# IFCM-Weighted SVDD Hardware Trojan Detection Method (Postprint)

**Authors:** Wei Yanhai, Li Xiongwei, Zhang Yang, Hu Xiaoyang, Zhang Kunpeng

**Date:** 2018-08-13T00:00:00+00:00

## Abstract

To address the challenges posed by the wide variety of hardware Trojan (HT) types, which make it difficult to obtain features of unknown Trojans, and the noise present in collected side-channel signals, this paper proposes a hardware Trojan detection method based on IFCM-weighted SVDD (IFCMW_SVDD). Traditional Support Vector Data Description (SVDD) suffers from the limitation of treating all samples equally under identical conditions when solving one-class classification problems, which necessitates distinguishing between primary and secondary samples during training according to the specific problem. The proposed method employs an improved fuzzy C-means (IFCM) algorithm to calculate the membership degree of "golden chip" side-channel signals, which is then used as a weight (W) coefficient for sample features. This enables the support vectors of the SVDD model constructed for hardware Trojan detection to characterize the "golden chip" signals while minimizing the description boundary as much as possible. Experimental results demonstrate that the proposed method achieves one-class hardware Trojan detection with improved detection accuracy and stability compared to the traditional SVDD algorithm.

## Full Text

## Hardware Trojan Detection Method Based on IFCM-Weighted SVDD

**Wei Yanhai[1], Li Xiongwei[1], Zhang Yang[1], Hu Xiaoyang[1], Zhang Kunpeng[2]** [1]. Equipment Simulation Training Center, Army Engineering University of PLA, Shijiazhuang Campus, Shijiazhuang 050003, China
[2]. Unit 66407 of PLA, Beijing 100093, China

**Abstract**

This paper proposes a hardware Trojan (HT) detection method based on Improved Fuzzy C-Means Weighted Support Vector Data Description (IFCMW_SVDD) to address two key challenges: the difficulty of obtaining features for unknown Trojans due to their vast variety, and the presence of noise in collected side-channel signals. Traditional SVDD has a limitation when solving one-class classification problems—it treats all training samples equally under the same conditions, whereas practical problems require distinguishing between primary and secondary samples during training. Our approach calculates the membership degree of "golden chip" side-channel signals using an Improved Fuzzy C-Means (IFCM) method and uses it as a weight coefficient for sample features. This enables the SVDD model' s support vectors to describe "golden chip" signals while minimizing the description boundary for hardware Trojan detection. Experimental results demonstrate that the proposed method achieves single-class hardware Trojan detection with improved accuracy and stability compared to traditional SVDD.

**Keywords:** hardware Trojan; side-channel signals; improved fuzzy C-means method (IFCM); support vector data description (SVDD); membership degree

## 0 Introduction

Hardware Trojans are micro-scale malicious circuit modules that compromise IC security by altering original designs, posing significant threats to integrated circuit (IC) applications. Current detection approaches include reverse engineering chip 解剖, logic functional testing, and side-channel analysis. While reverse engineering can achieve 100% detection rates [1], the process is time-consuming and labor-intensive. As market demand drives cost reduction, IC design, manufacturing, and packaging are increasingly outsourced, making chips more vulnerable to hardware Trojan insertion.

Side-channel signal-based detection offers a non-invasive alternative that avoids chip decapsulation [2]. By acquiring side-channel signals and performing feature transformation for differential comparison, this method can determine whether a device under test contains hardware Trojans. However, existing side-channel detection techniques are primarily designed for experimentally crafted Trojans, with each method targeting only specific Trojan types [3], yielding unsatisfactory results for unknown Trojans and one-to-many detection scenarios. Bao et al. [4] proposed a reverse engineering approach to identify Trojan-free chips using one-class Support Vector Machines (OCSVM) for hardware Trojan detection, but environmental influences on training samples cause model overfitting by expanding the decision boundary. Xu et al. [5] developed an intrusion detection model combining SVDD with clustering algorithms, applying k-means clustering to normal samples before data description for anomaly detection. Although effective on DARPA' 99 datasets, this approach struggles with complex, noise-sensitive Trojan signals. Niazmardi et al. [6] designed an SVDD-FCM al-

gorithm for remote sensing image classification with limited training data, but FCM tends to misclassify relatively dispersed samples as multiple categories.

To address these limitations, we analyze experimentally collected side-channel signal samples that follow a normal distribution in high-dimensional space. Leveraging SVDD' s capability for handling anomalous data and considering the shortcomings of FCM for single-cluster centers and high computational complexity of traditional SVDD, we propose an IFCM-weighted SVDD approach for hardware Trojan detection. Experimental results demonstrate that the improved SVDD model maintains robust detection performance across multiple unknown Trojans without requiring prior Trojan characteristics.

## 1 Power Consumption Detection Model and Problem Analysis

Side-channel signals primarily consist of electromagnetic and power consumption signals, with the latter offering higher accuracy and becoming the mainstream focus in hardware Trojan detection. In practice, each acquired power consumption signal is a dataset measuring chip power dissipation at different time instances based on n sampling points. With m sampling iterations, we obtain a power consumption matrix $X_{m\times n}$, where each n-dimensional signal vector can be viewed as a sample point in space determined by a covariance matrix and mean, forming a hyperellipsoid distribution.

Experimental analysis reveals that "golden chip" power current signals ($I_g$) comprise main path current ($I_e$) and noise components ($I_n$, including electronic noise $I_{el}$ and switching noise $I_{sw}$): $I_g = I_e + I_n$, where $I_n = I_{el} + I_{sw}$. For chips under test (containing hardware Trojan signals $I_{tr}$), the signal becomes $I_g = I_e + I_n + I_{tr}$. The detection principle involves comparing side-channel signals from golden chips and devices under test. When hardware Trojans consume significant circuit overhead, the corresponding $I_{tr}$ value is large enough for waveform observation to distinguish Trojan signals. However, for low-overhead Trojans, traditional methods cannot effectively differentiate signals even when ignoring noise effects, necessitating techniques like K-L projection [7] or K-means clustering analysis [8,9] for feature differentiation. For even smaller Trojans, K-L methods fail to identify effective orthogonal projection directions, and K-means clustering struggles with small inter-class distances and overlapping samples. SVDD demonstrates superiority by mapping samples to higher-dimensional spaces for classification, requiring only golden chip signals for detection.

Traditional C-means clustering partitions n signal samples into c predefined clusters, minimizing the sum of squared errors between each sample and its cluster mean:

$$J = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \|y - m_i\|^2$$

where $m_i$ is the mean of cluster $i$, and $\Gamma_i$ contains all samples assigned to cluster $i$. C-means is a hard classification method that often yields unsatisfactory results [10]. Fuzzy C-Means (FCM) alleviates this by introducing fuzziness: given a sample set $\{x_i, i = 1, 2, \ldots, n\}$ and predetermined cluster number $C$ (set to 1 for golden chip samples), with cluster centers $\{m_i, i = 1, 2, \ldots, c\}$, the membership function $\mu_j(x_i)$ represents the degree to which the $i$-th sample belongs to cluster $j$. The clustering loss function becomes:

$$J_b = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_j^b(x_i) \|x_i - m_j\|^2$$

where exponent $b > 1$ controls the fuzziness of clustering results. As $b$ increases, fuzziness intensifies; when $b \to \infty$, the algorithm yields completely fuzzy solutions where all cluster centers converge to the global mean. Empirically, $b$ is typically set around 2.

FCM imposes the constraint $\sum_{j=1}^{c} \mu_j(x_i) = 1$ for each sample $x_i$. To overcome FCM's limitation where discrete points maintain high membership values across clusters, IFCM modifies the constraint such that the sum of membership degrees across all samples for each cluster equals $n$:

$$\sum_{i=1}^{n} \mu_j(x_i) = n$$

Minimizing the loss function under this constraint yields the update rules:

$$\mu_j(x_i) = \frac{1}{\sum_{k=1}^{c} \left( \frac{\|x_i - m_j\|^2}{\|x_i - m_k\|^2} \right)^{1/(b-1)}}$$

$$m_j = \frac{\sum_{i=1}^{n} \mu_j^b(x_i) x_i}{\sum_{i=1}^{n} \mu_j^b(x_i)}$$

The IFCM algorithm proceeds as follows: (a) Set cluster number $c$ and parameter $b$; (b) Initialize cluster centers $m_i$ using C-means results; (c) Iteratively compute membership functions using current centers and update cluster centers until membership values stabilize.

## 2.2 SVDD Algorithm Analysis

SVDD is a one-class classification algorithm derived from Support Vector Machines (SVM) that trains exclusively on target samples without requiring non-target (Trojan) samples. It has demonstrated success in image pattern recognition [6] and equipment fault analysis [11,12]. The fundamental principle maps limited training samples to a higher-dimensional space via function $\Phi$ and constructs the minimal hypersphere that encloses target samples while excluding anomalies. The SVDD objective function is:

$$\min R^2 + \lambda \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \|\Phi(x_i) - c\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0$$

where $R$ is the hypersphere radius, $\lambda > 0$ is a penalty coefficient (smaller $\lambda$ values exclude more golden chip samples, requiring trade-offs between model size and accuracy), $\xi_i$ are slack variables measuring the importance of non-target or noisy samples, and $c$ is the sphere center.

This convex quadratic programming problem yields the Lagrangian:

$$L(R, c, \alpha_i, \xi_i) = R^2 + \lambda \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (R^2 + \xi_i - \|\Phi(x_i) - c\|^2) - \sum_{i=1}^{n} \beta_i \xi_i$$

Setting partial derivatives to zero gives:

$$\sum_{i=1}^{n} \alpha_i = 1, \quad c = \sum_{i=1}^{n} \alpha_i \Phi(x_i), \quad \alpha_i = \lambda - \beta_i$$

Substituting into the Lagrangian yields the dual problem:

$$\max \sum_{i=1}^{n} \alpha_i K(x_i, x_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq \lambda, \quad \sum_{i=1}^{n} \alpha_i = 1$$

where $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ is the kernel function. Support vectors are sample points with $\lambda > \alpha_i > 0$. The distance from any sample point to the center is:

$$R_i = \|\Phi(x_i) - c\| = \sqrt{K(x_i, x_i) - 2\sum_{j=1}^{n} \alpha_j K(x_i, x_j) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)}$$

Samples with $R_i \leq R$ are classified as golden chip signals; otherwise, they contain Trojans.

## 2.3 IFCM-Weighted SVDD Detection Model

Traditional SVDD treats all training samples equally when selecting support vectors. However, golden chip side-channel signals contain noise and discrete samples that can shift the training center. We introduce membership degree $\mu_j(x_i)$ as a weighting parameter $W_i$ (samples closer to cluster centers have higher membership), improving detection accuracy. The weighted objective function becomes:

$$\min R^2 + \lambda \sum_{i=1}^{n} W_i \xi_i \quad \text{s.t.} \quad \|\Phi(x_i) - c\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0$$

The modified Lagrangian is:

$$L(R, c, \alpha_i, \xi_i) = R^2 + \lambda \sum_{i=1}^{n} W_i \xi_i - \sum_{i=1}^{n} \alpha_i (R^2 + \xi_i - \|\Phi(x_i) - c\|^2) - \sum_{i=1}^{n} \beta_i \xi_i$$

The dual problem transforms to:

$$\max \sum_{i=1}^{n} \alpha_i K(x_i, x_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \quad \text{s.t.} \quad 0 \leq \alpha_i \leq \lambda W_i, \quad \sum_{i=1}^{n} \alpha_i = 1$$

Comparing the formulations reveals that weighting only modifies $\alpha_i$, where samples with higher membership degrees have larger Lagrange coefficients and are more likely to become support vectors, yielding a more reasonable model.

The IFCM-weighted SVDD algorithm proceeds as: (a) Input power consumption samples $X = \{x_i, i = 1, 2, \ldots, n\}$, set $c = 1$; (b) Average every ten samples to obtain processed data matrix; (c) Initialize cluster centers $m_i$ using C-means; (d) Iteratively compute membership functions and update golden chip cluster centers until convergence; (e) Apply IFCM weights to SVDD; (f) Standardize samples and construct SVDD model; (g) Compute distances to cluster center for unknown signals; (h) Classify as Trojan-containing if $R - R_i < 0$.

## 2.4 Algorithm Effectiveness Analysis

We validate the algorithm using two-dimensional training samples [Figure 1: see original paper]. The Gaussian kernel function $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2\sigma^2))$ is employed with grid search optimization for parameters $(\lambda, \sigma^2)$ in ranges $\lambda = [10^{-8}, 27]$ and $\sigma^2 = [10^{-5}, 2^{-4}]$, initialized at $\lambda = 2$, $\sigma^2 = 10^{-5}$. Figures 1(a), (b), and (c) illustrate SVDD, FCM_SVDD, and IFCMW_SVDD models, respectively. IFCMW_SVDD exhibits a tighter boundary with more support vectors selected more reasonably than the other

two methods. While FCM_SVDD shows two cluster centers (treating samples as two classes), IFCMW_SVDD correctly identifies a single center, addressing this limitation.

## 3.1 Experimental Configuration

We verify the method' s capability to detect multiple Trojans using ISCAS85 circuits, which allow configuring various Trojan scales. Combinational hardware Trojans of different gate counts were designed in c1908, c2670, c3540, c5315, c6288, and c7552 circuits [Figure 2: see original paper]. The Trojan comprises a trigger module and payload module. When the control signal is low, registers reset; when high, registers operate and output data to selectors. Upon reaching a specific state, the Trojan activation signal triggers the payload. Golden chip and Trojan signals were acquired (2ps time resolution, 1.4ms simulation time) with ten-fold averaging to reduce noise.

For further validation, Experiment 2 uses a SASEBO development board [Figure 3: see original paper]. The physical platform runs encryption algorithms on FPGA (with implanted Trojans for "Trojan" signal collection) while a Tektronix DPO4032 oscilloscope (350 MHz bandwidth) captures golden chip signals transmitted to a PC (CPU i5-6400). Eclipse+Pydev+Anaconda3 serves as the programming environment for SVDD and IFCM training, with MATLAB for model evaluation.

## 3.2 Experimental Results Analysis

Experiment 1 collected 1,000 power traces with 600 sampling points each for both golden chip and Trojan signals. Experiment 2 collected 5,000 traces with 10,000 sampling points, randomly selecting 10 traces from each class for comparison (Experiment 1: 600 points; Experiment 2: 1,400 points) [FIGURE:4, 5]. Direct observation reveals minimal fluctuations that cannot reliably distinguish signals, necessitating automated detection.

Detection rates for various Trojan scales are shown in . The proposed algorithm significantly outperforms SVDD, achieving 99.54% detection rate at 1.44% area overhead versus only 65.12% for traditional SVDD. While both methods' performance degrades with decreasing overhead, our approach maintains better stability.

Experiment 2 implanted a hardware Trojan occupying 2.5% of AES encryption circuit area on FPGA. Using 1,000 golden chip signals as training windows with grid search optimization, detection rate curves versus training sample size are shown in [Figure 6: see original paper]. SVDD and FCM_SVDD exhibit similar performance, substantially lower than our method.

We evaluate model effectiveness using precision ($P$), recall ($R$), and P-R curves:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}$$

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively. The P-R curve [Figure 7: see original paper] shows IFCMW_SVDD dominating others, with higher "balance points" (where precision equals recall) indicating superior performance.

## Conclusion

This paper proposes an IFCMW_SVDD hardware Trojan detection method that maps sample signals to high-dimensional space, constructing a hyperellipsoid using support vectors to isolate unknown Trojan side-channel signals outside the boundary. The approach requires only golden chip signals without Trojan analysis, uses weighting to mitigate overfitting/underfitting from improper model sizing, and retains only relevant support vectors to save storage space. Validation through simulation and side-channel experiments demonstrates effective detection of various hardware Trojans, providing new insights for future research.

## References

[1] Bao Chongxi, Yang Xie, Liu Yuntao, et al. Reverse engineering-based hardware trojan detection [M]// The Hardware Trojan War. Berlin: Springer, 2018: 269-288.
[2] Zhang Yang, Li Xiongwei, Chen Kaiyan, et al. Research of hardware trojan design and differential analysis based on fault injection [J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2014, 42(4): 68-71.
[3] Dupuis S, Flottes M L, Di Natale G, et al. Protection against hardware trojans with logic testing: proposed solutions and challenges ahead [J]. IEEE Design & Test, 2018, 35(2): 73-90.
[4] Bao Chongxi, Forte D, Srivastava A. On application of one-class SVM to reverse engineering-based hardware trojan detection [C]// Proc of the 15th International Symposium on Quality Electronic Design. Piscataway, NJ: IEEE Press, 2014: 47-54.
[5] Xu Jing, Shi Duanyin, Zhang Yajiang, et al. Model of IDS based on SVDD and cluster algorithm [J]. Control and Decision, 2010, 25(3): 441-444.
[6] Niazmardi S, Homayouni S, Safari A. An improved FCM algorithm based on the SVDD for unsupervised hyperspectral data classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2013, 6(2): 831-839.
[7] Agrawal D, Baktir S, Karakoyunlu D, et al. Trojan detection using IC fingerprinting [C]// Proc of Symposium on Security and Privacy. Washington DC: IEEE Computer Society, 2007: 296-310.

[8] Wang Bairen, Qu Ming. Hardware trojan detection method based on K-means clustering analysis [J]. Journal of Beijing Electronic Science and Technology Institute, 2016, 24(2): 84-87.
[9] Bao Chongxi, Forte D, Srivastava A. On reverse engineering-based hardware trojan detection [J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 2018, 35(1): 49-57.
[10] Bian Zhaoqi, Zhang Xuegong. Pattern recognition [M]. 2nd Edition. Beijing: Tsinghua University Press, 2000.
[11] Gryllias K, Qi Junyu, Mauricio A R, et al. A semi-supervised SVDD-based fault detection method for rolling element bearings [C]// Proc of the 1st World Congress on Condition Monitoring. 2017.
[12] Chen Muchen, Hsu Chunchin, Malhotra B, et al. An efficient ICA-DW-SVDD fault detection and diagnosis method for non-Gaussian processes [J]. International Journal of Production Research, 2016, 54(17): 5272-5287.

*Note: Figure translations are in progress. See original paper for figures.*