情報処理学会論文誌 Vol.61 No.12 1871–1880 (Dec. 2020)

意図的なエラーを付与することによる深層学習を用いたArbiter PUFへのクローニング攻撃の対策

八代 理紗[1,2,a)] 堀 洋平[2] 片下 敏宏[2] 崎山 一男[1]

受付日 2020年3月9日, 採録日 2020年9月10日

概要：本研究では，深層学習を用いたPUFへのクローニング攻撃の対策として，意図的なエラーを出力（レスポンス）に加える手法について有効性を明らかにする．Physically Unclonable Function（PUF）は，半導体素子のばらつきから回路固有の値を導出するセキュリティプリミティブである．近年，機械学習によってPUFの入力（チャレンジ）からレスポンスを予測する攻撃手法が提案されている．特に，深層学習を推定に用いることでレスポンスの予測成功率が高いクローンを容易に作製可能であると報告されている．そのため，深層学習を用いたクローニング攻撃に対する耐性向上が重要である．本論文では，クローニング攻撃のリソースであるチャレンジとレスポンスのペアに意図的なエラーを含め，深層学習により作製されたクローンのレスポンスの予測成功率を低下させることを狙う．より具体的には，PUFのレスポンスに意図的なエラーを一定の割合で加え，クローンの予測成功率と認証成功率に与える影響を明らかにする．255ビットのレスポンスによる認証を想定した実験の結果，5-XOR PUFに対して，意図的なエラーを31ビット加えることにより，クローンと正規品のPUFが識別できることを示す．

キーワード：physically unclonable function，深層学習，クローニング攻撃，意図的なエラー

Countermeasure against Deep Learning-based Cloning Attack on Arbiter PUF by Using Intentional Errors

Risa Yashiro[1,2,a)] Yohei Hori[2] Toshihiro Katashita[2] Kazuo Sakiyama[1]

Received: March 9, 2020, Accepted: September 10, 2020

Abstract: In this paper, we propose a countermeasure against deep learning-based cloning attacks on Physically Unclonable Function (PUF) by adding intentional errors to the output (response). PUF is a security primitive to obtain a unique value based on the variation of semiconductor devices. In recent years, however, the attack methods using machine learning to predict responses of PUF have been reported. In particular, by using deep learning, an attacker can easily create clones with high prediction accuracy. Therefore, it is important to prevent deep learning-based cloning attacks. The proposed method adds incorrect information (intentional errors) to challenge and response pairs to disturb the deep learning. In this paper, we show that our countermeasure is effective to decrease the prediction accuracy of clones. The experimental results show that adding a 31-bit intentional error to 5-XOR PUF can successfully distinguish clones from legitimate PUFs.

Keywords: physically unclonable function, deep learning, cloning attack, intentional error

1. はじめに

1.1 背景

近年，様々なモノがインターネットに接続され通信を行うInternet of Things（IoT）の普及にともない，約400億個のIoT機器がインターネットに接続されているとされ

[1] 電気通信大学大学院情報理工学研究科
The University of Electro-Communications, Department of Informatics, Chofu, Tokyo 182–8585, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki 305–8568, Japan
[a)] yashiro@uec.ac.jp

ている [1]．爆発的に増えた IoT 機器すべてを管理することは難しく，セキュリティリスクが増大しており深刻な問題となっている．セキュリティリスクを高める要因の 1 つとして，Integrated Circuit（IC）チップの模倣品があげられる．模倣 IC チップを利用することで，脆弱な秘密鍵の生成や誤動作を引き起こす可能性がある．もしそのような模倣 IC チップを実装した IoT デバイスが重要インフラで使用された場合，秘密情報の漏洩やシステム障害などのセキュリティインシデントを引き起こす可能性がある．そのため，模倣 IC チップの流通を防ぐ技術が必要である．

Physically Unclonable Function（PUF）[2], [3] は，模倣 IC チップの流通を防ぐために使われる技術の 1 つである．PUF は半導体のばらつきから IC チップ固有の出力（レスポンス）を導出する技術である．製造時ばらつきは物理的複製が困難である．しかし，チャレンジとレスポンスの関係性を数学的に複製するクローニング攻撃が報告されている [4], [5]．クローニング攻撃は，PUF のチャレンジとレスポンスの関係性を推定する．特に深層学習を用いた攻撃者は，クローニング攻撃によって高いレスポンスの予測成功率を実現できるため，PUF にとって大きな脅威である．ゆえに，クローニング攻撃に対する PUF の耐性を向上させることが重要である．

1.2 本研究の目的

本研究は，意図的なエラーをレスポンスに加えることによるクローニング攻撃の対策を検討する．攻撃者はクローニング攻撃のリソースとしてチャレンジとレスポンスのペア（CRP）を利用する．そのため，CRP において誤った情報，つまり意図的なエラーをレスポンスに加えることでクローンの予測成功率が低下すると考えた．遅延時間差パラメータをシミュレーションした PUF のレスポンスに意図的なエラーを加えてクローニング攻撃を行い，クローンの予測成功率の変化を検証した．使用したシミュレーション PUF は，実際の PUF と比べるとシンプルな構造の PUF を再現するものである．シンプルな構造の PUF は，動作に必要な遅延時間差の数が少なくなるため，攻撃者にとって推定が容易になり有利な条件となっていると考えられる．本論文では，より実際の PUF に近い構造を再現したシミュレーション PUF を作製し，意図的なエラーによってクローンの予測成功率の変化を検証した．

2. 先行研究

2.1 Physically Unclonable Function

PUF は，半導体素子のばらつきを基に固有の値を導出するため，ある PUF の出力は同じ入力であれば同一である．しかし，異なる PUF では，同じばらつきを持つ，つまり，同じ値を出力する PUF を物理的に複製することは困難である．PUF は，チャレンジ空間（入力の総数）の大きさによって confined PUF と extensive PUF の 2 つに分類できる[*1]．confined PUF はチャレンジ空間が小さく，extensive PUF はチャレンジ空間が大きい．

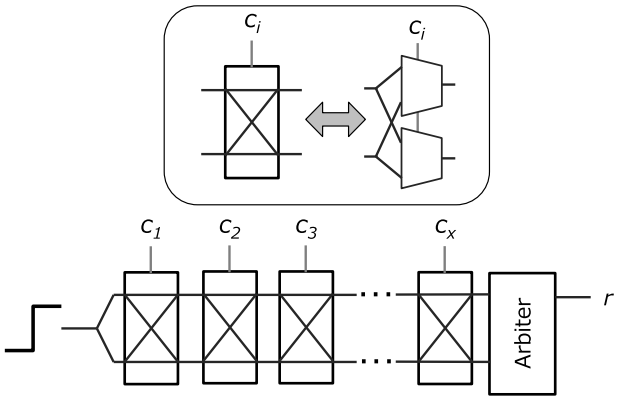


図 1 Arbiter PUF の構造
Fig. 1 Structure of Arbiter PUF.

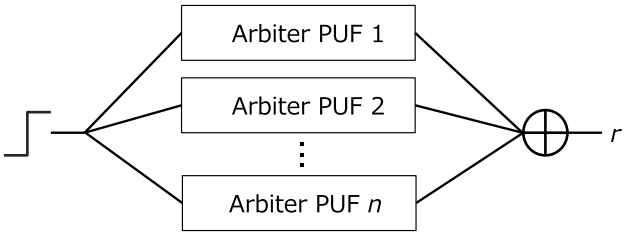


図 2 n-XOR PUF の構造
Fig. 2 Structure of n-XOR PUF.

2.2 Arbiter PUF と n-XOR PUF

extensive PUF の 1 つである Arbiter PUF [7] は，遅延時間差のばらつきを用いる PUF である．Arbiter PUF の構造を**図 1** に示す．伝搬経路は，チャレンジビットが 0 のときに直進，そうでなければ交差するように決定される．伝搬経路はセレクタペアごとに同じ構造であるため，理想的な状況では，各段の遅延時間は同じになる．しかし，実際には遅延時間は，配線長や閾値電圧の違いによってわずかに異なる値となり，チャレンジごとに固有の値となる．決定した伝搬経路に対して，立ち上がり信号を上の経路と下の経路に同時に入力し，どちらの経路を伝搬した信号の方が早かったか Arbiter 回路によって判定する．Arbiter 回路に入力された立ち上がり信号の速さによって，レスポンス 0/1 が決定される．

Arbiter PUF の挙動は正規品 PUF の CRP から遅延時間差を用いてモデル化することが可能だと報告されている [4], [5]．そこで，モデル化を難しくするために n-XOR PUF [8] が提案された．**図 2** に示すように，n-XOR PUF は n 個の Arbiter PUF から構成される．Arbiter PUF か

[*1] この呼び方は ISO/IEC 20897 で新たに定義されており，これまで weak PUF，strong PUF と呼ばれている分類と同じである [6].

ら出力された n ビットのレスポンスを XOR することによって，レスポンスが生成される．Arbiter PUF と比べて，n-XOR PUF は構造が複雑になるため，遅延時間差の推定に時間が必要になり，クローニング攻撃への脆弱性の改善が期待できるとされている．

2.3 Arbiter PUF で生じる遅延時間差のモデル化

Arbiter PUF の遅延時間差は数式によってモデル化することが可能であると報告されている [4], [5]．モデリング式は次のようになる．

x 段 Arbiter PUF の入力である x ビットのチャレンジの立ち上がり信号側から l 番目のチャレンジビットを $c_l \in \{0,1\}$ とする．l 段目で生じる遅延時間差は，$c_l = 0$ のとき δ_l^0，$c_l = 1$ のとき δ_l^1 とする．伝搬経路を表すパリティベクトル $(\vec{\Phi})$ は

$$\vec{\Phi}(\vec{C}) = (\Phi^1(\vec{C}), \ldots, \Phi^x(\vec{C}), \Phi^{x+1}(\vec{C}))^T \tag{1}$$

と表される．ここで $\Phi^l(\vec{C})$ は

$$\Phi^l(\vec{C}) = \begin{cases} (\prod_{i=l}^{x}(1-2c_i) & (l = 1, \ldots, x) \\ (1 & (l = x+1) \end{cases} \tag{2}$$

である．各パリティベクトルに対応する遅延時間差 $(\vec{w})$ を

$$\vec{w} = (w^1, w^2, \ldots, w^x, w^{x+1})^T \tag{3}$$

と定義する場合，w^i は

$$w^i = \begin{cases} (\delta_1^0 - \delta_1^1)/2 & (i = 1) \\ (\delta_{i-1}^0 + \delta_{i-1}^1 + \delta_i^0 - \delta_i^1)/2 & (i = 2, \ldots, x) \\ (\delta_x^0 - \delta_x^1)/2 & (i = x+1) \end{cases} \tag{4}$$

となる．すべての段のセレクタペアで発生した遅延時間差（Δ）は

$$\Delta = \vec{w}^T \vec{\Phi}. \tag{5}$$

と表すことができる．結果として，Arbiter PUF のレスポンス（r）は

$$r = \mathrm{sgn}(\Delta) = \begin{cases} 0 & (\Delta \leq 0) \\ 1 & (\Delta > 0) \end{cases} \tag{6}$$

となる．

2.4 クローニング攻撃

PUF のクローンを作製する攻撃をクローニング攻撃と呼ぶ．攻撃者は正規品の Arbiter PUF の CRP を取得し，遅延時間差のモデル化を行う．ここで遅延時間差の推定ができれば，未知のチャレンジに対するレスポンスを予測が可能，つまりクローンを作製できる．

n-XOR PUF は Arbiter PUF と比べて，環境ノイズによりビット反転したレスポンスが多く含まれる．構造の複雑さやビット反転したレスポンスの多さに起因し，Support Vector Machine（SVM）や Logistic Regression（LR）を使用したクローニング攻撃に時間がかかると報告されている [5]．しかし，環境ノイズによるビット反転の頻度は，Arbiter 回路に到達した際の遅延時間差に影響されることが指摘されている [9]．具体的には，遅延時間差の絶対値が 0 に近いほどビット反転が起きやすい．そのため，ビット反転の頻度はクローニング攻撃に役立つ情報となり，攻撃が容易になることが報告されている [10]．

Arbiter PUF は，その構成のシンプルさからレスポンスの固有性が低くなることがある．このような性質を改善するために提案された Arbiter PUF と類似する構造を持つ PUF がある [11], [12]．これらの PUF は，SVM や LR を使用したクローニング攻撃に対して耐性があることが報告されていた．しかし，近年では，深層学習をクローニング攻撃に用いることによって，クローンの予測成功率が高くなることが報告されている [13], [14], [15], [16], [17].

3. シミュレーション PUF を用いた関連研究の概要 [18]

3.1 提案された鍵共有システム

図 3 に関連研究 [18] で提案された鍵共有システムを示す．当該システムでは，被認証者は検証者から送られてきたチャレンジ（c）から，PUF のレスポンス（r'）に意図的なエラー（e）が追加された値 $(r'+e)$ を取得する．そして，$r'+e$ からヘルパーデータ（h）と秘密鍵（K）を生成し，ヘルパーデータ（h）を検証者に送信する．一方，検証者はデータベースに保存してあるレスポンス（r）に対して h を用いて，誤り訂正を行う．このとき，エラー数が誤り訂正能力より少なければ検証者は $r'+e$ を導出すること

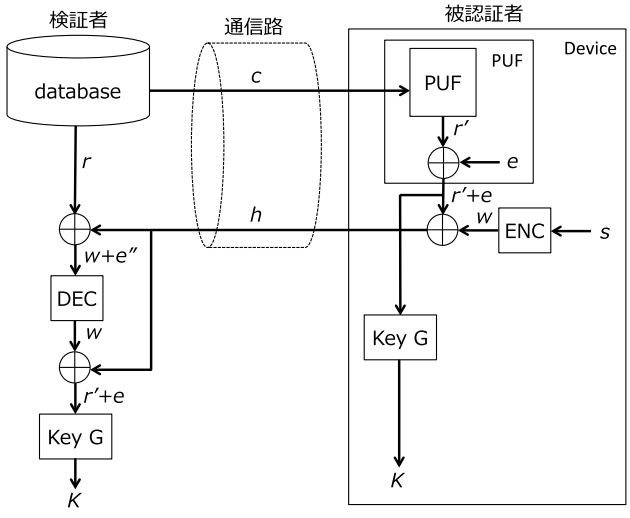


図 3 関連研究 [18] で提案された鍵共有システム
Fig. 3 Key sharing scheme suggested in Ref. [18].

ができる．$r'+e$ から鍵生成して K を生成する．つまり，検証者と被認証者で同じ鍵（K）を共有することができる．

3.2 提案されたシステムへの攻撃シナリオ

鍵共有システム [18] では，攻撃者はチャレンジ（c）と意図的なエラーが付与されたレスポンス $(r'+e)$ を取得して攻撃を行うことを想定している．c は検証者から被認証者に送信されているため，攻撃者は c を盗聴などによって取得できる．一方で $r'+e$ を攻撃者が取得できる環境は特殊な条件下である．そこで，攻撃者が $r'+e$ を取得できる攻撃シナリオについて考察する．

攻撃者は物理攻撃が可能であり，デバイス内の情報を取得できる状況を想定する．攻撃者はデバイス内の情報を取得できるため，デバイスから出力されていなくても $r'+e$ を取得できる．しかし一方で，攻撃者は r' や K を取得することで，予測成功率の高いクローンの作製や鍵共有後の通信内容を盗聴することが可能になる．いい換えると，もっと有益な攻撃リソースを攻撃者が取得できるため，$r'+e$ を用いてクローニング攻撃を行うとは考えにくい．そのため，この攻撃シナリオは，r' や K は取得できないが $r'+e$ は取得可能である特殊な状況の想定となる．

また，confined PUF を用いた鍵共有システムにおいて，c を再利用したときに攻撃者が h から r' を推定する攻撃 [19], [20] が提案されている．本論文が対象とする鍵共有システムは意図的 n エラーよってレスポンスが確率的に変わるが，CRP を再利用した場合は複数のレスポンスを集めてエラー e を消去することでこの攻撃が適用可能と考えられる．ただし，本論文が対象とする Arbiter PUF や XOR PUF などの extensive PUF は，チャレンジ空間がきわめて広く CRP を再利用する必要がないため，本論文では 1 度使用した CRP はデータベースから削除し再利用しないこととする．ゆえに，上述の攻撃は本論文では考慮しない．

3.3 クローニング攻撃に使用されたシミュレーション PUF

関連研究 [18] では，Santikellur ら [21] によって公開されたデータセットを使用している．このデータセットでは，2-XOR，3-XOR，4-XOR，5-XOR PUF それぞれ 1 つの PUF が公開されており，すべて Sahoo らが公開しているシミュレーションコード [22] を基に作成されている．しかし，Sahoo らによるシミュレーションで再現した Arbiter PUF は特殊な条件下の PUF である．具体的には，**図 4** に示すように，各段における遅延時間差パラメータが 1 つしか存在しない．実際の PUF ではチャレンジビットによって信号が伝搬する経路が違うため，遅延時間差が微小に異なる $(\delta_l^0 \neq \delta_l^1)$．チャレンジビットに関係なく遅延時間差が同じである可能性は低い．以下では，各段の遅延時間差が 1 パラメータで表される場合を δ-Arbiter PUF，2.3 節のモデルのように 2 パラメータで表される場合を $\delta^{0/1}$-Arbiter PUF などと表す．

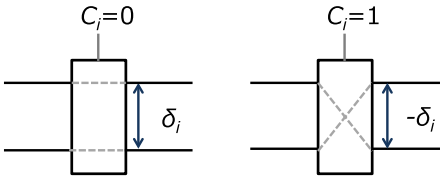


図 4 関連研究 [18] でクローニング攻撃に使用したシミュレーション PUF の遅延時間差

Fig. 4 Delay difference of simulation PUF used for cloning attack in Ref. [18].

4. 意図的なエラーによるクローンの予測成功率

4.1 関連研究 [18] をシンプルにしたシステムと想定する攻撃者

関連研究 [18] では，図 3 に示すような鍵共有システムを想定し，クローニング攻撃を行った．しかし，攻撃者が意図的なエラーを追加されたレスポンスを取得可能になるのは特殊な条件下である．そこで本論文では，**図 5** に示すような PUF を用いたシンプルな認証方式であるチャレンジレスポンス認証を想定する．本論文では 255 ビットで認証行う想定とする．検証者は事前に正規品の PUF から CRP を取得し，データベースに保存しておく．認証時にはデータベースからランダムに CRP を選択し，c を被認証者に送信する．被認証者は，PUF に c を入力し，得られる r' に意図的なエラー e を加えた $r'+e$ を取得する．$r'+e$ は検証者に送信され，検証者は保存してあるレスポンス r と比較する．このとき，r と $r'+e$ 間で異なるビット数が設定した閾値以下であれば認証が成功する．

チャレンジレスポンス認証では，リレー攻撃など様々な攻撃手法が考えられる．リレー攻撃は，攻撃者が検証者と PUF 間の通信を中継し，正規品の PUF のなりすましを行う．攻撃者が PUF へ通信できる場合，検証者から送信される c を PUF へ送信し，得られる r を検証者に送信することで，検証者と攻撃者間の通信が認証される．前述のように，Arbiter PUF や XOR PUF はチャレンジ空間がきわめて広いため，認証時に 1 度使用した CRP はデータベースから削除し，再利用しないと想定する．今回は深層学習を用いたクローニング攻撃のみを行う攻撃者を想定する．

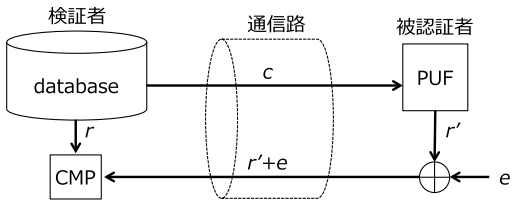


図 5 想定する認証方式

Fig. 5 Assumed authentication system.

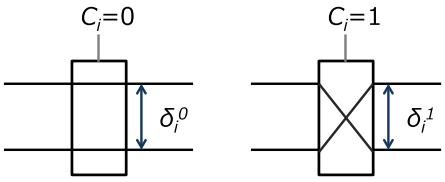


図 6 シミュレーション PUF の遅延時間差
Fig. 6 Delay difference of simulation PUF.

攻撃者は検証者と被認証者間の通信路を盗聴し，c と $r'+e$ を取得する．そして，取得した c と $r'+e$ からクローンを作製し，未知のチャレンジに対するレスポンスを予測して認証を試みる．

4.2 使用したシミュレーション PUF

関連研究では，Santikellur ら [21] によって作成されたオープンソースのデータセットをクローニング攻撃に使用していた．このデータセットは，各段の遅延時間差が 1 パラメータ（δ_l）で表されるシンプルな構造の δ-Arbiter PUF を再現しており，2-XOR，3-XOR，4-XOR，5-XOR PUF それぞれに対してデータセットが 1 つ公開されている．PUF が持つ遅延時間差はクローニング攻撃に大きく影響を与えると考える．たとえば公開データセットでは，チャレンジビットによる遅延時間差のない（δ_l）Arbiter PUF を再現しているが，実際の Arbiter PUF にはチャレンジビットに応じた遅延時間差（$\delta_l^{0/1}$）が存在すると考えられる．このとき，攻撃者が推定に必要な遅延時間差パラメータの数は増加すると考えられる．つまり，公開データセットは攻撃者にとって有利なデータセットになっている可能性がある．

そこで，今回使用するシミュレーション PUF は，関連研究と比較するために図 4 のように各段の遅延時間差パラメータが 1 つ（δ_l）の場合（以下 δ-PUF と呼ぶ）と**図 6** のように 2 つ（$\delta_l^{0/1}$）の場合（以下 $\delta^{0/1}$-PUF と呼ぶ）を想定する．$\delta^{0/1}$-PUF の場合，それぞれの遅延時間差は，実際の PUF から測定した値を参考に平均が 0，標準偏差が 10.75 の正規分布からランダムに決定した．δ-PUF は，チャレンジビットが 0 の遅延時間差（δ_i^0）のみを使用する．今回は，シミュレーション PUF 上に環境ノイズはなく Arbiter 回路で発生する遅延時間はない（レスポンスに偏りがない）条件とした．

4.3 クローニング攻撃の環境

δ-n-XOR PUF と $\delta^{0/1}$-n-XOR PUF（$2 \leq \mathrm{n} \leq 6$）をそれぞれ 10 個ずつシミュレーションで生成し，これに対してクローニング攻撃を行う．関連研究 [18] と同様に，255 ビットのレスポンスに対して平均で 0，7，15，23，31，42，47，55，63 ビットの意図的なエラーを付与し，クローンの予測成功率を評価する．付与する意図的なエラーのビット数は BCH 符号の誤り訂正能力に対応している．意図的なエラーはレスポンス 1 ビットに対して確率的に与える．たとえば，255 ビットに対して平均 7 ビットの意図的なエラーを加えるとは，レスポンス 1 ビットを 2.75%の確率で反転させることをいう．以下では，付与する意図的なエラーが平均 x ビットである場合も単に「x ビット」と表す．トレーニングデータセットは意図的なエラーを含む 500,000 CRP とし，テストデータセットは意図的なエラーがない 10,000 CRP とする．テストデータセットに意図的なエラーを加えない理由としては，意図的なエラーは乱数であり，どこに加わるか推測が困難なためである．つまり，攻撃者が作製するクローンの予測成功率が最も高くなるのは，意図的なエラーがないレスポンスを正しく予測できる場合になるためである．

今回の実験で使用した深層学習の詳細は付録 A.1 に記載する．

4.4 クローニング攻撃の結果

4.4.1 δ-PUF に対するクローニング攻撃

公開データセットと同様に遅延時間差パラメータが各段に 1 つ（δ_l）である δ-PUF（図 4）に対して，クローニング攻撃を行った結果を**図 7** と**表 1** に示す．図 7 は 10 個の PUF のクローンの予測成功率の平均を表しており，表 1 はクローンの予測成功率の最小値と最大値および信頼区間を表している．表 1 の中で，太字になっている部分はクローンの精度が 50±10%の値を表している．

仮に予測成功率が 10%の場合，予測したレスポンスをビット反転すれば 90%当たることになる．そのため，攻撃が成功するかどうかは予測成功率が 50%からどれだけ離れているかに依存する．つまり，太字になっている部分はクローンの攻撃成功率が低くなると期待されている場合である．

図 7 および表 1 より，クローンの予測成功率は大きく分けて以下の 3 種類に分けられる．

(a) 意図的なエラーを入れなくても予測成功率が 50%の場合，
(b) 意図的なエラーを入れても予測成功率が 90%以上で一定の場合，
(c) 意図的なエラーによって予測成功率が 90%前後から 50%前後に低下する場合．

(a) の場合

6-XOR PUF がこれに該当する．先行研究では行われなかった 6-XOR PUF に対して攻撃を行ったところ，意図的なエラーを付与しなくてもクローンの予測成功率が 50%となった．この場合，攻撃者は意図的なエラーの有無に関わらず有効なクローンを作製することができない．

(b) の場合

2-XOR PUF と 3-XOR PUF がこれに該当する．これら

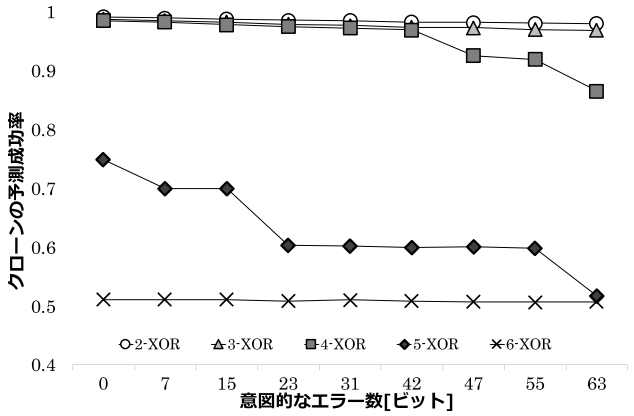


図 7 δ-PUF に対するクローンの平均予測成功率
Fig. 7 Average prediction rate of the clone of δ-PUF.

表 1 δ-PUF に対するクローンの予測成功率の最大値と最小値および信頼区間
Table 1 Maximum and minimum prediction rates and confidence inter vals of the clones of δ-PUF.

		0	7	15	23	31	42	47	55	63
2-XOR	max	99.21	99.05	98.94	98.82	98.73	98.59	98.47	98.40	98.32
	ave	99.13	98.97	98.77	98.59	98.53	98.28	98.23	98.09	97.99
	min	99.05	98.87	98.66	98.42	98.27	98.08	98.05	97.82	97.63
平均値の	上限	99.19	99.03	98.85	98.69	98.69	98.43	98.33	98.26	98.16
信頼区間	下限	99.08	98.91	98.70	98.48	98.37	98.13	98.13	97.92	97.81
3-XOR	max	98.91	98.61	98.44	98.08	98.13	97.55	97.48	97.26	96.63
	ave	98.78	98.54	98.25	97.89	97.72	97.38	97.31	96.98	96.85
	min	98.64	98.37	98.04	97.73	97.44	97.13	97.14	96.67	96.63
平均値の	上限	98.84	98.59	98.34	97.99	97.88	97.54	97.40	97.19	96.95
信頼区間	下限	98.69	98.40	98.11	97.75	97.49	97.28	97.18	96.80	96.68
4-XOR	max	98.77	98.49	98.36	98.02	97.92	97.52	97.24	96.44	96.18
	ave	98.52	98.28	97.83	97.53	97.26	96.97	92.53	91.94	86.55
	min	98.40	97.98	97.65	**55.98**	**50.92**	**54.88**	**53.30**	**54.09**	**53.64**
平均値の	上限	98.58	98.27	98.00	100.00	100.00	100.00	100.00	100.00	100.00
信頼区間	下限	98.39	97.98	97.45	85.22	77.92	80.62	77.83	67.31	66.82
5-XOR	max	98.34	98.02	97.81	97.27	96.58	96.41	96.38	96.02	**59.14**
	ave	74.90	69.96	69.91	60.33	60.13	**59.97**	60.02	**59.81**	**51.68**
	min	**50.99**	**50.91**	**50.89**	**50.62**	**50.22**	**50.34**	**50.37**	**50.32**	**50.30**
平均値の	上限	97.58	91.88	92.25	78.31	77.90	77.63	77.72	77.35	54.14
信頼区間	下限	52.23	48.05	47.80	42.36	42.35	42.32	42.31	42.28	29.23
6-XOR	max	**51.76**	**51.66**	**51.35**	**51.75**	**51.52**	**51.33**	**51.55**	**51.21**	**50.46**
	ave	**51.10**	**51.07**	**51.05**	**50.85**	**51.02**	**50.82**	**50.70**	**50.64**	**50.66**
	min	**50.81**	**50.80**	**50.88**	**50.14**	**50.50**	**50.18**	**50.30**	**50.00**	**50.18**
平均値の	上限	51.40	51.31	51.18	51.29	51.30	51.16	51.03	50.99	51.01
信頼区間	下限	50.81	50.84	50.91	50.40	50.75	50.48	50.37	50.29	50.32

の PUF では，関連研究 [18] と同様に意図的なエラーを加えてもクローンの予測成功率は低下しなかった．

(c) の場合

4-XOR PUF および 5-XOR PUF がこれに該当する．関連研究 [18] では，4-XOR PUF のレスポンスに意図的なエラーを加えても予測成功率は低下しなかった．しかし今回の実験では，付与する意図的なエラーが 0 から 42 ビットの場合に 4-XOR PUF のクローンの予測成功率は 98%であったが，47 ビットでは 93%，55 ビットでは 92%，そして 63 ビットでは 87%にまで低下させることができた．意図的なエラーを 47 ビット加えた際に，予測成功率が 50%近くまで低下するクローンが 1 つ存在した．

5-XOR PUF では，意図的なエラーを付与しない場合の平均予測成功率は 75%であったが，意図的なエラーを付与することで 52%–70%まで低下させることができた．23 ビットの意図的なエラーを加えることで予測成功率が 50%まで下がった PUF が 3 つ存在した．また，意図的なエラー数が 0 ビットでも予測成功率が 50%のクローンが 5 つ存在した．残りの 2 つの PUF では 55 ビットの意図的なエラーを加えても予測成功率が下がらなかった．

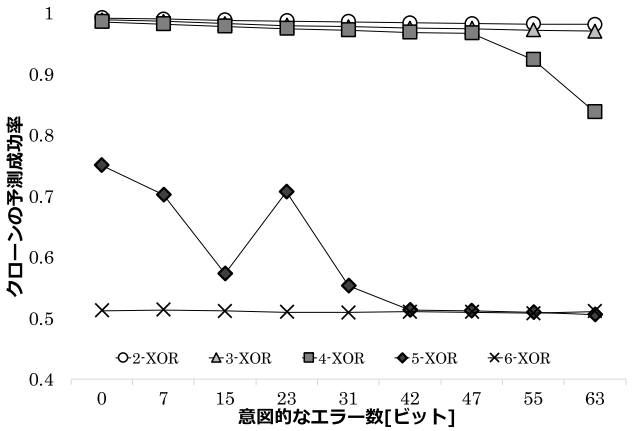


図 8 $\delta^{0/1}$-PUF に対するクローンの平均予測成功率
Fig. 8 Average prediction rate of $\delta^{0/1}$-PUF.

表 2 $\delta^{0/1}$-PUF に対するクローンの予測成功率の最大値と最小値および信頼区間
Table 2 Maximum and minimum prediction rates and confidence inter vals of the clones of $\delta^{0/1}$-PUF.

		0	7	15	23	31	42	47	55	63
2-XOR	max	99.21	99.09	98.85	98.77	98.71	98.51	98.43	98.44	98.15
	ave	99.13	98.95	98.74	98.62	98.52	98.32	98.20	98.09	98.03
	min	99.01	98.83	98.67	98.45	98.39	98.07	97.93	97.76	97.81
平均値の	上限	99.23	99.07	98.82	98.72	98.59	98.45	98.42	98.23	98.14
信頼区間	下限	99.03	98.82	98.65	98.43	98.29	98.08	97.93	97.63	97.86
3-XOR	max	98.93	98.69	98.35	98.09	97.91	97.60	97.59	97.37	97.19
	ave	98.80	98.58	98.20	97.87	97.68	97.44	97.35	97.11	96.96
	min	98.72	98.30	97.65	97.44	97.26	97.15	97.04	96.84	96.80
平均値の	上限	98.86	98.69	98.39	98.03	97.87	97.57	97.55	97.27	97.07
信頼区間	下限	98.73	98.47	98.01	97.70	97.50	97.31	97.14	96.94	96.85
4-XOR	max	98.56	98.30	98.03	97.62	97.30	96.99	96.82	96.66	96.39
	ave	98.46	98.09	97.76	97.37	97.14	96.73	96.63	92.29	83.75
	min	98.39	97.37	97.11	96.96	96.98	96.41	96.42	**55.68**	**52.17**
平均値の	上限	98.52	98.35	97.99	97.55	97.25	96.91	96.74	100.00	100.00
信頼区間	下限	98.40	97.84	97.53	97.20	97.04	96.56	96.51	80.38	65.37
5-XOR	max	98.62	97.95	97.94	97.28	96.85	**52.04**	**53.12**	**51.64**	**51.26**
	ave	74.95	70.15	**57.28**	70.70	**55.31**	**51.23**	**51.18**	**50.94**	**50.50**
	min	**51.03**	**50.65**	**51.14**	**50.64**	**49.20**	**50.64**	**50.47**	**50.21**	**49.21**
平均値の	上限	97.77	92.05	70.92	91.99	68.84	51.57	51.88	51.34	51.09
信頼区間	下限	52.14	48.26	43.65	49.41	41.79	50.89	50.48	50.55	49.91
6-XOR	max	**51.47**	**51.89**	**51.72**	**51.26**	**51.70**	**51.90**	**51.81**	**51.51**	**51.69**
	ave	**51.18**	**51.30**	**51.11**	**50.92**	**50.86**	**50.98**	**50.94**	**50.78**	**51.04**
	min	**51.03**	**50.65**	**51.14**	**50.64**	**49.20**	**50.64**	**50.47**	**50.21**	**49.21**
平均値の	上限	51.89	51.61	51.48	51.13	51.37	51.39	51.41	51.20	51.48
信頼区間	下限	50.98	50.98	50.74	50.72	50.35	50.57	50.47	50.36	50.60

4.4.2 $\delta^{0/1}$-PUF に対するクローニング攻撃

$\delta^{0/1}$-PUF（図 6）に対するクローンの予測成功率を**図 8** と**表 2** に示す．図 8 および表 2 より，予測成功率は δ-PUF と同様に (a)–(c) の傾向が見られる．

(a) の場合

6-XOR PUF がこれに該当する．6-XOR PUF は意図的なエラーを付与しなくてもクローンの予測成功率が 50%となった．この場合には攻撃者は有効なクローンを作製することができない．

(b) の場合

2-XOR PUF と 3-XOR PUF がこれに該当する．これらの PUF では，意図的なエラーを加えてもクローンの予測成功率はほぼ 90%以上で一定であり，低下しなかった．

(c) の場合

4-XOR PUF および 5-XOR PUF がこれに該当する．4-XOR PUF では付与する意図的なエラーが 0 から 47 ビットではクローンの予測成功率が 95%以上であったが，55 ビットでは 92%に，63 ビットでは 86%にまで低下させることができた．δ-PUF よりも多い 55 ビットの意図的なエラーを加えたときに予測成功率が 50%になるクローンが 1 つ存在した．また，63 ビットの意図的なエラーを加えたときには，予測成功率が 50%になるクローンが 3 つに増加した．

また，5-XOR PUF では，意図的なエラーを付与しない場合の予測成功率は 75%であったが，エラーを付与することによって 50–71%にまで低下させることができた．意図的なエラーが 0 ビットの時に予測成功率が 50%のクローンが 5 つ存在した．意図的なエラーを 42 ビット以上加えたときにすべての PUF で予測成功率が 50%前後となった．

5. 考察

5.1 チャレンジレスポンス認証の成否の考察

PUF を用いてチャレンジレスポンス認証を行うシステムにおいて，攻撃者がクローンを用いた場合の認証の成否について考察する．認証システムにはあらかじめ正規 PUF の CRP がデータベースに登録されているとし，認証に用いるレスポンスは 255 ビットとする．認証システムは閾値 θ を設定し，PUF から送られてくるレスポンスとデータベースに登録されているレスポンスのハミング距離が θ 以下であれば，認証成功とする．今回のシミュレーション上には環境ノイズは存在しないが，チャレンジレスポンス認証において環境ノイズの影響が大きいために環境ノイズの影響を考慮する．環境ノイズにより Arbiter PUF のレスポンスが反転する確率（BER: Bit Error Rate）は 0.03 とした．

(a) の場合

クローンの予測成功率が 50%である場合，255 ビット認証情報を用いると平均で 127.5 ビットの誤りが生じる．ただし，6-XOR PUF は今回の想定の中で最も n の数が大きいため，他の PUF よりも環境ノイズの影響を受けるレスポンス数が多くなる．Arbiter PUF の BER は 0.03 であるため，6-XOR PUF のレスポンスの BER は 0.1548 となり，環境ノイズの影響で生じる誤りレスポンス数は 39.5 ビットとなる．よって，127.5 ビットと 39.5 ビットの中間である 84 ビット前後に閾値を設定することで，正規品とクローンを正しく識別できる．また，意図的なエラーを入れると許容する誤りレスポンス数を増やすために閾値を大きくする必要があり，クローンを正規品と誤認証する可能性が高まるため意図的なエラーは入れない方がよい．

(b) の場合

クローンの予測成功率は意図的なエラーを 63 ビット入れても 2%しか下がらず，90%以上でほぼ一定である．一方で，63 ビットの意図的なエラーを許容するためには閾値を 63 ビットより大きくする必要があるが，予測成功率が 90%であるクローンにおいて予測を誤るビット数が 63 以下である確率は 0.999999999996156 となる．ゆえに，正規品とクローンをほとんど識別できないといえる．

(c) の場合

意図的なエラーを含ませることで，クローンが正規品と誤認証される状況から，正規品とクローンを正しく識別できる状況に改善させることができる．たとえば 5-XOR PUF では，意図的なエラーを 63 ビット付与した場合のクローンの予測成功率は最大でも 59.14%であった．この場合，クローンのレスポンス 255 ビット中に平均 104 ビットの誤りが含まれることになる．一方，環境ノイズによる BER が 0.03 である場合，正規 PUF のレスポンス 255 ビット中に平均 8 ビット程度の誤りが含まれることになる．ゆえに，63 ビットの意図的なエラーを付与する場合，閾値を 71（$= 63 + 8$）と 104 の間に適切に設定することで，正規 PUF とクローンを正しく識別できる可能性が高まる．

今回の評価では，実際の PUF の BER や，付与した意図的なエラーを環境ノイズが正してしまうケースなどを考慮できていない．これらを考慮した場合の予測成功率や認証成功率などの評価は今後の課題である．

5.2 モンテカルロシミュレーションによる正規 PUF とクローンの誤りレスポンス数の検証

クローン及び正規 PUF のレスポンスを確率的に生成するモンテカルロシミュレーションを行い，前述の (a)–(c) の場合における認証成功率を比較する．(a) では 6-XOR PUF，(b) では 3-XOR-PUF，(c) では 5-XOR PUF のクローンを評価対象とする．シミュレーションに用いるクローンは，学習に用いたレスポンスに (i) 意図的なエラーが含まれない場合，および (ii) 意図的なエラーを 255 ビットあたり 42 ビット含む場合を想定する．クローンの予測成功率は表 2 の最大値を用いる．すなわち，(a) では 6-XOR PUF の (i) 51.47%，(ii) 51.90%，(b) では 3-XOR PUF の (i) 98.93%，(ii) 97.60%，(c) では 5-XOR PUF の (i) 98.62%，(ii) 52.04%を用いる．正規 PUF では環境ノイズによる BER を 0.03 とし，XOR PUF の BER は 6-XOR PUF で 0.1548，3-XOR PUF で 0.0582，5-XOR PUF で 0.1330 となる．これらの BER とクローンの予測成功率を用いて，モンテカルロシミュレーションにより 255 ビット中の誤りレスポンス数を算出する試行を 1,000 回行った．

図 9，**図 10**，**図 11** に (a)–(c) における結果をそれぞれ示す．図における誤りレスポンス数は，意図的なエラーを除いた数値を示している．誤りレスポンス数の分布が正規 PUF とクローンで離れており，かつ正規 PUF の誤りビット数の方が小さい場合に，認証の閾値をその間に設定する

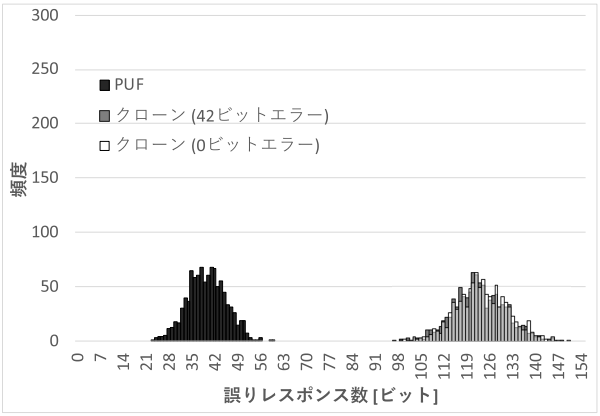


図 9 モンテカルロシミュレーションによる 6-XOR PUF を用いた認証の誤りレスポンス数

Fig. 9 The number of error responses of 6-XOR PUF in the authentication process simulated by Monte Carlo simulation.

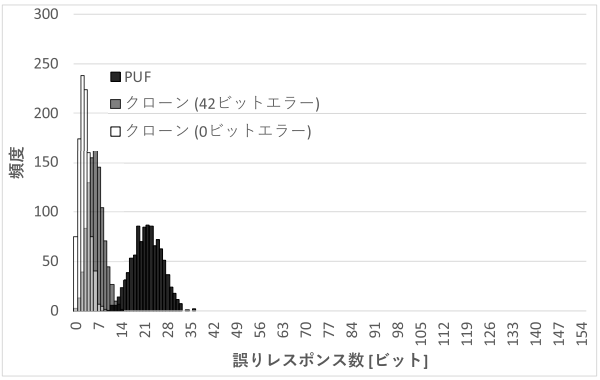


図 10 モンテカルロシミュレーションによる 3-XOR PUF を用いた認証の誤りレスポンス数

Fig. 10 The number of error responses of 3-XOR PUF in the authentication process simulated by Monte Carlo simulation.

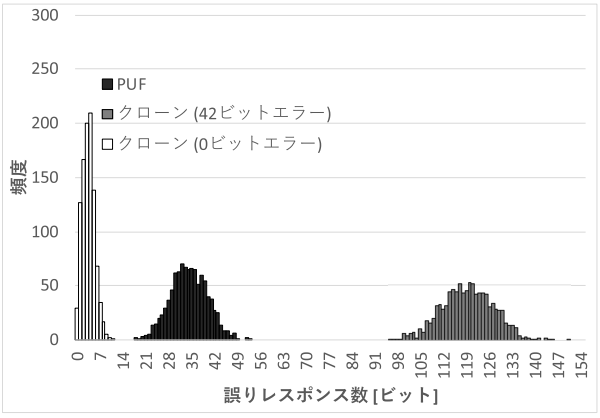


図 11 モンテカルロシミュレーションによる 5-XOR PUF を用いた認証の誤りレスポンス数

Fig. 11 The number of error responses of 5-XOR PUF in the authentication process simulated by Monte Carlo simulation.

ことで正規 PUF とクローンを識別しやすい状況となる．

(a) の場合のシミュレーション結果

図 9 が示すように，正規 PUF の誤りレスポンス数の最大値は 60，6-XOR PUF の (i) の場合の誤りレスポンス数の最小値は 99 であり，その差は 39 ビットとなった．ゆえに，この場合は誤りレスポンス数の分布が十分に離れているため，正規 PUF とクローンのを正しく識別できる可能性が高い．また，6-XOR PUF の (ii) の場合の誤りレスポンス数の最小値は 97 であり，正規 PUF との差は 37 ビットであった．しかし，正規 PUF のレスポンスには意図的なエラーが付与されて検証者に送られることを考慮する必要がある．(ii) では正規 PUF の 255 ビットレスポンスには意図的なエラーが平均 42 ビット付与されるため，クローンの誤りビット数と大きな差はなくなり，正規 PUF を正しく認証できなくなる可能性が高い．

(b) の場合のシミュレーション結果

(i) および (ii) 両方の場合で，正規 PUF の最大誤りレスポンス数の方が 3-XOR PUF のクローンの最小誤りレスポンス数よりも多く，その差は 36 ビットであった．ゆえに，この場合は正規 PUF を正しく識別することは困難である．

(c) の場合のシミュレーション結果

5-XOR PUF では，正規 PUF の最大誤りレスポンス数は 53 ビット，(i) の場合のクローンの最小誤りレスポンス数は 0 ビットであり，その差は 53 ビットであった．正規 PUF の方が誤りレスポンス数が多く，この場合は正規 PUF を正しく認証することはできない．

一方，(ii) の場合のクローンの最小誤りレスポンス数は 96 ビットであり，正規 PUF との最大誤りビット数との差は 44 ビットと，付与する意図的なエラーよりも大きかった．ゆえにこの場合，正規 PUF とクローンを正しく識別できる可能性が高い．

以上の結果から，適切な数の Arbiter PUF を組み合わせて n-XOR PUF とし，適切なビット数の意図的なエラーを付与することで，クローンが誤って認証されることを防ぎ正規 PUF のみを正しく認証することができるようになる．ゆえに，提案手法はクローニング攻撃に対して有効な防御策となりうる．

ただし今回は，クローンの予測成功率は攻撃者に有利となる最大値を用いたが，正規 PUF の BER は実環境の数値ではないうえ，誤りレスポンスビット数の平均を用いて考察を行った．また，誤りレスポンスビット数は確率的に決定されるが，平均より多く誤る場合は考慮されていない．しかし，仮に誤りビット数が今回の考察より増えた場合でも，本物拒否率や偽物受入率をどの程度許容するかを閾値 θ により適切に設定することで，提案システムが有効であるアプリケーションは存在すると考える．さらに，今回の実験では同一のレスポンスビットに環境ノイズと意図的エラーの両方が入る（打ち消しあう）場合は考慮されていな

い．BER，予測成功率，および閾値などの設定を様々に変えて提案システムがどのようなアプリケーションに対して有効であるかを検証することは今後の課題である．

6. まとめ

先行研究で行われていた意図的なエラーをレスポンスに加えてクローンの予測成功率を下げる手法を 2 種類のシミュレーション PUF に対して適用し，認証率を評価した．その結果，意図的なエラーを加えることでクローンの予測成功率を下げることが可能なことを示した．今回加えた意図的なエラーの範囲では，δ-5-XOR PUF に 63 ビット，$\delta^{0/1}$-5-XOR PUF に 42 ビットの意図的なエラーを加えることで正規品とクローンを判別できると考える．今後，意図的なエラーの範囲を拡張した場合にクローンの予測成功率がどうなるか，環境ノイズによりビット反転したレスポンスが存在するときに，クローンの予測成功率と認証成功率に与える影響を検討したい．

謝辞 この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務の結果得られたものを含みます．

参考文献

[1] 総務省：情報通信白書：進化するデジタル経済とその先にある Society 5.0，日経印刷 (2019).
[2] Pappu, R.: Physical One-way Functions, Ph.D. Thesis, Massachusetts Institute of Technology (2001).
[3] Pappu, R., Recht, B., Taylor, J. and Gershenfeld, N.: Physical One-Way Functions, *Science*, Vol.297, No.5589, pp.2026–2030 (2002).
[4] Lim, D.: Extracting Secret Keys from Integrated Circuits, Master's thesis, Massachusetts Institute of Technology (2004).
[5] Rührmair, U., Sehnke, F., Sölter, J., Dror, G., Devadas, S. and Schmidhuber, J.: Modeling Attacks on Physical Unclonable Functions, *Proc. 17th ACM Conference on Computer and Communications Security* (*CCS*), pp.237–249 (2010).
[6] Bruneau, N., Danger, J.-L., Facon, A., Guilley, S., Hamaguchi, S., Hori, Y., Kang, Y. and Schaub, A.: Development of the Unified Security Requirements of PUFs During the Standardization Process, *International Conference on Security for Information Technology and Communications*, pp.314–330, Springer (2018).
[7] Gassend, B., Clarke, D., Van Dijk, M. and Devadas, S.: Silicon Physical Random Functions, *Proc. 9th ACM Conference on Computer and Communications Security* (*ACM*), pp.148–160 (2002).
[8] Suh, G.E. and Devadas, S.: Physical Unclonable Functions for Device Authentication and Secret Key Generation, *Proc. 44th Annual Design Automation Conference* (*DAC*), pp.9–14 (2007).
[9] Delvaux, J. and Verbauwhede, I.: Side Channel Modeling Attacks on 65nm Arbiter PUFs Exploiting CMOS Device Noise, *Proc. International Symposium on Hardware-Oriented Security and Trust* (*HOST*), pp.137–142 (2013).
[10] Becker, G.T.: The Gap Between Promise and Reality: On the Insecurity of XOR Arbiter PUFs, *Proc. International Workshop on Cryptographic Hardware and Embedded Systems* (*CHES*), pp.535–555 (2008).
[11] Machida, T., Yamamoto, D., Iwamoto, M. and Sakiyama, K.: A New Arbiter PUF for Enhancing Unpredictability on FPGA, *The Scientific World Journal*, Vol.2015 (2015).
[12] Fruhashi, K., Shiozaki, M., Fukushima, A., Murayama, T. and Fujino, T.: The Arbiter-PUF with High Uniqueness Utilizing Novel Arbiter Circuit with Delay-Time Measurement, *2011 IEEE International Symposium of Circuits and Systems* (*ISCAS*), pp.2325–2328, IEEE (2011).
[13] Yashiro, R., Machida, T., Iwamoto, M. and Sakiyama, K.: Deep-Learning-Based Security Evaluation on Authentication Systems Using arbiter PUF and Its Variants, *Proc. International Workshop on Security* (*IWSEC*), pp.267–285 (2016).
[14] Awano, H., Iizuka, T. and Ikeda, M.: PUFNet: A Deep Neural Network Based Modeling Attack for Physically Unclonable Function, *Proc. IEEE International Symposium on Circuits and Systems* (*ISCAS*), pp.1–4 (2019).
[15] Khalafalla, M. and Gebotys, C.: PUFs Deep Attacks: Enhanced Modeling Attacks Using Deep Learning Techniques to Break the Security of Double Arbiter PUFs, *Proc. Design, Automation And Test in Europe* (*DATE*), pp.204–209 (2019).
[16] Santikellur, P., Bhattacharyay, A. and Chakraborty, R.S.: Deep Learning based Model Building Attacks on Arbiter PUF Compositions, Cryptology ePrint Archive, Report 2019/566 (2019). available from ⟨https://eprint.iacr.org/2019/566⟩.
[17] Yashiro, R., Hori, Y., Katashita, T. and Sakiyama, K.: Deep Learning Attack against Large n-XOR PUFs on 180nm Silicon Chips, *2020 International Workshop on Nonlinear Circuits, Communications and Signal Processing* (*NCSP*) (2020).
[18] Yashiro, R., Hori, Y., Katashita, T. and Sakiyama, K.: A Deep Learning Attack Countermeasure with Intentional Noise for a PUF-Based Authentication Scheme, *International Conference on Security for Information Technology and Communications*, pp.78–94, Springer (2019).
[19] Delvaux, J., Gu, D., Verbauwhede, I., Hiller, M. and Yu, M.-D.M.: Efficient fuzzy extraction of PUF-induced secrets: Theory and applications, *International Conference on Cryptographic Hardware and Embedded Systems*, pp.412–431, Springer (2016).
[20] Ueno, R., Suzuki, M. and Homma, N.: Tackling Biased PUFs Through Biased Masking: A Debiasing Method for Efficient Fuzzy Extractor, *IEEE Trans. Comput.*, Vol.68, No.7, pp.1091–1104 (2019).
[21] Santikellur, P., Bhattacharyay, A. and Chakraborty, R.S.: Modeling_of_APUF_Compositions, available from ⟨https://github.com/Praneshss/Modeling_of_APUF_Compositions⟩.
[22] Sahoo, D.P., Nguyen, P.H., Jin, C. and Mahmood, K.: DA_PUF_Library, available from ⟨https://github.com/scluconn/DA_PUF_Library⟩.
[23] Chollet, F. et al.: Keras (2015), available from ⟨https://keras.io⟩.

付　　録

A.1 使用した深層学習の詳細

クローニング攻撃に使用した深層学習のハイパーパラメータを**表 A·1** に示す．これらのハイパーパラメータは公開データセット [21] に対する予測成功率が低下しない，かつ複数施行した中で最も高い予測成功率のときのパラメータとした．深層学習に使用したマシンの詳細を**表 A·2** に示す．使用したハイパーパラメータおよびマシンでは，1 つのデータセットを学習するのに約 24 分の時間を要した．また，以下に keras 2.2.4 [23] で使用したコードの一部を記載する．

```
#ニューラルネットワークのモデル
model = Sequential(),
model.add(Dense(5000, activation='tanh', input_shape=(129,)))
model.add(Dropout(0.5))
model.add(Dense(1000, activation='sigmoid'))
model.add(Dropout(0.2))
model.add(Dense(500, activation='tanh'))
model.add(Dropout(0.2))
model.add(Dense(200, activation='sigmoid'))
model.add(Dropout(0.2))
model.add(Dense(100, activation='sigmoid'))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='mean_squared_error',
optimizer=Adam(), metrics=['accuracy'])

#過学習の傾向があるためコールバックを使用
ES = EarlyStopping(monitor='val_acc', patience=0, verbose=0)
CP = ModelCheckpoint("model.hdf5", monitor='val_acc',
verbose=0, save_best_only=True)
history = model.fit(x_train, y_train, batch_size=1000,
epochs=200, verbose=0, validation_data=(x_test, y_test),
callbacks=[ES, CP])
```

表 A·1 深層学習に使用するハイパーパラメータ
Table A·1 Hyper parameter used in the deeplearning.

隠れ層	$h1$	$h2$	$h3$	$h4$	$h5$
ユニット数	5,000	1,000	500	200	100
活性化関数	tanh	sigmoid	tanh	sigmoid	sigmoid
Dropout	0.5	0.2	0.2	0.2	–

表 A·2 深層学習に使用したマシン
Table A·2 Machine spec used in the deeplearning.

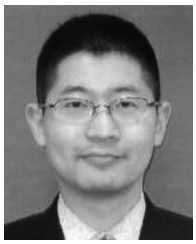
OS	Windows 10 Pro
CPU	Intel(R) Core(TM) i9-9900K
GPU	Nvidia GeForce RTX 2080 SUPER
ライブラリ	keras 2.2.4

八代 理紗

平成 27 年東海大学情報通信学部通信ネットワーク工学科卒業．平成 29 年電気通信大学大学院情報理工学研究科博士課程前期課程修了．同年より同大学院博士課程後期課程在籍．電子情報通信学会学生会員．

堀 洋平

平成 11 年筑波大学第三学群工業システム学類卒業．平成 15 年同大学院博士課程修了．博士（工学）．産業技術総合研究所特別研究員，中央大学専任研究員を経て，平成 22 年産業技術総合研究所研究員．平成 27 年より同研究所主任研究員，現在に至る．ハードウェアセキュリティおよび FPGA を利用した専用ハードウェアシステムの研究開発に従事．IEICE，IEEE，IACR 各会員．



片下 敏宏

平成 18 年筑波大学大学院システム情報工学研究科修了．同年産業技術総合研究所特別研究員．平成 20 年産業技術総合研究所任期付研究員，現在に至る．高速演算，セキュリティに関するハードウェアおよびソフトウェア設計の研究に従事．博士（工学）．

崎山 一男

平成 6 年大阪大学基礎工学部電気工学科卒業．平成 8 年同大学大学院修士課程修了．同年（株）日立製作所入所．平成 15 年カリフォルニア大学ロサンゼルス校・電気工・修士課程修了．平成 19 年ルーベン大学電気工・博士課程修了．平成 20 年電気通信大学准教授．平成 25 年より同教授，現在に至る．博士（工学）．国際暗号研究学会（IACR），IEEE 各会員．