早稲田大学大学院 基幹理工学研究科

博 士 論 文 概 要

論 文 題 目

Hardware-Trojan Detection Methods Utilizing Machine Learning Based on Hardware-Specific Features
ハードウェア固有の特徴にもとづく機械学習を利用したハードウェアトロイ検出

申 請 者

Kento	HASEGAWA
長谷川	健人

情報理工・情報通信専攻 情報システム設計研究

2019 年 12 月

Hardware devices have been widely used in our world. For example, most of the people have smart phones which contain high-performance processors and various types of sensors. Motor vehicles are equipped with numerous numbers of hardware devices to perform advanced driver-assistance system. As for the industry field, the foundries producing semiconductors are highly automated with industrial robots. To describe the highly-automated industries by smart devices, the term 'Industry 4.0' is introduced by German government. Japanese government also promotes 'Society 5.0' where people resolve various society challenges by incorporating the technological innovations of the fourth industrial innovation. As exemplified above, people no longer lead highly convenient lives without hardware devices, and aim to develop information-oriented society more than ever before.

There exist potential risks at any steps on hardware design and manufacturing. The production process of hardware devices is roughly divided into two steps: the design step and manufacturing step. Hardware specification and circuit designs are determined at the design step. Due to the rapid and low-cost production, third-party intellectual properties (3PIPs) are often integrated to products. Since recent integrated circuits (ICs) are highly integrated and contain billions of gates, the modification by an attacker may be easily hidden inside the products. As for the manufacturing step, circuit designs are often fabricated and assembled at the overseas fabrications because the hardware market becomes globalized. Several fabrications may be untrusted, and a backdoor may be inserted into their products. As discussed above, there exist potential risks at any steps on hardware design and manufacturing. A malicious function inserted at a hardware circuit is often call as a 'hardware Trojan (HT).' How to detect hardware Trojans is a serious concern. Note that, existing hardware design and manufacturing processes typically include test steps, but the test processes just check the functionality and validity of the products. Since typical test processes do not take the threats of hardware Trojans into account, we have to develop a hardware-Trojan detection scheme. In this dissertation, we aim to find out how to tackle the threats of hardware Trojans.

Under the circumstances, security threats to the hardware devices have been pointed out. Threats on hardware devices are now raising and becoming reality. How to tackle the problem is a major concern in the IoT era.

The methodologies defeating hardware Trojans have been studied recent years. The methodologies can be classified into two categories: a prevention methodology and a detection methodology. With the prevention methodology, hardware designs are altered to be difficult to insert or further modify the circuit by third parties. Hardware logic encryption approaches thwart insertion of hardware Trojans by encrypting hardware designs. Physical unclonable functions (PUF) are often used to generate secret keys for logic encryption. Camouflaging (or obfuscating) approaches are also applied to protect hardware designs. The prevention methodology is effective to protect hardware designs so as not to modify them by third parties. On the other hand, the detection methodology aims to catch hardware Trojan circuits.

The detection approaches are further classified into two categories: a destructive approach and a non-destructive approach. The destructive approach generally adopts destructive reverse-engineering techniques to depackage an IC and performs optical analysis. Though this approach is useful to physically analyze the manufactured ICs, the tested ICs cannot be shipped anymore. Meanwhile, the non-destructive approach does not destruct ICs. Since the non-destructive approach just analyzes the design or manufactured ICs without destruction,

this approach can be easily integrated to existing design and manufacturing process. In this dissertation, we focus on the non-destructive approaches to defeat hardware Trojans.

As discussed above, we aim to defeat hardware Trojans adopting non-destructive approaches. Several non-destructive methods can be taken on the design or manufacturing step. The non-destructive methods on the design step analyze hardware designs including 3PIPs. Formal verification and code analysis are often taken. The non-destructive methods on the manufacturing step analyze manufactured hardware products. Functional tests and side-channel analysis are often taken. Most of the existing methods take model-based approaches, and therefore detectability of unknown threats has to be discussed.

Recently, machine learning has attracted the interest of researchers as a breakthrough in data mining, and it is expected to overcome security-related challenges. In fact, existing methods demonstrate that machine learning can be used to find out malicious behaviors. However, existing machine-learning-based methods require 'Golden model' where no hardware Trojan is definitely inserted. Developing a sophisticated machine-learning-based hardware-Trojan detection method is not only a challenging problem but also a promising research to realize highly automated society. The major problems to leverage machine learning for hardware Trojan detection are how to extract effective features to identify Trojan nets and normal nets, and how to detect hardware Trojans without Golden models.

In this dissertation, we leverage machine learning algorithms to the non-destructive hardware-Trojan detection methods. First, in Chapters 2, and 3, we aim to detect hardware Trojans at gate-level netlists utilizing machine learning algorithms such as support vector machine (SVM), neural networks (NNs), and random forests with extracting effective features from a net in gate-level netlists. This is the first work to leverage machine learning at gate-level netlists. Next, in Chapter 4, we aim to detect malicious behaviors based on power analysis utilizing one of the unsupervised machine learning algorithms. This is the first discussion on how to detect malicious behavior based on power analysis by anomaly detection algorithm.

In this dissertation, we propose machine-learning-based hardware-Trojan detection methods based on hardware-specific feature values. This dissertation is organized according to the following chapters.

Chapter 1 [Introduction] describes the backgrounds and overview of this dissertation.

Chapter 2 [Hardware Trojan Classification Utilizing Machine Learning] proposes a hardware-Trojan classification method at gate-level netlists to identify hardware-Trojan infected nets (or Trojan nets). In this chapter, we have a preliminary discussion on how to apply machine learning to hardware Trojan detection, and then we evaluate the effective feature values for hardware Trojan detection. As a preliminary discussion on the hardware Trojan detection at a gate-level netlist, we extract the five hardware-Trojan features from each net in a netlist. These feature values are complicated so that we cannot give the simple and fixed threshold values to them. Hence, we secondly represent them to be a five-dimensional vector and learn them by using SVM or NN. The experimental results with Trust-HUB benchmarks demonstrate that our method increases the true positive rate compared to the existing state-of-the-art results in most of the cases. Based on the preliminary discussion, we propose effective Trojan-net features for supervised machine-learning-based hardware-Trojan detection and their application to a

random forest classifier. We first propose 51 Trojan-net features which describe well Trojan nets. After that, we pick up random forest as one of the best candidates for machine learning and optimize it to apply to hardware-Trojan detection. Based on the importance values obtained from the optimized random forest classifier, we extract the best set of 11 Trojan-net features out of the 51 features which can effectively classify the nets into Trojan ones and normal ones, maximizing the F-measures. By using the 11 Trojan-net features extracted, our optimized random forest classifier has achieved at most 100% true positive rate as well as 100% true negative rate in several Trust-HUB benchmarks and obtained the average F-measure of 79.3% and the accuracy of 99.2%, which realize the best values among existing machine-learning-based hardware-Trojan detection methods.

Chapter 3 [Application of the Hardware-Trojan Detection Utilizing Machine Learning] proposes three applications of machine-learning-based hardware-Trojan detection. First, we propose a machine-learning-based hardware-Trojan detection method for gate-level netlists using multi-layer neural networks. We classify the nets in an unknown netlist into a set of Trojan nets and that of normal nets using multi-layer neural networks based on the 11 Trojan-net features proposed in Chapter 2. By experimentally optimizing the structure of multi-layer neural networks, we can obtain an average of 84.8% true positive rate and an average of 70.1% true negative rate while we can obtain 100% true positive rate in some of the benchmarks, which outperforms the existing methods in most of the cases. Second, we propose a Trojan-invalidating circuit, and implement it on an FPGA board. The implementation results demonstrate that the implemented Trojan-invalidating circuit successfully prevent from activating a hardware Trojan. Third, we propose a reinforcement of the hardware-Trojan detection utilizing machine learning. Since existing machine-learning-based hardware-Trojan detection methods are performed in the feature spaces, the proposed method considers boundary net structures between normal nets and Trojan nets and compensates the first machine-learning-based detection results based on them. The experimental results demonstrate that our proposed method successfully improve the detection results compared to the existing method.

Chapter 4 [Malicious Behavior Detection Based on Power Analysis] proposes an anomaly behavior detection method utilizing power analysis for low-cost micro-controllers. Our method accurately measures power consumption of the target device, and then classifies its waveform into the sleep-mode part, in which a micro-controller saves power, and into the active-mode part, in which a micro-controller works in a normal operation. After that, we obtain the duration time and consumed power from each active-mode period as feature values. Finally, we detect abnormal behavior based on the obtained feature values utilizing an outlier detection method without Golden models. We empirically evaluate the proposed method utilizing two micro-controllers, and the experimental results demonstrate that our proposed method successfully detects abnormal behaviors.

Chapter 5 [Conclusion] summarizes this dissertation and gives several future directions on machine-learning-based hardware-Trojan detection. In conclusion, we find out that hardware Trojan detection utilizing machine learning based on hardware-specific features has a future prospect. However, there still remain several tasks to be done. Enhancing the classification performance of hardware Trojan detection and implementation to the real world are our future works.

早稲田大学　博士（工学）　学位申請　研究業績書

氏 名　長谷川　健人　　印

（2019 年　12 月　現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
a. 論文 学術誌 原著論文	1. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, "Empirical Evaluation and Optimization of Hardware-Trojan Classification for Gate-Level Netlists based on Multi-Layer Neural Networks," IEICE Trans. on Fundamentals of Electronics, Communications and Computer Science, Vol. E101-A, No. 12, pp. 2320-2326, Dec. 2018. 2. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “Trojan-net Feature Extraction and Its Application to Hardware-Trojan Detection for Gate-level Netlists Using Random Forest,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Science, Vol. E100-A, No. 12, pp. 2857–2868, Dec. 2017. 3. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “A Hardware-Trojan Classification Method Using Machine Learning at Gate-level Netlists based on Trojan Features,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E100-A, No. 7, pp. 1427–1438, Jul. 2017.
c. 講演 国際会議	4. **K. Hasegawa**, R. Ishikawa, M. Nishizawa, K. Kawamura, M. Tawada, and N. Togawa, “FPGA-based Heterogeneous Solver for Three-Dimensional Routing,” in Proc. Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, China, Jan. 2020. 5. **K. Hasegawa**, K. Takasaki, M. Nishizawa, R. Ishikawa, K. Kawamura, and N. Togawa, “Implementation of a ROS-Based Autonomous Vehicle on an FPGA Board,” in Proc. International Conference on Field-Programmable Technology (FPT), Tianjin, China, Dec. 2019. 6. K. Nozawa, **K. Hasegawa**, S. Hidano, S. Kiyomoto, K. Hashimoto and N. Togawa, “Adversarial Examples for Hardware-Trojan Detection at Gate-Level Netlists,” in Proc. International Workshop on Attacks and Defenses for Internet-of-Things (ADIoT), Luxembourg, Luxembourg, Sep. 2019. 7. ○ **K. Hasegawa**, K. Chikamatsu and N. Togawa, “Empirical Evaluation on Anomaly Behavior Detection for Low-Cost Micro-Controllers Utilizing Accurate Power Analysis,” in Proc. IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 54-57, Rhodes Island, Greece, Jul. 2019. 8. M. Nishizawa, **K. Hasegawa**, and N. Togawa, "Capacitance Measurement of Running Hardware Devices and its Application to Malicious Modification Detection," in Proc. IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 362-365, Chengdu, China, Oct. 2018. 9. T. Inoue, **K. Hasegawa**, M. Yanagisawa, and N. Togawa, "Designing Subspecies of Hardware Trojans and Their Detection Using Neural Network Approach," in Proc. IEEE 8th International Conference on Consumer Electronics in Berlin (ICCE-Berlin), pp. 156-159, Berlin, Germany, Sep. 2018. 10. ○ (招待講演) **K. Hasegawa**, Y. Shi, and N. Togawa, “Hardware Trojan Detection Utilizing Machine Learning Approaches,” in Proc. IEEE International Workshop on Hardware Security and Trust (HSAT), pp. 1891-1896, Elizabeth, NJ, USA, Aug. 2018. 11. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “Detecting the Existence of Malfunctions in Microcontrollers Utilizing Power Analysis,” in Proc. IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 97-102, Platja d'Aro, Spain, Jul. 2018. 12. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, "A Trojan-invalidating Circuit Based on Signal Transitions and Its FPGA Implementation,” in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, Florence, Italy, May 2018. 13. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “A Hardware-Trojan Classification Method Utilizing Boundary Net Structures,” in Proc. IEEE International Conference on Consumer Electronics (ICCE), pp. 103-106, Las vegas, NV, USA, Jan. 2018. 14. T. Inoue, **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “Designing Hardware Trojans and Their Detection based on a SVM-based Approach,” in Proc. IEEE International Conference on ASIC (ASICON), pp. 811–814, Guiyang, China, Oct. 2017. 15. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “Hardware Trojans Classification for Gate-level Netlists Using Multi-Layer Neural Networks,” in Proc. IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 227–232, Thessaloniki, Greece, Jul. 2017.

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
	16. ○ **K. Hasegawa**, M. Yanagisawa, and N. Togawa, “Trojan-feature Extraction at Gate-level Netlists and Its Application to Hardware-Trojan Detection Using Random Forest Classifier,” in Proc. IEEE International Symposium on Circuits and Systems, pp. 2154–2157, Baltimore, MD, USA, May 2017. 17. ○ **K. Hasegawa**, M. Oya, M. Yanagisawa, and N. Togawa, “Hardware Trojans Classification for Gate-level Netlists based on Machine Learning,” in Proc. IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 203–206, Sant Feliu de Guixols, Spain, Jul. 2016.
国内学会	18. 西澤 誠人,長谷川 健人, 戸川 望, “非正規挿入デバイス検知のための電気容量監視装置とその実験的評価,” 信学技報, vol. 119, no. 260, HWS2019-66, pp. 53-58, 大阪市, Nov. 2019. 19. 野澤 康平, 長谷川 健人, 披田野 清良, 清本 晋作, 橋本 和夫, 戸川 望, “ニューラルネットワークを用いたハードウェアトロイ識別に対する敵対的サンプル攻撃の実証評価,” 信学技報, vol. 119, no. 260, HWS2019-64, pp. 41-46, 大阪市, Nov. 2019. 20. 長谷川 健人, 戸川 望, " IoT デバイス管理基盤の一考察," 電子情報通信学会 ソサイエティ大会, p. 139, 豊中市, Sep. 2019. 21. (査読あり) 長谷川 健人, 近松 聖, 戸川 望, "スリープ状態をもつ組込みシステムを対象とした電力解析にもとづく異常動作検知とその実証的評価," 情報処理学会 DA シンポジウム 2019 論文集, pp. 93-98, 加賀市, Aug. 2019. 22. (ポスター発表) 西澤 誠人, 石川 遼太, 長谷川 健人, 川村 一志, 多和田 雅師, 戸川 望, "配置配線のためのアンサンブルソルバシステム," 情報処理学会 DA シンポジウム 2019 ポスター発表, 加賀市, Aug. 2019. 23. (招待講演) 長谷川 健人, 野澤 康平, 披田野 清良, 清本 晋作, 橋本 和夫, 戸川 望, "ハードウェアセキュリティにおける AI 活用と攻撃," 電子情報通信学会総合大会, pp. SS-58-59, 新宿区, Mar. 2019. 24. 井上 智貴, 長谷川 健人, 戸川 望, "周辺ネットの特徴量を考慮した二段階のニューラルネットワークによるハードウェアトロイ検出手法," 情報処理学会研究報告, 西之表市, Mar. 2019. 25. 野澤 康平, 長谷川 健人, 披田野 清良, 清本 晋作, 橋本 和夫, 戸川 望, "ニューラルネットワークを用いたハードウェアトロイ識別に対する敵対的サンプル攻撃に関する一考察," 暗号と情報セキュリティシンポジウム予稿集, 大津市, Jan. 2019. 26. (査読あり) 西澤 誠人, 長谷川 健人, 柳澤 政生, 戸川 望, "低電力化電気容量検出装置を用いた動作中の不正デバイス検知," 情報処理学会 DA シンポジウム 2018 論文集, pp. 112-117, 加賀市, Aug. 2018. 27. (査読あり) 長谷川 健人, 柳澤 政生, 戸川 望, "マイクロコントローラのスリープ状態に着目した消費電力にもとづく悪意のある機能の発現検知," 情報処理学会 DA シンポジウム 2018 論文集, pp. 118-123, 加賀市, Aug. 2018. 28. (ポスター発表) 石川 遼太, 西澤 誠人, 長谷川 健人, 川村 一志, 多和田 雅師, 戸川 望, "ナンバーリンクソルバのための FPGA 協調システム," 情報処理学会 DA シンポジウム 2018 ポスター発表, 加賀市, Aug. 2018. 29. 井上 智貴, 長谷川 健人, 柳澤 政生, 戸川 望, “亜種ハードウェアトロイの設計とそのニューラルネットワークを用いた検出,” 信学技報, vol. 118, no. 83, VLD2018-36, pp. 173-178, 札幌市, Jun. 2018. 30. 長谷川 健人, 柳澤 政生, 戸川 望, “暗号回路に挿入されたハードウェアトロイとその抑止回路の FPGA 実装,” 信学技報, vol. 117, no. 273, VLD2017-53, pp. 139–144, 熊本市, Nov. 2017. 31. 井上 智貴, 長谷川 健人, 柳澤 政生, 戸川 望, “トリガ条件の異なるハードウェアトロイの設計と SVM を用いた検出,” 信学技報, vol. 117, no. 273, VLD2017-51, pp. 133–138, 熊本市, Nov. 2017.

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
	32. (査読あり) 長谷川 健人, 柳澤 政生, 戸川 望, “ネットの周辺情報を考慮した機械学習によるハードウェアトロイ識別,” 情報処理学会 DA シンポジウム 2017 論文集, pp. 127–132, 加賀市, Aug. 2017. 33. (ポスター発表) 長谷川 健人, 石川 遼太, 寺田 晃太朗, 川村 一志, 多和田 雅師, 戸川 望, “組込みデバイスと FPGA を用いたナンバーリンクソルバの設計と実装,” DA シンポジウム 2017, 加賀市, Aug. 2017. 34. 長谷川 健人, 柳澤 政生, 戸川 望, “ネットの特徴量を用いた多層ニューラルネットワークによるハードウェアトロイ識別,”情報処理学会研究報告, Vol. 2017-SLDM-179, No. 23, pp. 1–6, 久米島町, Mar. 2017. 35. (査読あり) 長谷川 健人, 柳澤 政生, 戸川 望, “Random Forest を用いたネットリスト特徴選択と機械学習によるハードウェアトロイ識別,” 情報処理学会 DA シンポジウム 2016 論文集, pp. 8–13, 加賀市, Sep. 2016. 36. (ポスター発表) 寺田 晃太朗, 長谷川 健人, 川村 一志, 多和田 雅師, 戸川 望, “機械学習と FPGA を用いたナンバーリンクソルバ,” DA シンポジウム 2016, 加賀市, Sep. 2016. 37. 長谷川 健人, 柳澤 政生, 戸川 望, “ニューラルネットを利用したネットリストの特徴にもとづくハード ウェアトロイ識別,” 信学技報, vol. 116, no. 94, VLD2016-7, pp. 1–6, 弘前市, Jun. 2016. 38. 長谷川 健人, 大屋 優, 柳澤 政生, 戸川 望, “SVM を利用したネットリストの特徴に基づくハードウェアトロイ識別,” 信学技報, vol. 115, no. 338, VLD2015-58, pp. 135–140, 長崎市, Dec. 2015.
e. その他雑誌記事	39. 長谷川 健人, 戸川 望, “スパイチップはあるのか - ハードウェアセキュリティの必要性,” 情報処理, vol. 60, no. 1, pp. 4-6, Jan. 2019. 40. 川村 一志, 長谷川 健人, 多和田 雅師, 戸川 望, “機械学習と FPGA を用いた配線問題解法への取り組み,” 情報処理, vol. 59, no. 3, pp. 228--231, Mar. 2018.
業績賞等	41. DA シンポジウム アルゴリズムデザインコンテスト 特別賞, DA シンポジウム 2019, Aug. 2019. 42. DA シンポジウム アルゴリズムデザインコンテスト 特別賞, DA シンポジウム 2018, Aug. 2018. 43. 第 33 回電気通信普及財団賞（テレコムシステム技術学生賞）最優秀賞, Mar. 2018. 44. 情報処理学会 山下記念研究賞, Mar. 2018. 45. IEEE CEDA All Japan Chapter Academic Research Award, Nov. 2017. 46. DA シンポジウム アルゴリズムデザインコンテスト 最優秀賞, DA シンポジウム 2017, Aug. 2017. 47. IEEE CEDA All Japan Chapter Academic Research Award, Aug. 2017. 他, 9 件
研究費・助成金	48. 早大理工総研－キオクシア(旧・東芝メモリ) 若手奨励研究, Sep. 2018 – Feb. 2020, 総額 50 万円. 49. 日本学術振興会 特別研究員奨励金, Apr. 2018 – Mar. 2020, 総額 210 万円 (2018 年度: 110 万円, 2019 年度: 100 万円). 50. 電気通信普及財団, 平成 29 年度 6 月期海外渡航旅費援助, Jun. 2017, 総額 29 万円.
特許	51. (発明者) 披田野 清良, 清本 晋作, 長谷川 健人, 戸川 望, (出願人) KDDI 株式会社, 学校法人早稲田大学, ``学習装置、学習方法及び学習プログラム", 特願 2019-140107, (出願日) 2019 年 7 月 30 日. 52. (発明者) 戸川 望, 長谷川 健人, (出願人) 学校法人早稲田大学, ``検出方法及び検出装置," 特願 2018-113649, (出願日) 2018 年 6 月 14 日.