Research on the Detection Method of HOG-SVM for Doping Modified Hardware Trojan

Chaolong Sun\textsuperscript{1,1*}, Lei Li\textsuperscript{1,2*}, Wanting Zhou\textsuperscript{1}, Xianjun Tan\textsuperscript{1}, Yuanhang He\textsuperscript{2}

\textsuperscript{1}UESTC, Chengdu, Sichuan, 611731, China
\textsuperscript{2}Science and Technology on Communication Security Laboratory, Institute of Southwest Communication, Chengdu, Sichuan, 610041, China

\textsuperscript{*}E-mail: aoyasun@163.com
\textsuperscript{*}E-mail: lilei_uestc@uestc.edu.cn

Abstract. With the globalization of the IC industry, hardware security has become an important topic, especially hardware Trojans (HT) hidden in ICs. In this paper, we presented the method based on histogram of oriented gradient (HOG) and support vector machine (SVM) to detect doping modified HTs, which is generally inserted by modifying the layout of some standard cells and is very difficult to detect. In order to address this problem, this paper firstly takes the standard layout without changing the doping polarity as the Trojan free (TF) layout, and regards the modified layout with changing the doping polarity as the HT layout, and then classifies the experimental data set into HT or TF by extracting the HOG features and using SVM training classification method. Experimental results demonstrate that the experimental accuracy of this method can reach 92.8%.

1. Introduction

With the rapid development of integrated circuit (IC) design and manufacturing technology, the increasingly popular third-party technical services on the one hand reduce the cost of chip fabrication and shorten the period of putting them into the market, on the other hand increase the risk of chip security performance and reduce the reliability of ICs. Under the condition of IC design and manufacturing globalization, some additional logic units of the malicious circuit might be inserted in ICs, which are not defined in the original design specifications, also known as Hardware Trojan (HT), in the process of IC design and manufacturing. Yang et al. [1] verified that remote attackers can realize remote control of user's electronic products through preinstalled HTs. The malicious design and manufacture of HT can change or increase some functions of integrated circuits, while relying on such electronic devices with HTs will bring different degrees of loss to users. According to the implementation level, HTs can be divided into system level, register transfer level (RTL) and physical layout level. These types of HTs generally change the structure of circuits and transistors. Among these three types, there is a kind of HT that changes the doping polarity of the transistors [2]. Since it is difficult to be detected by the traditional detection methods [3-6] including bypass analysis and function test, it is regarded one of the most difficultly detected HTs [2].
Some researchers proposed to use advanced optical imaging system to build HT detection systems based on reverse engineering (RE). Bao [7] proposed an innovative and robust RE scheme to identify the Trojan free (TF) ICs and applied K-means clustering to RE-based Trojan detection approach and the accuracy of 83.6% can be obtained. Nasr [8] presented an automatic and robust solution for the step of layout identification and used a machine learning technique to learn weak Haar-like features to simplify steps with the accuracy of 83.3%.

In this paper, we presented a histogram of oriented gradient (HOG) and support vector machine (SVM) detection method from the view of layout to detect doping modified HTs. Firstly, we chose the HT that changes the doping polarity to make layout pictures. Secondly, machine learning method is used to extract the HOG features in the layout. Finally, we used SVM method to realize two classifications, distinguish the HT and TF pictures in the dataset, which can achieve 92.8% accuracy.

2. Conceptual and theoretical analysis

In this chapter, we will focus on the HT model which changes the doping polarity in the active area, and the machine learning algorithm which extracts the HOG’s features of pictures and classifies them by SVM.

2.1. Hardware Trojan model

The HT model refers to the hardware model with specific Trojan characteristics. According to Becker G T’s [2] theory of changing doping polarity, this paper selects the standard logic units commonly used in SMIC13 process library, and redesigns some standard HT logic units for HT detection experiments.

(a) Unmodified (b) Trojan

Figure 1. Figure of an unmodified inverter and of a Trojan inverter [2].

Changing the doping polarity type of HT is not to change the material type and structure inside the transistor, but by changing the dopant polarity of the active area and then change the function of the transistor to keep its output at a certain value. Take the inverter as an example. The layout of the standard inverter is shown in Figure 1 (a), which is composed of one n-MOS and one p-MOS. The cross section of the standard n-MOS and p-MOS transistors is shown in Figure 2. In order to make the inverter output always high, for p-MOS, the p-dopant(positive dopant) mask of this p-MOS section is changed by the n-dopant(negative dopant) mask, making drain and VDD connected, as shown in Figure 2 (a). For n-MOS, change the source's n-dopant to p-dopant to form an inverse PN junction to disconnect drain from VSS, as shown in Figure 2 (b). Then the output of Trojan inverter is equivalent to directly connecting with VDD. The modified Trojan inverter is shown in Figure 1(b). We can also modify other standard logical units to build a HT model by this method.

2.2. Machine learning algorithm

In this chapter, we will use the HOG feature extraction algorithm to extract the features of the layout pictures, and then use SVM classification algorithm to classify the picture.
2.2.1. Extraction of HOG features. The HOG algorithm is proposed by Dalal et al. [9]. The core idea is that for a picture, the directional density distribution of its gradient or edge can well describe the characteristics of the local target area. The HOG feature extraction process is as follows:

- Reduce the influence of light and shade, normalize the color image;
- The gradient in the two-dimensional coordinate direction is calculated by the gradient template, which represents the horizontal and vertical gradient value of the pixel point;
- All pixels are weighted, and the weight is determined by gradient, and statistical voting is carried out to establish direction histogram;
- The cell units are combined to form a larger area, and the R-HOG rectangular interval is used to form a new feature vector, which is normalized, and the features in all blocks are gathered together to form the HOG features.

2.2.2. SVM classification algorithm. SVM [10] is a kind of generalized linear classifier which classifies data according to supervised learning method. Its decision boundary is the maximum margin hyperplane of learning samples, also called hyperplane. In this paper, linear SVM is used to solve the problem of image classification. First, input m sample data that can be divided linearly \( \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\} \), Where x is the eigenvector of N dimension, y is the binary output, and the value is +1 or -1; the output of SVM model is parameters \( \omega \), \( b \) and classification decision function. The steps are as follows:

- The construction constraint optimization problem is as follows (1):

\[
J(\omega) = \frac{1}{2} \|\omega\|^2_2, \text{s.t. } y^{(i)}(\omega^T \cdot x^{(i)} + b) \geq 1, i = 1, 2, \ldots, m
\]

(1)

- Using the Karush-Kuhn-Tucker condition to get the formula (2), the Lagrange Multiplier is introduced to get the optimal solution \( \beta^* \):

\[
L(\omega, b, \beta) = \frac{1}{2} \|\omega\|^2_2 + \sum_{i=1}^{m} \beta_i [1 - y^{(i)}(\omega^T \cdot x^{(i)} + b)], \beta \geq 0
\]

(2)

- Update parameters \( \omega, b \) as formula (3) (4):

\[
\omega^* = \sum_{i=1}^{m} \beta_i^* y^{(i)} x^{(i)}
\]

(3)

\[
b^* = y^* - \sum_{i=1}^{m} \beta_i^* y^{(i)} x^{(i)^T} x^i
\]

(4)

- Construct the final classifier as follows (5):

\[
f(x) = \text{sign}(x) = (\omega^* \cdot x + b^*)
\]

(5)

3. Experiment and results

In this chapter, we will use our own HT model dataset and classify it into two types: HT and TF through the HOG-SVM classification method, and count the accuracy of the detection algorithm.

3.1. Get the dataset

We select the standard logical unit in smic13 process library, and design the layout of HT by changing the dopant polarity of the active area. In Virtuoso layout drawing software, we select some basic logic units in Table 1 to draw HT layout. After drawing the layout of the HT model, in order to make the dataset, we extract the pictures of the Trojan horse unit in the layout, and set its resolution to 1000 * 1000, and the picture format to JPG format. The layout of some Trojan units is shown in Figure 3. In the same way, we also extract the layout pictures of the basic logic units in the process library as a HT free type.
Figure 3. Layout of some Trojan units.

Table 1. HT model modification unit.

| Category          | Unit Name               |
|-------------------|-------------------------|
| Combinational logic | AND, AO, NAND, NOR, OR, XOR, INV, XNOR |
| Sequential logic  | DFF, DFFR, DFFSR        |
| Other logic       | DLY, BUF                |

3.2. Machine learning algorithm flow

We train and test the dataset through the HOG-SVM algorithm. The experimental environment is Python 3.6, and the algorithm implementation flow is as follows:

- Get a list of pictures. The experimental dataset is randomly divided into two parts: training set and test set, in which the number of pictures ratio is 8:2;
- Extract and save the HOG features. The HOG method is called and the parameters are adjusted continuously to minimize the training time and improve the training accuracy. The parameters to be adjusted are: orientations (number of orientation bins), pixels (size (in pixels) of a cell), cells (number of cells in each block);
- Training and testing. The experimental data type is linear and classifiable, and the SVM training model is linear support vector classification for training and testing.

3.3. Experimental results

In order to make a detailed evaluation of the algorithm performance, after determining the more reasonable system parameters, different numbers of training samples are selected for training in the experiment, and the training results are shown in the following Table 2.

Table 2. Accuracy and time on the dataset.

| Experimental data (number of train data/test data) | Recognition time(s) | Recognition accuracy |
|---------------------------------------------------|----------------------|----------------------|
| 100/56                                            | 17.7                 | 89.3%                |
| 141/56                                            | 50.4                 | 91.1%                |
| 182/56                                            | 84.7                 | 92.8%                |
| 223/56                                            | 123.6                | 92.8%                |

It can be seen from the above table that the success rate of detecting the layout of the HT by the HOG-SVM method is 92.8%, while Bao [7] offered the accuracy of 83.6% and Nasr [8] provided the accuracy of 83.3%, as shown in Table 3. Considering the limited experimental dataset samples, the experiment accuracy can be accepted. It proved the feasibility of the presented method. In the future work, we will add more samples to improve the experiment accuracy.
Table 3. Comparison with prior work.

| Paper | Algorithm       | Accuracy                  |
|-------|-----------------|---------------------------|
| This paper | HOG+SVM         | 92.8%                     |
| Nasr [8] | Haar+Adaboost   | 83.3% (Trojan-Parametric) |
| Bao [7]  | K-means+SVM     | 83.6% (Trojan-Parametric) |

4. Conclusion

With the IC more and more widely used in all walks of life, IC security has become more and more important. In this paper, we presented a HOG-SVM detection method based on chip layout, and achieved the accuracy of 92.8%. It proves that this algorithm is a robust method to detect the doping modified HT’s layout and it also provides a new idea for HT detection methodology.

Acknowledgments

Our work was supported by Key Laboratory Fund-61421030304, Military Financial Fund, and NSAF-U1630133.

References

[1] Yang K, Hicks M, Dong Q, et al. A2: Analog Malicious Hardware[C]// 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016.
[2] Becker G T, Regazzoni F, Paar C, et al. Stealthy dopant-level hardware Trojans: extended version[J]. Journal of Cryptographic Engineering, 2014, 4(1):19-31.
[3] Xue M, Wang J, Hu A. Adaptive optimization of two-class classification-based hardware Trojan detection method (Regular paper, in Chinese) [J]. Chinese Journal of Computers, 2018, 41(2):439-451.
[4] Nourian M A, Fazeli M, Hely D. Hardware Trojan Detection Using an Advised Genetic Algorithm Based Logic Testing[J]. Journal of Electronic Testing, 2018.
[5] De Souza, Faria G, Kim H Y. Differential audio analysis: a new side-channel attack on PIN pads[J]. International Journal of Information Security, 2018.
[6] Nejat A, Hely D, Beroulle V. ESCALATION: Leveraging Logic Masking to Facilitate Path-Delay-Based Hardware Trojan Detection Methods[J]. Journal of Hardware and Systems Security, 2018.
[7] Bao C, Forte D, Srivastava A. On Reverse Engineering-Based Hardware Trojan Detection[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 35(1):1-1.
[8] Nasr A A, Abdulmageed M Z. Automatic Feature Selection of Hardware Layout: A Step toward Robust Hardware Trojan Detection[J]. Journal of Electronic Testing, 2016, 32(3):357-367.
[9] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.
[10] Cortes, Corinna, Vapnik, Vladimir. Support-vector networks[J]. Machine Learning, 20(3):273-297.