SCIS 2013 The 30th Symposium on Cryptography and Information Security
Kyoto, Japan, Jan. 22- 25, 2013
The Institute of Electronics, Information and Communication Engineers

機械学習による遅延時間差検出型アービターPUF モデルを用いた認証方式
Authentication using RG-DTM PUF Model by Machine Learning

小川 昂佑* Kousuke Ogawa　汐崎 充† Mitsuru Shiozaki　藤野 毅†† Takeshi Fujino

あらまし 複製困難な製造ばらつき情報を抽出してチップ固有の値を生成する Physical Unclonable Function (PUF)を用いた認証が注目されている．この PUF は与えるチャレンジに対する出力レスポンスの一致度から簡易認証はできるが，認証に必要なチャレンジレスポンス・ペア（CRP）を全て記憶しておくのは非現実的である．例えば，本論文で用いる遅延時間差検出型アービターPUF(RG-DTM PUF)では，セレクタ段数 128 の場合には単純に 2^{128} もの CRP を記憶しておく必要がある．そこで，通常認証時には出力されないアービターPUF の内部遅延情報に対する機械学習攻撃を用いて PUF の認証時の出力を計算するモデルを生成することを提案する．この PUF モデルは CRP を記憶する代わりにチャレンジに対するレスポンスを計算するためのパラメータを記憶しておくため，小さな記憶容量で全ての CRP を記憶することができる．本論文では，検討した認証用 PUF モデルと，シミュレーションデータと 180nm CMOS プロセスで試作したチップの実測データを用いて認証率を評価した結果，また機械学習攻撃との兼ね合いについてまとめる．

キーワード PUF, 認証，機械学習，SVM，遅延時間差検出型アービターPUF

1 はじめに

近年, IC チップはサイドチャネル攻撃等の攻撃によってメモリに保持している秘密情報が窃取され，偽造・複製される危険性が指摘されている．そこで，秘密情報をメモリに保持せず，複製困難な製造ばらつき情報からチップ固有の値を生成する Physical Unclonable Functions (PUFs)を用いた認証が注目されている[1,2]. PUF は入力信号 (チャレンジ)に応じてデバイス固有の物理情報を抽出し，出力信号 (レスポンス)に変換するチャレンジ&レスポンス方式のデバイスである．LSI に実装される PUF[3]は，トランジスタや配線のサイズ， 閾値電圧などの製造ばらつきによる物理量の差異を抽出してデバイス固有のレスポンスを生成する．製造ばらつきはランダムに生じ人工的に制御することが困難なので PUF が生成するレスポンスは偽造・複製が困難で予測することのできないデバイス固有の情報となる．この PUF のデバイス固有情報を簡易認証に利用する場合, 与えたチャレンジによって生成されたレスポンスと，予め認証サーバ等に登録されていた CRP との一致度で認証を行うが，認証サーバ等に全ての CRP を保存するには非常に多くの記憶容量が必要であり非現実的である．例えば, Lee らが提案したアービターPUF[4]はセレクタ段数によって生成できる CRP は決まるが，セレクタ段数が 128 の場合には単純に 2^{128} もの CRP が生成でき，全てを記憶するのは難しい．また，PUF には別の問題がある．CRP を一定数入手できれば，機械学習を用いて同一 PUF をシミュレートする「モデル」が生成でき，未使用も含めて全ての CRP が予測可能となる安全面の問題である．実際に，アービターPUF は Support Vector Machine (SVM)を用いた機械学習攻撃により攻撃されることが報告されている[5].

本研究室は機械学習攻撃に耐性を持つ遅延時間差検出型アービターPUF(RG-DTM PUF)を提案してきた[6]. この RG-DTM PUF は CRP から機械学習攻撃するのは困難であるが，通常認証時には出力されない内部遅延情報とマルチクラス機械学習を用いれば PUF のモデル化で実現できると考えられる．この点に着目して，本論文

* 立命館大学大学院理工学研究科，〒525-8577，滋賀県草津市野路東 1-1-1, Graduate School of Science and Technology, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan, ri002073@ed.ritsumei.ac.jp,
† 立命館大学総合理工学研究機構，〒525-8577，滋賀県草津市野路東 1-1-1, Research Organization of Science & Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan, mshio@fc.ritsumei.ac.jp
†† 立命館大学理工学部，〒525-8577，滋賀県草津市野路東 1-1-1, Department of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan, fujino@se .ritsumei.ac.jp

では認証サーバの CRP データベースに代わる少ない記憶容量で済む認証用 PUF モデルと，この PUF モデルを用いた認証方法を提案する．

本論文の構成は以下の通りである．まず，第 2 章ではこれまで我々が提案してきた RG-DTM PUF と CRP を用いた機械学習攻撃に対する耐性評価結果についてまとめる．そして，第 3 章で新たに提案する機械学習を用いた PUF モデルによる認証方法について述べる．第 4 章では 180nm CMOS プロセスで試作したチップを使用し，生成した PUF モデルとの認証を評価した結果を示す．最後に，第 5 章でまとめと今後の課題について述べる．

2 RG-DTM PUF

2.1 遅延時間差検出方式

RG-DTM PUF は図 1 に示すように，セレクタを多段接続したセレクタチェーン回路と遅延時間差検出方式を適用したアービター回路で構成される．

セレクタチェーン回路は，チャレンジ信号 C に応じて IN からアービター回路までの 2 つの等価な経路を選択する．各セレクタ回路は等価に設計されているが，製造ばらつきによって全て異なる遅延時間を生じるため，チャレンジ毎に異なる遅延時間差を生じる．セレクタ段数が N 段の時，チャレンジ信号の組み合わせ数は 2^N となるので，2^N 種類の遅延時間差情報が得られる．

従来のアービターは 2 つの経路のうち，どちらの信号伝達が速かったかを判定し 0/1 のレスポンスを生成していたのに対して（図 2 左参照），RG-DTM PUF では図 2 右のように遅延時間差を複数区間に分割し，各区間にレスポンス 0/1 を割り振り，アービター回路により遅延時間差がどの区間に属するかを測定した結果からレスポンスを生成する．

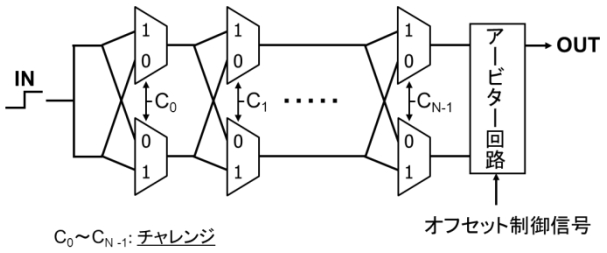


図 1　アービターPUF の回路図とレスポンス生成方法

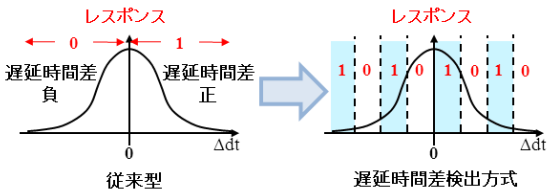


図 2　遅延時間差検出型アービターPUF

2.2 機械学習攻撃に対する耐性

機械学習攻撃は，PUF の CRP を幾つか手に入れることで，他の全ての CRP が予測できる攻撃であり，アービターPUF は Support Vector Machine(SVM)による機械学習攻撃によって脆弱であることが報告されている．この SVM は機械学習の一種であり，学習データから 2 つのクラスを識別する超平面をもとめることができる[7,8]. アービターPUF を例に機械学習攻撃の原理を説明する．セレクタ段数がN段のアービターPUF において，m 段目のセレクタのチャレンジ c_n^m が 0 の時の遅延時間差を δ_m^0，チャレンジが 1 の時の遅延時間差を δ_m^1 とする．このとき SVM でのアービターPUF の遅延時間のモデルを次式で表す．

$$\left.\begin{aligned} w^1 &= \frac{\delta_1^0 - \delta_1^1}{2}, w^i = \frac{\delta_{i-1}^0 + \delta_{i-1}^1 + \delta_i^0 - \delta_i^1}{2}, \\ w^{N+1} &= \frac{\delta_N^0 + \delta_N^1}{2} (i = 2, \cdots, N) \end{aligned}\right\} \quad (1)$$

またチャレンジ $\vec{C}_n = c_n^1, \cdots, c_n^N$ としたとき，チャレンジのモデルを次式で表す．

$$\phi^i = \prod_{l=i}^{N} (1 - 2c_n^l), \phi^{N+1} = 1 (i = 1, \cdots, N) \quad \cdots (2)$$

式(1)(2)よりセレクタチェーンで生じる遅延時間差は $\vec{w}^T \vec{\phi}_n$ で表すことができる．例えば，セレクタ段数 3 段のアービターPUF にチャレンジ 001 を与えた時，$\vec{w}^T \vec{\phi}_n = -\delta_1^0 - \delta_2^0 + \delta_3^1$ となり，アービター回路までの遅延時間差が正しく導出できることがわかる．アービター回路の出力レスポンス 0 のときを $t_n = -1$，レスポンス 1 のときを $t_n = 1$ と表現すると，アービター回路は sgn 関数で表すことができる．sgn 関数は括弧内が正のとき 1，負のとき-1 を出力する関数である．以上より，アービターPUF の動作は次式で示すことができる．

$$t_n = \mathrm{sgn}(\vec{w}^T \vec{\phi}_n) \quad \cdots (3)$$

機械学習攻撃では，窃取して集めた CRP より $\vec{\phi}_n$ と t_n をそれぞれ計算し，SVM により遅延時間モデル $\vec{w}$ を予測する．正しい遅延時間 $\vec{w}$ が導出できれば，アービターPUF の動作が計算できるので，あるチャレンジ c_n に対するレスポンス t_n を予測することも可能となる．

RG-DTM PUF の場合，図 2 に示すような遅延時間差検出方式にある時間差領域を分割する概念がないため，sgn 関数で動作を表すことができない．そのため，既存の機械学習攻撃では攻撃できない．機械学習攻撃に対す

る耐性をシミュレーションにより示すため，RG-DTM PUF に対して SVM を用いた機械学習攻撃を行った結果を図 3 に示す．ここでは学習データ用として 100,000 個の CRP を生成，それとは別にチャレンジに対して出力されるレスポンスの正解率（予測率）を計算するために 10,000 個 CRP を生成した．遅延時間差検出における分割数が 2 分割，4 分割のとき，学習 CRP 数の増加に伴って予測率が高くなっているのがわかる．特に 2 分割は 100,000CRP もあれば，98%以上と殆ど正確に出力レスポンスが予測できる．しかし，8 分割であれば学習 CRP 数に関わらず，予測率が 50%とランダムにレスポンスを選択する確率となり機械学習攻撃が成功しないことがわかる．

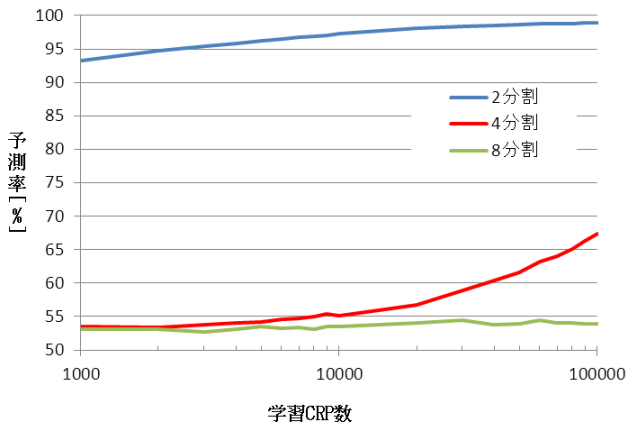


図 3　RG-DTM PUF に対する学習攻撃
(シミュレーション)

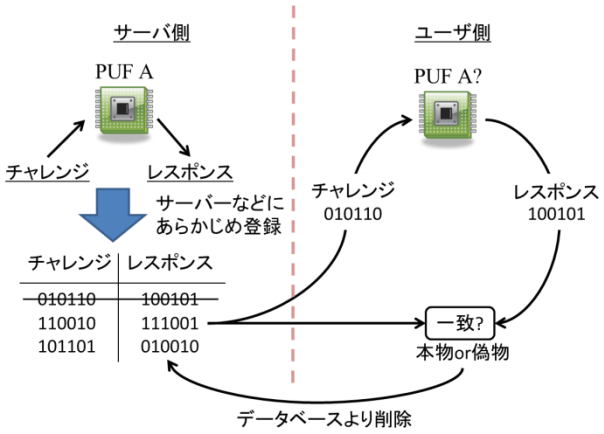


図 4　データベースを用いた PUF の認証方式

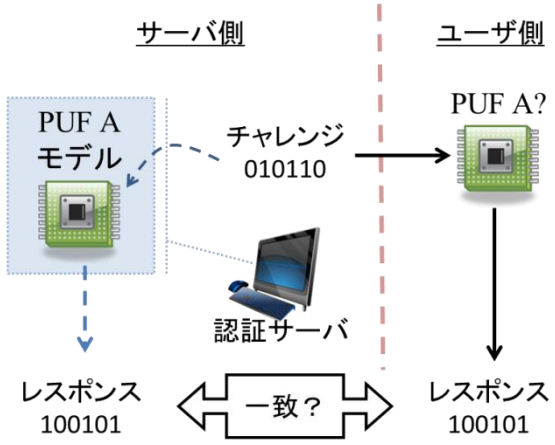


図 5　モデルを用いた認証方式

3　RG-DTM PUF の機械学習を用いた認証

3.1　機械学習を用いたモデルによる認証

PUF を用いた簡易認証方式は図 4 に示すように，あるチャレンジを PUF に与えて得られたレスポンス列があらかじめ認証サーバに保存しておいた CRP データベースと一致するかを調べて正誤判定する．そして，悪意のあるものが通信を傍受した場合に備え，一度認証に使用した CRP は以降使用しないようにする．そのため，N 段の RG-DTM PUF の全ての CRP を使用する場合，2^N もの CRP を記憶するための膨大な記憶容量が必要となる．そこで，図 5 のように CRP のデータベースの代わりに，機械学習により生成した PUF モデルを認証サーバに保存し，PUF と PUF モデルの両方に同じチャレンジを入力して，出力されたレスポンス同士の一致度から認証することを試みた．このとき，認証サーバには PUF モデルのパラメータのみを保存すれば良いので，少ない記憶容量で全ての CRP が認証に使用することができる．

3.2　RG-DTM PUF の認証用モデルの生成

RG-DTM PUF は機械学習攻撃に対して耐性をもつ上，単純に CRP から PUF モデルを生成していては安全面に問題がある．そこで，認証時には出力されない RG-DTM PUF の内部遅延情報を使用した PUF モデルの生成方法を提案する．使用する内部遅延情報は遅延時間差がどの区間だったかの詳細情報である．レスポンスは 0/1 に割り当てられた区間のどちらに入っているかを検出しているのに対して，PUF モデルを生成する際には図 6 に示すように遅延時間差が何番の領域だったか調べ，領域番号を内部遅延情報とする．

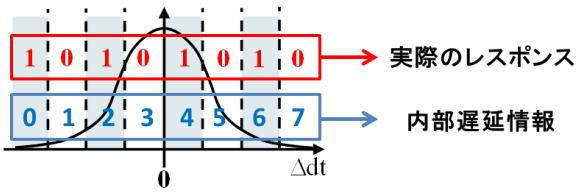


図 6　内部遅延情報

次に，内部遅延情報の取得方法について述べる．時間差検出型アービターは図 7 に示すようなセンスアンプの両端にキャパシタが接続された回路を使用している．両端のキャパシタをオン／オフすることで 0/1 の判定基準にオフセットを設け，遅延時間差がどの区分かを検出する．図 8 に示すように，まず初期状態（全てのキャパシタがオフ）の出力が”1”だったとする．この結果より遅延時間差情報が正方向であるのがわかる．次に，キャパシタをオンさせて 0/1 の判定基準にオフセットを設けて出力を得る．オフセットを順番に変更し続ければ，出力が”1”から”0”に変わるので，出力が変化した時のオフセット量から図 6 の内部遅延情報は簡単に得ることができる．

このような複数領域に分けられた内部遅延情報が得られればマルチクラスの機械学習が適用可能となる．簡単にマルチクラスを用いれば PUF モデルが生成できることを示す．図 9 に示すように 4 つのクラスに分割する場合，下記の 3 つの sgn 関数で表される分類器を用いれば良い．

$$t_{A\_n} = \mathrm{sgn}(\vec{w}^T \vec{\phi}_n)$$
$$t_{B\_n} = \mathrm{sgn}(\vec{w}^T \vec{\phi}_n + \alpha)$$
$$t_{C\_n} = \mathrm{sgn}(\vec{w}^T \vec{\phi}_n - \alpha)$$

どのクラス（0～3）に該当するかはt_{A_n}，t_{B_n}，t_{C_n}の組み合わせで示すことでき，該当するクラス情報とチャレンジより遅延時間$\vec{w}$と分類境界αを予測する．内部遅延情報はクラス（0～3）を意味するのでマルチクラスの機械学習を用いれば RG-DTM PUF のモデル化が行える．

このモデル生成時のみ内部遅延情報の出力を行い，認証に使用する時には 1/0 のレスポンスのみを出力することで，機械学習攻撃への高い耐性を保ちつつ，モデルを用いた認証が可能となる．

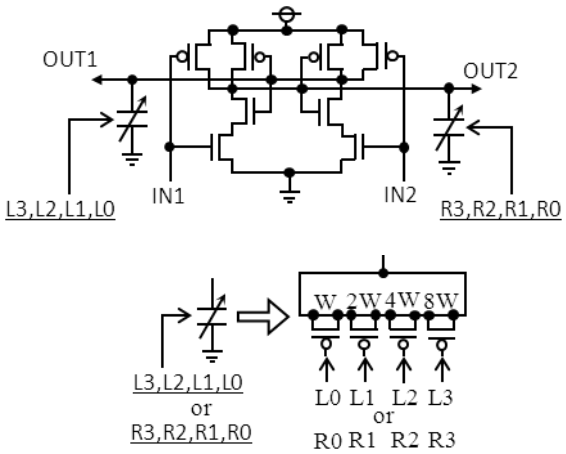


図 7　遅延時間差検出型アービター回路

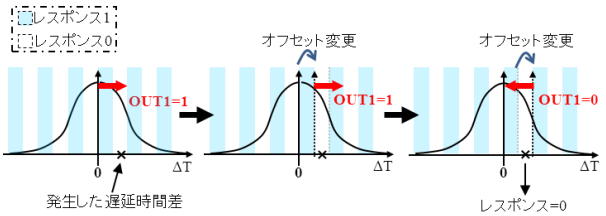
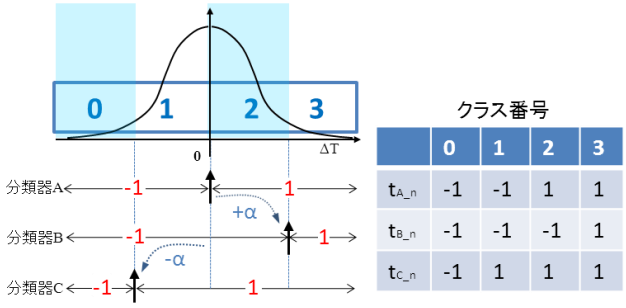


図 8　内部遅延情報の生成



	0	1	2	3
t_{A_n}	-1	-1	1	1
t_{B_n}	-1	-1	-1	1
t_{C_n}	-1	1	1	1

図 9　マルチクラス機械学習におけるモデル例

3.3　シミュレーションによる認証結果

提案手法を用いて RG-DTM PUF モデルが生成できるかを示すために，セレクタ段数 32 段，分割数 8 の RG-DTM PUF のシミュレーションデータを使ってモデル化を行い，認証評価を行った．モデル化はマルチクラス SVM を用いた機械学習により生成した．評価結果を図 10 に示す．RG-DTM PUF モデルが生成した CRP が正しく予測できる確率（予測率）は学習数 100,000 個で約 95%に達しているのがわかる．この結果は 1 ビットのレスポンスの正誤確率なので，256 ビット長の ID を生成したときには 5%の 13 ビット程度は誤認すると考えられる．そのため，認証に使用する際には誤りビットを許容する必要がある．256 ビット長の ID を使用したとき，実際に何ビット許容する必要があるのかを調べた結果を図 11 に示す．正しい RG-DTM PUF と誤認を調べるために 9 個の異なる RG-DTM PUF を用いて評価を行った．横軸は許容する誤りビット数を示し，縦軸は 1000 回認証したときの平均認証率を示す．誤り許容ビットを 21 ビットのとき，正しい RG-DTM PUF は認証率 100%となった．誤り許容ビットを 98 ビット以上にすると異なる RG-DTM PUF も認証誤って認識してしまうことがわかった．つまり，誤り許容ビットの閾値を 21～98 ビットの間に設定していれば，正しく認証できることがわかった．

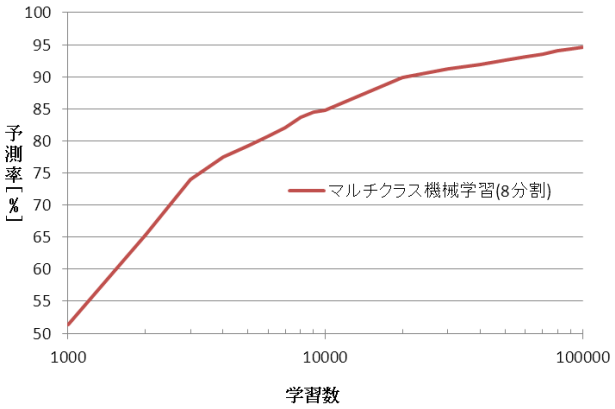


図 10　RG-DTM PUF に対するマルチクラス学習攻撃（シミュレーション）

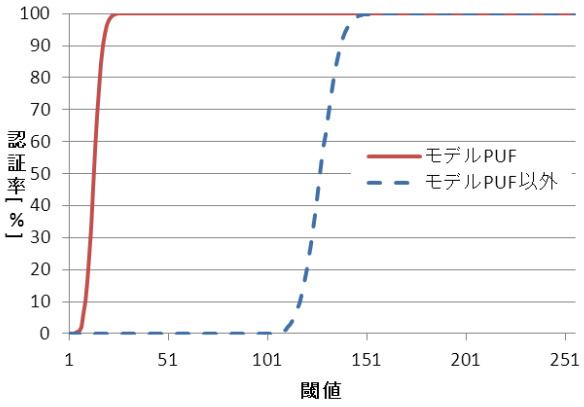


図 11　RG-DTM PUF のモデルによる認証率（シミュレーション）

4　試作チップを用いた認証評価

シミュレーションデータにより認証評価を行ったが，実測では環境変化等により不安定なレスポンスが存在する．そこで RG-DTM PUF のモデルによる認証を，180nm CMOS プロセスで試作したチップを用いて認証評価を行った．使用した RG-DTM PUF のセレクタ段数は 32 段，PUF モデルは SVM を用いた機械学習によって生成した．

4.1　分割数と機械学習

まず始めに，今回使用した 9 個のチップ全てに対して SVM を用いた機械学習攻撃を行い，攻撃耐性を評価した．評価結果を図 12 に示す．学習用に 100,000 個の CRP，それとは別に予測率を計算するために 10,000 個の CRP を生成した．分割数 4 のときは最小予測率 64%，最大予測率 92%まで達し，9 個全てのチップに対して機械学習攻撃ができることがわかった．一方，分割数 8 であれば全てのチップでおよそ予測率 50%となり，機械学習攻撃に対して耐性を持っていることがわかった．結果より機械学習攻撃に耐性を持たせるには 8 分割以上が必要であることがわかった．

次に，マルチクラス SVM を用いた機械学習により PUF モデルを生成した．分割数は機械学習攻撃結果を受けて 8 分割を利用した．最もモデル生成が困難で予測率が低かった結果を図 13 に示す．学習数が 100,000 個で最低でも予測率約 85%に達しており，シミュレーション結果ほどではないにしても，概ね認証できると考えられる．

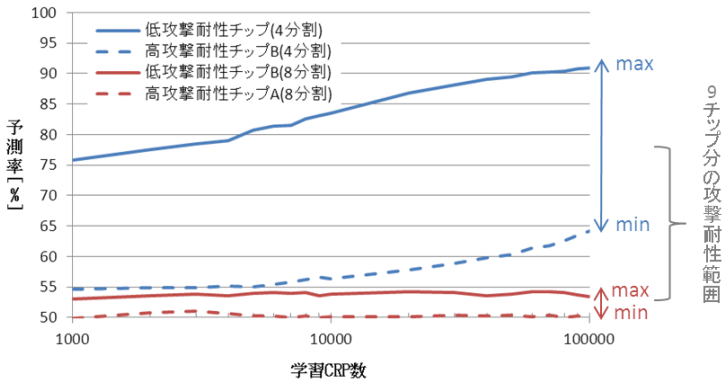


図 12　4,8 分割 RG-DTM PUF に対する学習攻撃

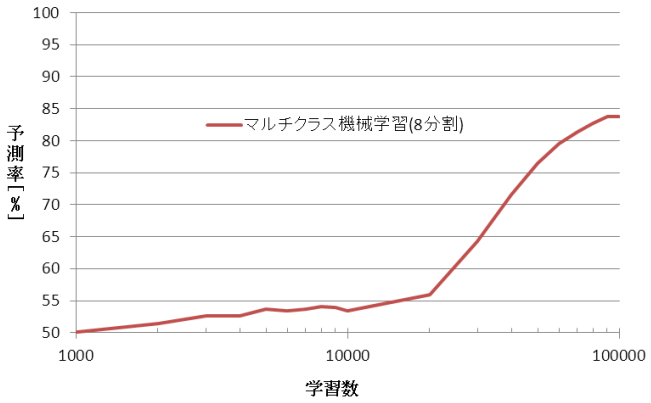


図 13　8 分割 RG-DTM PUF マルチクラス機械学習

4.2　誤り許容ビット数と認証率

9 個の実チップを用いて認証率を調査した．全てのチップにランダムに生成した 256 種類の同一チャレンジ信号を与えて，256 ビットの ID を生成した．異なるチップの ID 間のハミング距離を計算したユニーク性と，同一チャレンジを 100 回入力して得られた ID 間のハミング距離を計算した再現性を評価した結果を図 14 に示す．この時，各チップの正しい ID はマルチクラスの機械学習から生成した PUF モデルから出力される ID とした．ユニーク性は平均 126.88 ビット，標準偏差 8.81 ビットとなり，理想値である平均 128 ビット，標準偏差 8 ビッ

トとほぼ一致したが，再現性は平均 65.1 ビットと理想値 0 ビットより少し大きい結果となった．

次に，得られたユニーク性と再現性の結果より，本物を偽物とする確率(False Negative Rate：FNR)と，偽物を本物とする確率(False Positive Rate：FPR)を算出した．その結果を図 15 に示す．誤り許容ビット数を 71 ビット近辺にすれば辛うじて FNR と FPR の両方が 0.001ppm となるので，認証に使用するときは誤り許容ビット数を 71 ビット程度に設定すれば良い．過去に RG-DTM PUF 単体の性能を評価したとき，これらの結果もより良いデータが得られていたが，マルチクラス機械学習により生成された PUF モデルの誤認率も含んでいるために FNR と FPR が高くなっていると考えられる．

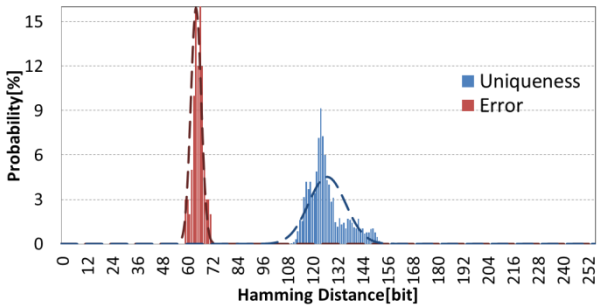


図 14　ユニーク性と再現性

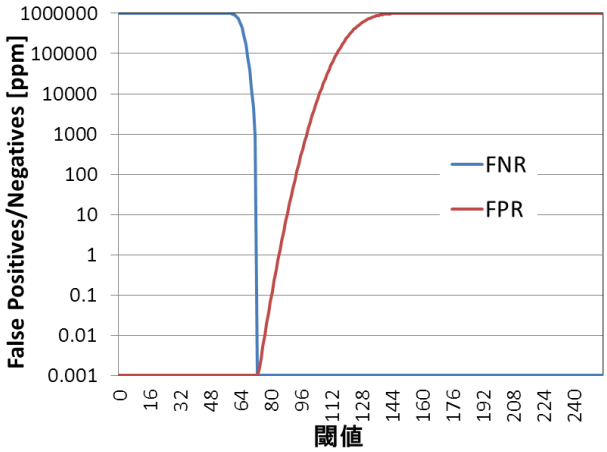


図 15　FNR と FPR

実際に 9 チップ分の RG-DTM PUF モデルと各チップから出力された ID を用いて認証テストを行った．誤り許容ビットと認証率の関係を示した代表結果を図 16 に示す．認証試行回数は 1,000 回である．誤り許容ビット数 59 以上で正しいチップは 100%の確率で認証できることがわかった．また，誤り許容ビット数 98 以上にすると異なるチップも誤認してしまう．そのため，誤り許容ビットの閾値 59～97 ビットであることがわかり，図 15 で得られた 71 ビットが妥当な結果であることが確認できた．そして，誤り許容ビットを 71 ビットに設定して 9 個のチップ全てに対して認証テストを行った結果を表 1 にまとめる．RG-DTM PUF モデル（行方向）と実チップ（列方向）が一致している場合だけ認証率 100%となり，異なるチップを誤認する確率は 0%となり，RG-DTM PUF のモデルを用いた認証ができていることが確認できた．

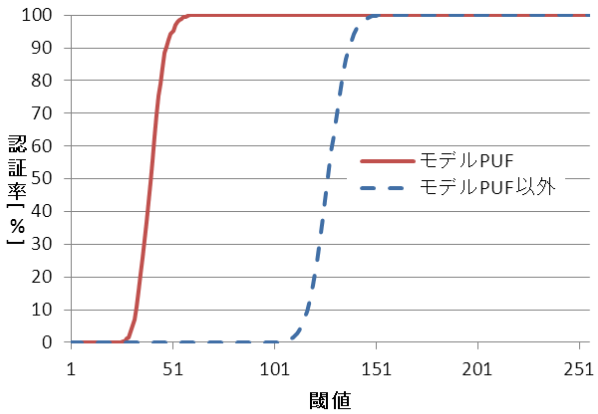


図 16　RG-DTM PUF モデルを用いた認証率

表 17　閾値 71 における認証率[%]

チップ	RG-DTM PUF モデル								
	A	B	C	D	E	F	G	H	I
A	100	0	0	0	0	0	0	0	0
B	0	100	0	0	0	0	0	0	0
C	0	0	100	0	0	0	0	0	0
D	0	0	0	100	0	0	0	0	0
E	0	0	0	0	100	0	0	0	0
F	0	0	0	0	0	100	0	0	0
G	0	0	0	0	0	0	100	0	0
H	0	0	0	0	0	0	0	100	0
I	0	0	0	0	0	0	0	0	100

5　まとめと今後

機械学習攻撃への耐性が高く，少ない記憶容量で実現できる RG-DTM PUF のモデルを用いた認証を提案した．セレクタ段数 32 段，8 分割 RG-DTM PUF の機械学習攻撃に対する耐性を示した上で，内部遅延情報とマルチクラスを用いた機械学習を用いて PUF モデルを用いれば，誤り許容を 21～98 ビットで正しい認証ができることを示した．また，180nm CMOS プロセスを用いて試作した RG-DTM PUF チップから RG-DTM PUF モデルと 256 ビットの ID を生成して FNR，FPR を算出したところ，誤り許容ビットを 71 ビットのときに 0.001ppm と最も誤認率が低いことがわかった．そこで，9 個の実チップに対して誤り許容ビット 71 で認証テス

トをしたところ，100%正しい認証ができていることが確認できた.

今回，SVM のみで PUF モデルの生成を検討したが，今後は他の機械学習を適用した PUF モデルの比較検討を行っていく.

謝辞

本研究は JST，CREST「ディペンダブル VLSI システムの基盤技術」の一環として行われた．180nm CMOS プロセスでのチップ試作は東京大学大規模集積システム設計教育研究センターを通じてローム(株)の協力で行われた．関係各位に感謝いたします.

参考文献

[1] R. Pappu, “Physical One-Way Functions,” PhD thesis, Massachusetts Institute of Technology, March 2001. Available at http://pubs.media.mit.edu/pubs/papers/01.03.pappuphd.powf.pdf

[2] G. Edward Suh, and Srinivas Debadas, “Physical Unclonable Functions for Device Authentication and Secret key Generation,” Annual ACM IEEE Design Automation Conference, pp.9-14, 2007.

[3] B. Gassend, D. Clarke, M. van Dijk, and S. Devadas, “Silicon Physical Random Functions,” CCS2002, pp.148-160, 2002.

[4] Jae W. Lee, D. Lim, B. Gassend, G. E. Suh, M. van Dijk, and S. Debadas, “A Technique to Build a Secret Key in Integrated Circuits for Identification and Authentication Applications,” In Proceedings of the IEEE VLSI Circuits Symposium, pp.176-179, 2004.

[5] U. Ruhrmair, F. Sehnke, J. Solter, G. Dror, S. Devadas, J. Schmidhuber, ” Modeling Attacks on Physical Unclonable Functions,” in Proceedings of ACM Conference on Computer and Communications Security, pp.237-249, 2010.

[6] Furuhashi Kota, Shiozaki Mitsuru, Fukushima Akitaka, Murayama Takahiko, Fujino Takeshi, “The arbiter-PUF with high uniqueness utilizing novel arbiter circuit with Delay-Time Measurement,” ISCAS2011, pp.2325-2328, 2011.

[7] V. Vapnik, ”Statistical Learning Theory,” Wiley, New York. 1998.

[8] C. M. Bishop et al., ”Pattern Recognition and Machine Learning,” Springer New York:, 2006.