, G. Dror, J. Schmidhuber, W. Burleson, and S. Devadas, "PUF modeling attacks on simulated and silicon data," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1876–1891, 2013.

[23] G. T. Becker, "On the pitfalls of using arbiter-PUFs as building blocks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 8, pp. 1295–1307, 2015.

[24] N. Wisiol, C. Mühl, N. Pirnay, P. H. Nguyen, M. Margraf, J.-P. Seifert, M. van Dijk, and U. Rührmair, "Splitting the interpose PUF: A novel modeling attack strategy," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 97–120, 2020.

[25] D. P. Sahoo, P. H. Nguyen, D. Mukhopadhyay, and R. S. Chakraborty, "A case of lightweight PUF constructions: Cryptanalysis and machine learning attacks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 8, pp. 1334–1343, 2015.

[26] D. P. Sahoo, D. Mukhopadhyay, R. S. Chakraborty, and P. H. Nguyen, "A multiplexer-based arbiter PUF composition with enhanced reliability and security," *IEEE Transactions on Computers*, vol. 67, no. 3, pp. 403–417, 2017.

[27] M. S. Alkatheiri and Y. Zhuang, "Towards fast and accurate machine learning attacks of feed-forward arbiter PUFs," in *IEEE Conf. on Dependable and Secure Computing*, 2017, pp. 181–187.

[28] A. O. Aseeri, Y. Zhuang, and M. S. Alkatheiri, "A machine learning-based security vulnerability study on XOR-PUFs for resource-constraint internet of things," in *IEEE International Congress on Internet of Things (ICIOT)*, 2018, pp. 49–56.

[29] P. Santikellur, A. Bhattacharayay, and R. S. Chakraborty, "Deep learning based model building attacks on arbiter PUF compositions," *Cryptology ePrint Archive*, vol. 2019, p. 566, 2019.

[30] M. A. Alamro and K. T. Mursi, "Machine learning attack on a multiplexer PUF variant using silicon data: a case study on rMPUFs," in *IEEE 6th International Conf. on Computer and Communication Systems (ICCCS)*, 2021, pp. 1017–1022.

[31] K. T. Mursi, B. Thapaliya, Y. Zhuang, A. O. Aseeri, and M. S. Alkatheiri, "A fast deep learning method for security vulnerability study of XOR PUFs," *Electronics*, vol. 9, no. 10, p. 1715, 2020.

[32] J. Delvaux and I. Verbauwhede, "Fault injection modeling attacks on 65 nm arbiter and RO sum PUFs via environmental changes," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 6, pp. 1701–1713, 2014.

[33] G. T. Becker and R. Kumar, "Active and passive side-channel attacks on delay based PUF designs," *Cryptology ePrint Archive*, vol. 2014, p. 287, 2014.

[34] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[35] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[36] A. Mahmoud, U. Rührmair, M. Majzoobi, and F. Koushanfar, "Combined modeling and side channel attacks on strong PUFs," *Cryptology ePrint Archive*, vol. 2013, p. 632, 2013.

[37] Y. Liu, Y. Xie, C. Bao, and A. Srivastava, "A combined optimization-theoretic and side-channel approach for attacking strong physical unclonable functions," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 1, pp. 73–81, 2017.

[38] Y. Gao, M. van Dijk, L. Xu, W. Yang, S. Nepal, and D. C. Ranasinghe, "TREVERSE: TRial-and-Error Lightweight Secure ReVERSE Authentication With Simulatable PUFs," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 419–437, 2022.

[39] M. Gao, K. Lai, and G. Qu, "A highly flexible ring oscillator PUF," in *Proc. of the 51st Annual Design Automation Conf.*, 2014, DOI:10.1145/2593069.2593072.

[40] R. Hesselbarth, F. Wilde, C. Gu, and N. Hanley, "Large scale RO PUF analysis over slice type, evaluation time and temperature on 28nm Xilinx FPGAs," in *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2018, pp. 126–133.

[41] K. Fukunaga and T. E. Flick, "Classification error for a very large number of classes," *IEEE transactions on pattern analysis and machine intelligence*, no. 6, pp. 779–788, 1984.

[42] M.-D. Yu, M. Hiller, J. Delvaux, R. Sowell, S. Devadas, and I. Verbauwhede, "A lockdown technique to prevent machine learning on PUFs for lightweight authentication," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 3, pp. 146–159, 2016.

[43] J. Zhang and C. Shen, "Set-based obfuscation for strong PUFs against machine learning attacks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 288–300, 2020.

[44] A. Aghaie and A. Moradi, "Inconsistency of simulation and practice in delay-based strong PUFs," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 520–551, 2021.

[45] F. Ganji, S. Tajik, and J.-P. Seifert, "PAC learning of arbiter PUFs," *Journal of Cryptographic Engineering*, 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 Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2021.

[47] J. Shi, Y. Lu, and J. Zhang, "Approximation attacks on strong PUFs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 10, pp. 2138–2151, 2019.

[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 *International Conference on Smart Card Research and Advanced Applications*. Springer, 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 Transactions on Information Forensics and Security*, vol. 17, pp. 2719–2731, 2022.