様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １３ 日現在

機関番号：３３９１９
研究種目：研究活動スタート支援
研究期間：2019～2021
課題番号：１９Ｋ２４３５７

研究課題名（和文）多重防御機構を備えたセキュアで騙されないAIエンジンの開発

研究課題名（英文）Development of a secure and non-fooled AI engine with multiple defense mechanisms

研究代表者
野崎　佑典（Nozaki, Yusuke）

名城大学・理工学部・助教

研究者番号：６０８４７９５３

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：よりよい社会を実現するためにAI技術の利用が注目されている。AIの社会実装を進める上で，AI利用時の安全性を確保することは非常に重要である。そこで本研究では，セキュアなAIシステムを実現するための研究開発を行った。主に，モデル抽出攻撃とEvasion攻撃について研究開発を進め，それぞれどのような攻撃によってどのような脅威が生じるかを明らかにし，対策技術の研究開発を進めた。これらの研究開発によって，AIシステムの安全性を向上させた。

研究成果の学術的意義や社会的意義
研究成果の学術的意義は，実デバイスに実装したAIエンジンでのいくつかの攻撃に対する脅威を明らかにした点と，開発した対策手法によって，AIエンジンに対する攻撃への安全性を向上できる点である。また，本研究で研究開発した手法によって， AIシステムのセキュリティを向上させることができるため，AIの社会実装を推進することができる点で社会的意義が大きい。

研究成果の概要（英文）：To realize a better society, the AI technology has been attracted attention. It is important to ensure the safety of AI to promote social implementation of AI. Therefore, this study conducted research and development to realize a secure AI system. This study revealed the threats of model extraction attacks and evasion attacks and researched the countermeasure methods. The proposed methods improved the security of AI system.

研究分野：情報セキュリティ

キーワード：AIエンジン　セキュリティ

## １．研究開始当初の背景

Society5.0 の実現のために Artificial Intelligence（AI）技術が注目されている。一方で，AI の社会実装における課題として，AI を使用することの安全性とセキュリティリスクが挙げられている。AI の安全性に関して，AI の誤判断は現実世界へ与える影響が大きく騙されない AI が必要である。AI のセキュリティリスクに関して，AI で利用する学習データには個人情報や企業機密が含まれるだけでなく，モデル生成は豊富な計算資源を必要とするため，モデル情報を保護するためのセキュアな AI が必要である。実際に AI に対して，AI の誤判断を誘発させる Evasion 攻撃や，モデル情報を複製するモデル抽出攻撃が報告されている。また，近年では AI 動作時の処理時間や漏洩電磁波などを用いた攻撃も報告されており，攻撃技術はより高度化している。一方で，これらのサイドチャネル情報を利用した攻撃に対する対策手法の研究は進んでいない。

## ２．研究の目的

本研究では，実デバイスに実装することを想定したセキュアな AI エンジンを新たなに開発する。この AI エンジンでは，エッジデバイスの耐タンパ技術と Physically Unclonable Function（PUF）技術を組み合わせた防御機構を構築することで，攻撃への安全性を向上させる。本研究によって，AI 利用におけるセキュリティリスクを低減させ，AI の社会実装を推進する。このように，本研究では AI の社会実装を推進させることで，Soceity5.0 の実現（よりよい社会の実現）に貢献する。

## ３．研究の方法

セキュアな AI システムを実現するために，本研究では耐タンパ AI 実装と PUF 指向 AI 実装を開発する。

耐タンパ AI 実装に関しては，まず対策について検討するために AI がどのような攻撃に対して脆弱であるかを明らかにする必要がある。そのため，AI デバイスでのサイドチャネル情報を利用したモデル抽出攻撃の研究開発を進める。また，攻撃原理を明らかにした後は，その攻撃に対する対策手法を開発する。この開発手法では，これまでに開発してきたエッジデバイス向けの様々な耐タンパ技術をベースとし，AI デバイスを指向した耐タンパ実装を開発する。

PUF 指向 AI 実装では，AI に誤認識を誘発させるような攻撃の脅威を明らかにするために，攻撃原理の解明を進める。そして解明した攻撃手法に合わせた対策技術の開発を進める。具体的には，これまでに検討を進めてきた AI デバイスと PUF 技術を併用させたアーキテクチャをベースに対策技術を確立する。

## ４．研究成果

全体として，当初の目標をおおむね達成することができた。まず，耐タンパ AI 実装については，AI として Multi-Layer Perceptron（MLP）を対象に実装実験を行った。その結果，消費電力を利用したモデル抽出攻撃に対して脆弱であることを明らかにした。また，対策手法に関しては，実際の AI の演算とその時に生じるサイドチャネル情報の相関について，時間軸方向での攪乱を引き起こすシャッフリング対策を開発した。開発した対策手法を図 1 に示す。図 1 に示すように，AI の積和演算の乗算部分に関して，乱数によってその計算順序をランダム化する対策を開発した。対策アーキテクチャに対する攻撃結果を図 2 に示す。図 2 に示すように，重み情報（モデル情報）の推定において，正解値である「5」の相関値は小さくなっており，正しく値が推定できていないことが確認できる。このように，対策によって攻撃時の相関値を低下させ，モデル抽出攻撃に対する耐性を向上させることに成功した。

また，PUF 指向 AI 実装については，AI デバイスでの脅威として，AI に誤認識させる攻撃に関する研究を進め，AI ハードウェアでのトロイ攻撃の脅威を明らかにした。この攻撃では，予めハードウェアの構成要素に対して，誤った推論を行うような回路を混入させる。この回路は内部情報を書き換えることで実現するため，回路オーバーヘッド無く実現することができる。また，機能テストでも発見することは難しく，攻撃者が設定したある入力トリガによって，攻撃者が意図した動作を行う。本研究で明らかにした脅威では，画像認識システムを対象に，攻撃者が設定したノイズ情報を含む入力画像が与えられたときに，認証システムをパスするような攻撃を実施した。実装実験では，対象の回路を Field Programmable Gate Array（FPGA）に実装して，その脅威を定量的に評価した。また，対策手法として，Neural Network（NN）と PUF を併用させたアーキテクチャである NN PUF に関する検討を進めた。具体的には NN PUF に関して，様々なデバイスでの実装評価や安全性評価に加えてユニーク性を向上させるための実装手法について研究開発した。また，PUF を用いた認証システムにおける安全性を向上させるための認証方式も開発した。

これらの研究成果については，関連する国内の研究会や国際会議で発表するだけでなく，査読付学術論文誌で公開した。

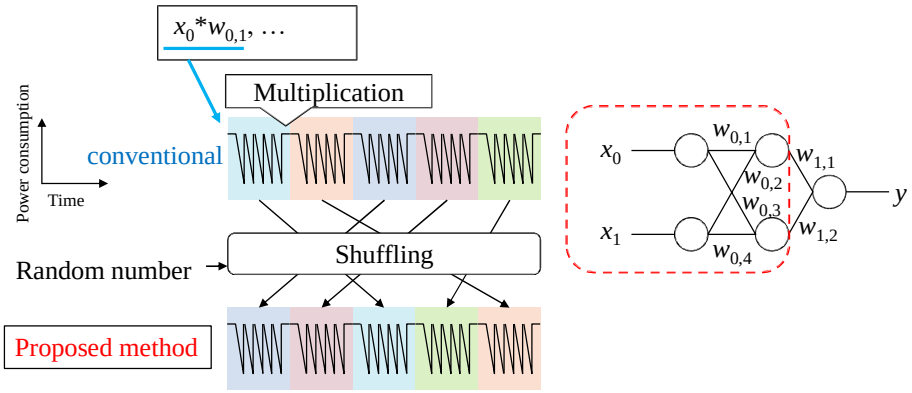


**図 1　耐タンパ AI 実装**

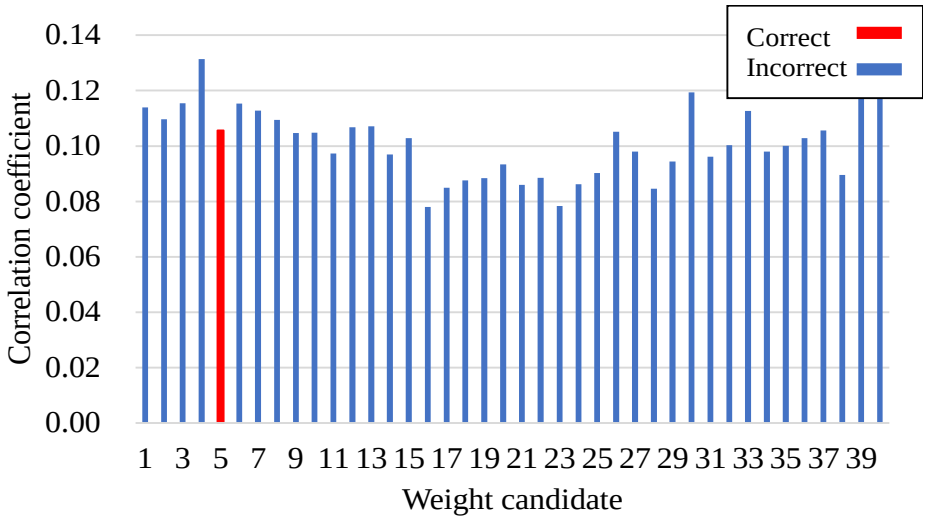


**図 2　サイドチャネル情報を利用したモデル抽出攻撃に対する安全性評価結果**

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

| | |
|---|---|
| １．著者名<br>野崎佑典，吉川雅弥 | ４．巻<br>63 |
| ２．論文標題<br>秘密分散法を利用したPUFのセキュア認証方式とその評価 | ５．発行年<br>2022年 |
| ３．雑誌名<br>情報処理学会論文誌 | ６．最初と最後の頁<br>840～848 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.20729/00217481 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>野崎佑典，竹本修，池崎良哉，吉川雅弥 | ４．巻<br>141 |
| ２．論文標題<br>LUTベースのAIデバイスに対する階層的なハードウェアトロイとその評価 | ５．発行年<br>2021年 |
| ３．雑誌名<br>電気学会論文誌C | ６．最初と最後の頁<br>1234～1240 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1541/ieejeiss.141.1234 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　6件）

| |
|---|
| １．発表者名<br>Y. Nozaki and M. Yoshikawa |
| ２．発表標題<br>Shuffling Countermeasure against Power Side-Channel Attack for MLP with Software Implementation |
| ３．学会等名<br>2021 IEEE the 4th International Conference on Electronics and Communication Engineering (ICECE 2021)（国際学会） |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>Y. Nozaki and M. Yoshikawa |
| ２．発表標題<br>Neural Network Based Glitch Physically Unclonable Function |
| ３．学会等名<br>2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON 2021)（国際学会） |
| ４．発表年<br>2021年 |

１．発表者名
Y. Nozaki and M. Yoshikawa

２．発表標題
Performance Evaluation of AI Authentication Device Implemented on SAKURA-G

３．学会等名
7th IEEE International Conference on Applied System Innovation (ICASI 2021)（国際学会）

４．発表年
2021年

１．発表者名
野崎佑典，竹本修，池崎良哉，吉川雅弥

２．発表標題
FPGA向けAIモジュールに対する電力解析の検討

３．学会等名
2021年電気学会電子・情報・システム部門大会

４．発表年
2021年

１．発表者名
野崎佑典，吉川雅弥

２．発表標題
NN PUFのユニーク性を向上させるレスポンス生成手法

３．学会等名
令和3年度電気・電子・情報関係学会東海支部連合大会

４．発表年
2021年

１．発表者名
野崎佑典，吉川雅弥

２．発表標題
軽量認証暗号SPARKLEの耐タンパ性評価

３．学会等名
情報処理学会CDS研究会

４．発表年
2021年

１．発表者名
野崎佑典，吉川雅弥

２．発表標題
多層パーセプトロンへの電力サイドチャネル対策の検討

３．学会等名
第50回東海ファジィ研究会

４．発表年
2021年

１．発表者名
野崎佑典，吉川雅弥

２．発表標題
SAKURA-Gに実装したAI向け認証デバイスの性能評価

３．学会等名
第49回東海ファジィ研究会

４．発表年
2021年

１．発表者名
Y. Nozaki, S. Takemoto, Y. Ikezaki, and M. Yoshikawa

２．発表標題
LUT oriented Hardware Trojan for FPGA based AI Module

３．学会等名
6th IEEE International Conference on Applied System Innovation (ICASI 2020)（国際学会）

４．発表年
2020年

１．発表者名
野崎佑典，竹本修，池崎良哉，吉川雅弥

２．発表標題
AI推論器のLUT構造に着目したハードウェアトロイ

３．学会等名
電子情報通信学会HWS研究会

４．発表年
2020年

１．発表者名
野崎佑典，竹本修，池崎良哉，吉川雅弥

２．発表標題
FPGA向けAI推論器に対するハードウェアトロイの検討

３．学会等名
令和2年度電気・電子・情報関係学会東海支部連合大会

４．発表年
2020年

１．発表者名
Y. Nozaki and M. Yoshikawa

２．発表標題
Tamper Resistance Evaluation of MLP with Software Implementation against Power Consumption based Model Extraction

３．学会等名
2020 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP’20)（国際学会）

４．発表年
2020年

１．発表者名
野崎佑典，吉川雅弥

２．発表標題
ソフトウェア実装したMLPに対する電力解析の検討

３．学会等名
第48回東海ファジィ研究会

４．発表年
2020年

１．発表者名
Y. Nozaki and M. Yoshikawa

２．発表標題
Tamper Resistance Evaluation of TWINE Implemented on 8-bit Microcontroller

３．学会等名
3rd International Conference on Software Engineering and Information Management (ICSIM 2020)（国際学会）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|