Automatic Feature Selection and Improved Classification in SICADA Counterfeit Electronics Detection

Lauren Milechin, Eric Koziel, Michael Vai and Roger Khazan
MIT Lincoln Laboratory
Lexington, MA, USA

Keith Bergevin and Phil Comer
Defense Microelectronics Activity (DMEA)
McClellan, CA, USA

Contact Author: lauren.milechin@ll.mit.edu

Abstract: *Counterfeiters seeking financial gain can introduce misrepresented or recycled microelectronic components to both government and commercial supply chains. This reduces system reliability and trust, and currently has few comprehensive and practical solutions. The SICADA methodology was developed to detect such counterfeit microelectronics by collecting power side channel data and applying machine learning to identify counterfeits. This methodology has been extended to include a two-step automated feature selection process and now uses a one-class SVM classifier. We describe this methodology and show results for empirical data collected from several types of Microchip dsPIC33F microcontrollers*.

Keywords: counterfeit detection; device authentication; supply chain security; hardware identity

Introduction and Background

Many government programs outlast the typical support lifetimes of their required components. Untrusted third parties can target this disparity for financial gain by selling counterfeits of obsolete and rare parts. These counterfeits may be similar devices modified to look like the desired part, or could be a recycled part. We initially developed and described the Side Channel Authenticity Discriminate Analysis (SICADA) [1] methodology to classify counterfeit devices using power side channels and detect such types of misrepresented parts. SICADA has advantages over many typical counterfeit detection methods in that it does not require device destruction, can be used on legacy devices, and is not overly expensive and time consuming [2]. This methodology has been improved with an automated feature selection process and a one-class unsupervised counterfeit identification method.

SICADA works by comparing features derived from the power waveforms of a suspect device to those from known authentic devices (referred to as the golden set). It attempts to classify the unknown device as counterfeit or legitimate based on differences in these observations. The goal is to have a practical system that can be deployed in the environments where the authenticity of a device needs to be determined, such as distribution centers and end users. This analysis approach is a machine learning process that involves four main parts: data collection, feature generation, feature selection, and classification. The first two steps are described in [1], and our methods for feature selection and classification are described in this work.

Machine learning classification methods can be incredibly sensitive to the features provided to them. A good algorithm can produce bad results if it is supplied with too many, noisy, or highly correlated features. Feature reduction techniques can be separated into two main categories: selection and projection. Projection techniques involve transforming features into a lower-dimensional subspace, whereas selection involves choosing individual features without transformation. A popular projection technique, PCA (Principle Component Analysis), is often used to produce a smaller set of transformed features. PCA involves a transformation that maximizes variation among given data points. While this provides good separation between data points, this variation maximization occurs indiscriminately, and can separate data within the same class. Further, projection techniques in general provide transformed features, which can make it difficult to interpret which original features are causing the separation of the data points.

Feature selection instead determines the most important features to separate a given dataset. This can provide insight into the features themselves. In our case, it allows for analysis on the best discriminating features in our set and provides insight into the physical phenomena that define the differences. This further allows for fine-tuning the data collection process for more accurate counterfeit identification.

Once the final feature set is determined, the classification step itself is a challenge. Not only is this an anomaly detection problem, but the model must not assume it has seen all possible types of counterfeits in its training data or that any counterfeits it has seen are necessarily representative of those it may come in contact with in the future. SICADA assumes that a "golden set" (a set of known legitimate devices) exists, and therefore lends itself well to unsupervised one-class classifiers.

In the remaining sections, we first describe in more detail our new methodology for feature selection and classification. We then describe how we tested our methods and discuss the results from those experiments.

This material is based upon work supported by the Office of the Secretary of Defense under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001.
Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of the Secretary of Defense.

Methodology

Data Collection: Empirical power measurements are gathered by monitoring electrical current fluctuations on device power inputs. This is done by observing voltage drop across a sense resistor with an oscilloscope. Other inputs to the device include a fixed test program and an external clock. Our test program was designed to exercise different parts of the device circuit, and the external clock is used to provide measurement consistency.

Feature Generation: We process the collected data using the same feature generation approach described in [1] and adapted from [3]. Power traces are broken down by clock cycle and transformed into Hilbert analytical signals, from which the instantaneous amplitude, phase, and frequency are derived. The first four statistical moments (mean, variance, kurtosis, and skewness) are gathered from each of these waveforms. This creates 12 distinct features for each clock cycle. For example, one feature is the mean instantaneous amplitude of the 10^{th} clock cycle. In this way, we can enumerate the shape characteristics of the clock events of our test program.

Feature Selection: There are two challenges that arise from the large number of features produced by SICADA during feature generation: the many highly linearly correlated features and sheer number of features, many of which play a minimal role in discriminating unlike devices. Therefore, we have two steps in SICADA's automated feature selection: first the removal of highly correlated features and second the selection of features that separate the known and unknown devices.

The first step addresses the first challenge, the large number of highly correlated features. This step occurs immediately after the features for the golden set are generated, and only needs to be performed once for each golden set. We remove the strongly linearly correlated features by calculating the absolute value of the correlation coefficient, |ρ|, for each pair of features using data from the golden set. The value of |ρ| can be calculated as follows:

$$|\rho| = \left| \frac{cov(X_i, X_j)}{\sqrt{\sigma_i \sigma_j}} \right|$$

where $cov(X_i, X_j)$ is the sample covariance of features X_i, X_j, and σ_i, σ_j are the standard deviations of X_i, X_j, respectively. The correlation coefficient is a number between -1 and 1, which measures the linear correlation between two random variables. Correlation coefficients close to 1 indicate strong positive correlation, close to -1 indicate strong negative correlation, and correlation coefficients close to 0 imply little to no linear correlation between random variables. After the correlation coefficients are calculated, one feature from each pair of features with |ρ| greater than some threshold is eliminated. With a threshold of .9, we remove roughly 85% of the features, depending on the chosen golden set device type.

The second feature selection step selects the final features that will be used for classification. Essentially, we grow a decision tree on both the data from the golden and suspect devices, labeling the golden and suspect device data as two separate classes, and use the features chosen for the decision tree as our final feature set. Since this involves data from both the golden set and the suspect device, it must be done once for each suspect device.

The basic properties of decision trees make them a good way to quickly select discriminating features. Decision trees fundamentally operate by selecting the features that best separate data given class labels. At each point in the tree, a true/false question is asked about one feature of the data that arrives there. All the data points for which the result is true go down one branch, and the remaining go down the other branch. During training, the choice of which true/false question asked at any given node will be one that best separates the training data that arrives at that node. Training is complete when each training data point is successfully classified, if possible, and so only features required to fully separate the data will be used in the decision tree. Decision trees have long been shown to produce good feature subsets for classification [4]. We therefore use the features chosen in the decision tree as our final feature set for the classifier.

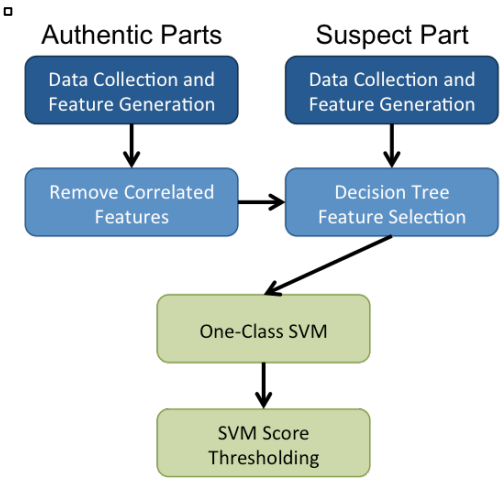

Figure 1: Basic SICADA Methodology. Data collection and feature generation is the first step for both the known authentic parts and the part in question. Following that is the feature selection process and the classification steps.

Classification: We previously investigated several classifiers for SICADA, and have settled on an unsupervised one-class Support Vector Machine (SVM) classifier. It is primarily an outlier detection method, and it works best when finding a small number of outliers in a larger dataset, as described in [5] and shown in [6]. This property suits our problem well, where we have a few samples from a suspect device that we are comparing to samples from multiple golden devices. The one-class SVM generates a boundary around the majority of the data (this

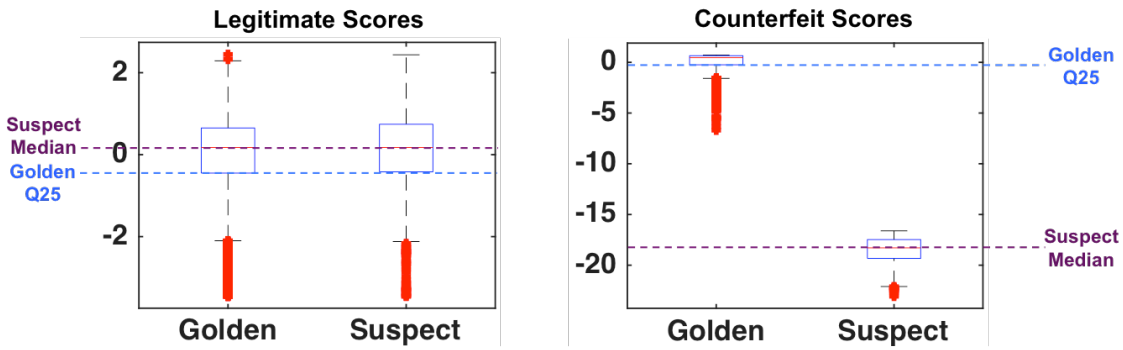

Figure 2: SVM Score Comparisons. Each figure shows boxplots of the SVM scores. Here, we are declaring a counterfeit if the median (50th quantile) of the suspect device SVM scores is less than the 25th quantile of the golden device set SVM scores. On the left, the suspect device is correctly declared legitimate, while the device tested on the right is declared a counterfeit.

is a percentage specified by the user) and gives each point a score: positive scores for points within the boundary, with larger positive scores for points closer to the center, and negative scores for those outside the boundary, larger in magnitude for those further from the boundary.

SICADA uses the one-class SVM by inputting both golden and suspect data into the classifier, and setting the outlier percentage to the percentage of suspect samples out of the total number of samples. We can make a counterfeit declaration based on the relative distributions of the resulting scores. Intuitively, we can say a device is counterfeit if most of the scores for its samples are negative or lower than the majority of the scores for the known device. However, if the scores of the suspect and golden device samples are similarly distributed, the two devices are indistinguishable, and the suspect device could be legitimate (see Figure 1). Specifically, we declare a counterfeit device if the i^{th} quantile of the suspect scores is less than the j^{th} quantile of the golden scores, where $i > j$. Otherwise, we say the device is legitimate.

Experimental Setup

Data used in this paper is identical to the data collected for [1]. Empirical measurements were taken from Microchip's dsPIC33F family of microcontrollers (Table 1). We used j12 parts as the base device, while j32 and j128 parts had larger memory sizes. The j128 devices have more comparators, DACs, and timers compared to either j12 or j32 devices. Each type is further broken down by grade, where I-type parts are standard industrial grade and E-type parts are extended temperature grade parts. Most device types had samples from different date codes to avoid matching against only specific lots. Note that we have made a differentiation between the j12i and j12ib devices. The j12i parts are several years older and are a different silicon revision than the other parts used in this study.

Power information was collected via custom sensing circuits and captured on an oscilloscope. Each microcontroller was loaded with a program to execute 1000 iterations of a consistent operational loop utilizing a mixture of arithmetic, register, and memory operations. This operational loop comprised 53 individual clock cycles. Each device was clocked via an external pulse generator at 10 MHz to avoid timing inconsistencies in factory calibration settings. To reduce the possibility of temporary environmental factors affecting the results of our evaluation, we collected data on three separate dates. The average results mentioned are averages of the evaluation metrics of these separate datasets.

Table 1: Microchip dsPIC33F Type List

Label	Device	Quantity
j12i	J12GP-202 I	9
j12ib	J12GP-202 I	5
j12e	J12GP-202 E	6
j32i	J32GP-202 I	18
j32e	J32GP-202 E	6
j128i	J128GP-802 I	22
j128e	J128GP-802 E	15

Our evaluation scheme involves many pairwise comparisons. Four representative parts for each device type (shown above in Table 1) are chosen to make up the golden models. This is because we had as few as five or six devices for some types, and through testing it seemed that at least four devices were required for good performance. Then, for each golden set, we run the feature selection and classification steps for each of the remaining devices. This yields 539 pairwise comparisons, 486 of which are different-type comparisons and should be identified as counterfeit, and 53 same-type comparisons that should be classified as legitimate.

For the first step of feature selection, we use a correlation coefficient threshold of 0.9, which removes about 85% of the generated features. The further number of features removed during the decision tree feature selection varies between each pairwise comparison. However, the number of features remaining for legitimate test devices tends to be much higher than the number of features remaining for counterfeit test devices. The reason for this is fairly intuitive: a legitimate device is much harder to differentiate from the golden set, and so requires a large, severely overfit tree to classify. The suspect and golden set score quantiles (i and j) used for comparison are 52.5% and 65%,

respectively. These were chosen from a series of ROC curves sweeping these thresholds looking to maximize the true positive rate (catching a counterfeit), minimize false positive rate (misclassifying a legitimate part as a counterfeit), and be somewhat robust to changes in the data.

The SICADA methodology, starting with feature generation, was implemented using MATLAB ®, and uses decision tree and SVM included in the Statistics Toolbox.

Results

We evaluate the results in terms of identifying a true counterfeit as a positive detection. Below in Table 2, we show results broken down by golden type, or the known authentic part type. This table includes results from the first data collection, however these results are typical for each collection. Table 3 shows the overall results from the three collections.

Table 2: Results for one data collection summarized for each golden device type.

Golden Type	Total Counterfeit	Total Legitimate	True Positive Rate	False Positive Rate
j12i	72	5	98.61%	0.00%
J12ib	76	1	94.74	0.00
j12e	75	2	100.00	0.00
j32i	63	14	90.48	0.00
j32e	75	2	76.00	0.00
j128i	59	18	77.97	0.00
j128e	66	11	66.67	0.00

In general, it seems that counterfeits imitating the j12 parts are most successfully detected, followed by j32 devices, and finally j128 devices. This was consistent across all three data collections. Further, the counterfeits that were missed when compared to the j12ib devices were all j12i parts, which are an older version of the same device.

Most missed counterfeits involved differentiating between environmental grades of the j128 and j32 devices. This could be because there are twice as many or more these than the others, across a couple more date codes, and so have more opportunity for varying signatures. In this first run, all but ten of the missed counterfeit suspect parts were of these types, and mainly are missed when compared with their environmental grade (I-type vs. E-type) counterpart. It is possible that a larger golden set for these device types or separating out the date codes as individual types will improve finding counterfeits of these types. More work can be done on differentiating the environmental grades.

Overall, in each collection on the three separate dates, we catch over 80% of the counterfeits, and on average capture 88.96% of the counterfeits. Our false positive rate is very low, so we rarely misclassify legitimate devices as counterfeits.

Table 3: Overall results for each data collection.

Data Collection	True Positive Rate	False Positive Rate
1	86.83%	0.00%
2	96.50	9.43
3	83.54	0.00

Conclusion

We have extended the SICADA methodology to include an automated feature selection process and one-class SVM classification. The two-step feature selection process allows us to eliminate feature pairs that are highly correlated, and adaptively select features to use for each new suspect device. By using an unsupervised one-class SVM, we are able to identify counterfeits with only four known devices for training. Through our experiments, where similar, but different-type devices are used to represent counterfeits, we found we were able to successfully identify these counterfeits 88.86% of the time on average and miss-classified legitimate devices 3.14% of the time.

In future work, we look to test this methodology on a wider range of microelectronics parts, namely FPGAs. We aim to conduct more thorough analysis on the features that are important discriminators of different parts, and identify those that cause us to misidentify a legitimate device.

References

1. Koziel, E., Thurmer, K., Milechin, L. et al. Side Channel Authenticity Discriminant Analysis for Device Class Identification. In *Government Microciruit Applications & Critical Technology Conference* (2016).
2. Government Accountability Office. *Suspect Counterfeit Electronic Parts Can Be Found on Internet Purchasing Platforms*. GAO-12-375. (2012).
3. Cobb, W., Lapse, E., Baldwin, R., Temple, M., and Kim, Y. Intrinsic Physical-Layer Authentication of Integrated Circuits. *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1 (February 2012).
4. Kohavi, R. and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, vol. 97, no. 1 (1997), p. 273-324.
5. Khan, S. S. and Madden, M. G. A survey of recent trends in one class classification. In *Irish conference on Artificial Intelligence and Cognitive Science* (Berlin 2009), Springer, p. 188-197.
6. Wang, K. and Stolfo, S. J. One-class training for masquerade detection. In *Workshop on Data Mining for Computer Security, Melbourne, Florida* (2003), p. 10-19.