A Systematic Evaluation of Profiling Through Focused Feature Selection

Picek, Stjepan; Heuser, Annelie ; Jovic, Alan; Batina, Lejla

DOI
10.1109/TVLSI.2019.2937365

Publication date
2019

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Very Large Scale Integration (VLSI) Systems

A Systematic Evaluation of Profiling Through Focused Feature Selection

Stjepan Picek, *Senior Member, IEEE*, Annelie Heuser, Alan Jovic, *Member, IEEE*, and Lejla Batina, *Senior Member, IEEE*

***Abstract*— Profiled side-channel attacks consist of several steps one needs to take. An important, but sometimes ignored, step is a selection of the points of interest (features) within side-channel measurement traces. A large majority of the related works start the analyses with an assumption that the features are preselected. Contrary to this assumption, here, we concentrate on the feature selection step. We investigate how advanced feature selection techniques stemming from the machine learning domain can be used to improve the attack efficiency. To this end, we provide a systematic evaluation of the methods of interest. The experiments are performed on several real-world data sets containing software and hardware implementations of AES, including the random delay countermeasure. Our results show that wrapper and hybrid feature selection methods perform extremely well over a wide range of test scenarios and a number of features selected. We emphasize L1 regularization (wrapper approach) and linear support vector machine (SVM) with recursive feature elimination used after chi-square filter (Hybrid approach) that performs well in both accuracy and guessing entropy. Finally, we show that the use of appropriate feature selection techniques is more important for an attack on the high-noise data sets, including those with countermeasures, than on the low-noise ones.**

***Index Terms*— Feature selection, guessing entropy (GE), machine learning (ML), profiled side-channel attacks (SCAs), random delay countermeasure.**

I. Introduction

PROFILED side-channel attacks (SCAs) have received a lot of attention in recent years because this type of attack defines the worst case security assumptions. Besides the more traditional choice of template attack (TA), a number of machine learning (ML) techniques have been investigated in this context [1]–[3]. The common knowledge from those results suggests that profiled side-channel analysis can be extremely powerful for key recovery, with ML being a highly viable choice. Contrarily, feature selection, in particular, the usage of ML-based techniques, did not receive significant attention. Early works on TAs introduced sum of squared pairwise T-differences (SOST)/sum of squared differences (SOSD) [4] as feature selection methods, and consequently, most of the follow-up works assume that this step has somehow been performed in a satisfactory, if not optimal, manner. A common strategy often also suggests using the Pearson correlation for this purpose (see [2] and [5]).

First, we ask a question on the importance of the number of features in a data set. For a fixed number of training samples, the predictive power of a classifier algorithm eventually reduces as the dimensionality (the number of features) of the problem increases. Consequently, for scenarios with a large number of features, we need to use more training examples, where that number increases exponentially with the number of dimensions. This results in the so-called curse of dimensionality (and the closely related Hughes effect). When discussing features (also known as points of interest, points in time, variables, and attributes), we can distinguish among relevant, irrelevant, and redundant features. A meaningful separation in these categories is very important when optimizing the attack strategy and can be divided into the following general directions:

1) feature selection—where the most important subsets of features are selected;
2) dimensionality reduction—where methods like principal component analysis (PCA) transform the original features into new features;
3) deep learning techniques like convolutional neural networks that perform implicit feature selection.

The last two techniques can be very successful but they do not provide information about the original features. Such techniques either completely transform the features or use them in a manner too complicated to be understood by human experts. Moreover, deep learning could often have no performance advantage against "standard" ML if the number of measurements is not very large [6]. Note that, in this article, we do not consider comparisons with deep learning techniques, but we refer interested readers to [7]–[9].

There are many articles considering profiled SCA, where the number of features is fixed and the analysis is conducted by considering only the changes in the number of traces or by selecting a more powerful classifier (see [10] and [11]). It is indeed somewhat surprising that the SCA community (until now) did not take a closer look at the feature selection part of the classification process. Similar to the powerful classification methods coming from the ML domain, there are also feature selection techniques one could utilize. To the best of the authors' knowledge, there exists one work that focuses on the feature selection for profiled SCA, but it does not consider ML techniques and it compares only methods known for side-

Manuscript received January 29, 2019; revised June 1, 2019 and July 23, 2019; accepted August 11, 2019. This work was supported in part by the Technology Foundation STW through the Dutch Government under Project 13499 TYPHOON—VIDI. *(Corresponding author: Alan Jovic.)*

S. Picek is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: s.picek@tudelft.nl).

A. Heuser is with the French National Center for Scientific Research (CNRS), IRISA, 35000 Rennes, France (e-mail: annelie.heuser@irisa.fr).

A. Jovic is with the Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia (e-mail: alan.jovic@fer.hr).

L. Batina is with the Institute for Computing and Information Sciences, Radboud University, 6525 XZ Nijmegen, The Netherlands (e-mail: lejla@cs.ru.nl).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2019.2937365

channel analysis either as feature selection techniques or as distinguishers [12]. Note that, in leakage detection (see [13]), one is identifying data-dependent but not necessarily model-agnostic leakage information. Consequently, detecting features (points in the power trace) is a task orthogonal to leakage detection, as leakage detection [according to, e.g., test vector leakage assessment (TVLA)] may not necessarily lead to a successful key recovery. One approach could be as follows: first, use leakage detection to identify possible leakages in the trace, then analyze the corresponding operation, in particular, determine if the model is key sensitive, and finally use feature selection in combination with the underlying model for a profiled distinguisher.

In this article, we concentrate on feature selection techniques but we also investigate PCA to give insights into performance differences between feature selection and dimensionality reduction techniques. More precisely, we investigate how the efficiency of SCA distinguishers can increase due to feature selection techniques. For this, we employ several feature selection techniques ranging from "simple" ones like the Pearson correlation, which is a *de facto* standard in the side-channel community, to more complex approaches such as various wrapper and hybrid methods used in ML. To the best of the authors' knowledge, the use of such advanced techniques has never been reported in the context of SCA before.

We show that feature selection is an important step in profiled attacks. We give insights on its use for the following goals: 1) faster training of models; 2) reducing model complexity; 3) improving model performance (when suitable features are selected); 4) reducing overfitting; and 5) "correcting" the covariance matrix in TA when the number of features is too large with respect to the number of traces.

A. Our Contributions

1) We introduce a novel approach of using ML techniques for the important problem of feature selection in SCA.
2) We demonstrate the potential of wrapper and hybrid methods in SCA as they often perform the best for feature selection on the examined data sets.
3) We show how to overcome some previously identified shortcomings of TAs by the ML techniques, which not only solves the problems but also improves upon the performance of templates as well.
4) We show that our feature selection methods may also be used for dimensionality reduction, having similar or better results than PCA in most cases.
5) All our results are verified on the real-world data sets in different settings. The analysis is explained in detail. In total, we consider and run more than 600 experimental scenarios in this work.

B. Previous Work

Ever since the seminal work of Chari *et al.* [14] introducing TAs, efforts were put into optimizing those and enlarging their scope. The observation on the profiling, i.e., training phase in TAs, has naturally led to ML techniques and their potential impact on the key recovery phase.

With that respect, a number of ML techniques has been investigated (see [1]–[3]). The results suggested the unquestionable potential of ML techniques for templates and, as such, they were stimulating for further research. However, the limitations of ML approaches were unveiled and their full potential remained unclear. Lerman *et al.* [15], in particular, compared TAs and ML on dimensionality reduction. They concluded that TAs are the method of choice as long as a limited number of features can be identified in leakage traces containing most of the relevant information. Accordingly, an increase in the number of points of interest favors ML methods. Our results show that the answer is not so simple, i.e., it depends on several factors, such as the number of features, classifiers, implementation details, and so on.

Regarding the feature selection problem in SCA, there were very few attempts and works devoted to this topic, as some simple techniques were considered sufficient. Early works introduced SOST/SOSD [4] as feature selection methods and the majority of follow-up papers skipped this step completely. One strategy also suggested using the Pearson correlation for this purpose [1], [2], [5], which is an obvious solution, but does not answer the question on whether we can do better.

Some authors noticed the importance of finding adequate time points in other scenarios. Reparaz *et al.* [16] introduced a technique to identify tuples of time samples before key recovery for multivariate differential power analysis (DPA) attacks. Here, typically, the attacker is confronted with a masked implementation, requiring higher order attacks (hence, multiple features corresponding to the right time moments, e.g., when a mask is generated and manipulated). Zheng *et al.* [12] looked into this specific feature selection question but left ML techniques aside. Picek *et al.* [17] considered several ML techniques for profiling attacks and investigated the influence of the number of features in the process by applying information gain feature selection. Finally, we also question the previous results on dimensionality reduction as our comparison of ML feature selection and PCA [18] (which is feature extraction) favors the former.

II. Background

A. Notation

Calligraphic letters (e.g., $\mathcal{X}$) denote sets, capital letters (e.g., X) denote random variables taking values in those sets, and the corresponding lowercase letters (e.g., x) denote their realizations. Let k^* be the fixed secret cryptographic key (byte) and the random variable T is the plaintext or ciphertext of the cryptographic algorithm, which is uniformly chosen. The measured leakage is denoted as X and we are particularly interested in multivariate leakage $\vec{X} = X_1, \ldots, X_D$, where D is the number of time samples or features (attributes) in ML terminology.

Considering a powerful attacker who has a device and the knowledge on the secret key implemented, a set of N profiling traces $\vec{X}_1, \ldots, \vec{X}_N$ is used to estimate the leakage model beforehand. Note that this set is multi-dimensional (i.e., it has dimension $D \times N$). In the attack phase, the attacker then measures additional traces $\vec{X}_1, \ldots, \vec{X}_Q$ from the device under attack to recover the unknown secret key k^*.

B. Data Sets

We use three data sets that we consider to be a representative sample of commonly encountered scenarios. More precisely, one data set is without countermeasures and with a small amount of noise, which is a relatively easy scenario for profiled attack when the number of measurements is sufficient. Next, we consider one data set without countermeasures but with a large amount of noise. There, we are approaching more realistic scenarios where profiled techniques have problems in reaching high performance. Finally, the last data set has a countermeasure in the form of random delays, which represents a realistic scenario for evaluation. We do not consider data sets with masked implementations since we assume that the mask is different in the training and testing phase, which makes feature selection more complex.

1) DPAcontest v4 Data Set [19]: The fourth version of the DPAcontest data set provides measurements of a masked AES software implementation. As the mask is known, one can easily turn it into an unprotected scenario. As this is a software implementation, the most leaking operation is not the register writing but the processing of the S-box operation, and thus the attack targets the first round. Hence, the leakage model is

$$Y(k^*) = \texttt{Sbox}[P_{b_1} \oplus k^*] \oplus \underbrace{M}_{\text{known mask}} \tag{1}$$

where P_{b_1} is a plaintext byte and we choose $b_1 = 1$. Compared to the measurements from the second version of the data set, the SNR is much higher with a maximum value of 5.8577. For our experiments, we start with a preselected window of 3500 features from the original trace (we simply preselect all features around the S-box operation).

2) AES_HD Data Set: This data set is chosen in order to target an unprotected implementation of AES-128 encryption specification. The core of AES-128 was written in VHDL in a round-based architecture, taking 11 clock cycles for each encryption. A universal asynchronous receiver–transmitter (UART) module is wrapped around the core to enable external communication. The module is designed to allow accelerated measurements so avoid any dc shift due to environmental variation over prolonged measurements. The total area footprint of the design contains 1850 lookup tables (LUTs) and 742 flip-flops. Xilinx Virtex-5 field-programmable gate array (FPGA) of a SASEBO GII evaluation board was used to implement the design. Side-channel traces were measured using a high sensitivity near-field electromagnetic (EM) probe, which was placed over a decoupling capacitor on the power line. Measurements were sampled on the Teledyne LeCroy Waverunner 610zi oscilloscope. A suitable and commonly used (HD) leakage model, when attacking the last round of an unprotected hardware implementation, is the register writing in the last round [19], that is,

$$Y(k^*) = HW(\underbrace{\texttt{Sbox}^{-1}[C_{b_1} \oplus k^*]}_{\text{previous register value}} \oplus \underbrace{C_{b_2}}_{\text{ciphertext byte}}) \tag{2}$$

where C_{b_1} and C_{b_2} are two ciphertext bytes, and the relation between b_1 and b_2 is given through the inverse ShiftRows operation of AES. $b_1 = 12$ was chosen, which resulted in $b_2 = 8$, as it is one of the easiest bytes to attack. The obtained measurements that form the data set are relatively noisy and the resulting model-based SNR (signal-to-noise ratio), i.e., $(\text{var}(\text{signal})/\text{var}(\text{noise})) = (\text{var}(y(t, k^*))/\text{var}(x - y(t, k^*)))$ has a maximum value of 0.0096. In total, 500 000 traces were captured corresponding to 500 000 randomly generated plaintexts, each trace with 1250 features. However, not all the traces were used for training and testing the model. The evaluation details are given in Section IV. As this implementation leaks in the HD model, we denote this implementation as AES_HD. The data set is publicly available at https://github.com/AESHD/AES_HD_Data set.

3) Random Delay Data Set [20]: As our third use case, we use an actual protected implementation to prove the potential of our approach. Our target is a software implementation of AES on an 8-bit Atmel AVR microcontroller with implemented random delay countermeasure, as described by Coron and Kyzhvatov [20]. We mounted our attacks against the first AES key byte by targeting the first S-box operation. The data set consists of 50 000 traces of 3500 features each. For this data set, the SNR has a maximum value of 0.0556. This data set is publicly available at https://github.com/ikizhvatov/randomdelays-traces.

C. Profiled Attacks and Guessing Entropy

In this section, we introduce the methods we use in the classification tasks. Note that we opted to work with only a small set of techniques, since we aim to explore how to find the best possible subset of features, while the classification task should be considered as just a means of comparison among feature selection methods. Consequently, we try to be as "method-agnostic" as possible and we note that for each set of features, one could probably find a classification algorithm performing slightly better. As noted in [21], there is no need to include many classifiers to obtain the best solutions. Generally, one of the best classifiers suffices, which is certainly the random forest (RF) algorithm. We use RF for classification in all the experiments since it provides stable and accurate results [3], [17]. Also, the linear kernel support vector machine (SVM) is used because of its efficiency and accuracy for wrapper and hybrid-based feature selection, as explained in continuation. As mentioned in Section I-B, the TA (i.e., TA classifier) is the traditional method of choice in SCA, especially when the number of features is small. Consequently, we use TA classifier and its pooled version [22] for comparison with RF.

1) Random Forest: RF is a well-known ensemble decision tree learner [23]. Decision trees choose their splitting attributes from a random subset of k attributes at each internal node. The best split is taken among these randomly chosen attributes and the trees are built without pruning. RF is a stochastic algorithm because of its two sources of randomness: bootstrap sampling and attribute selection at node splitting. Learning time complexity for RF is approximately $\mathrm{O}(I \cdot k \cdot N \cdot \log N)$, where I is the number of trees in the forest, k is the number of features considered at each node in each tree (usually $k = \sqrt{D}$, D being the total number of features), and N is the number of samples. We use RF as the classifier of choice for multiclass classification in our work. This is mainly in line with available research [21], where it is expected that RF will perform among the best classifiers. RF is used in all evaluations of the reduced sized feature sets.

2) Support Vector Machines: SVM is a kernel-based ML family of methods used to accurately classify both linearly separable and linearly inseparable data [24]. The basic idea when the data are not linearly separable is to transform them to a higher dimensional space by using a transformation kernel function. In this new space, the samples can usually be classified with higher accuracy. We use SVM with the linear kernel as the classification algorithm for wrapper and hybrid-based feature selection (see Sections III-B and III-C). Linear kernel SVM is used instead of a polynomial- or radial-based SVM, because advanced feature selection approaches require the construction of many models, which is computationally intensive and, therefore, unsuitable for nonlinear kernel function-based SVM. The time complexity range for linear kernel SVM is O(DN), which is significantly less than O(DN^3) for the time complexity of radial kernel SVM. We note that utilizing a linear kernel is an efficient choice when the number of dimensions is high (as in our case) or when we can assume that there is a linear separation between data.

3) Template Attack: The TA relies on the Bayes theorem and considers the features as dependent. In the state-of-the-art, TA relies mostly on a normal distribution. Accordingly, TA assumes that each $P(\vec{X} = \vec{x} | Y = y)$ follows a (multivariate) Gaussian distribution that is parameterized by its mean and covariance matrix for each class Y. Choudary and Kuhn [22] propose to use only one pooled covariance matrix averaged over all classes Y to cope with statistical difficulties and thus a lower efficiency. Besides the standard approach, we additionally use this version of the TA in our experiments. The time complexity for TA is O(ND^2) in the training phase and O($|\mathcal{Y}|D^2$) in the testing phase ($|\mathcal{Y}|$ is the number of classes).

D. Guessing Entropy

After running profiled attacks, we obtain accuracy as the measure of performance for our classifiers. Since this measure can be often misleading in SCA, especially in the Hamming weight (HW) scenario [7], we also use the guessing entropy (GE) to properly assess the performance of our feature selection and classification techniques [25]. A side-channel adversary $A_{E_K,L}$ conducts experiment $\mathsf{Exp}_{A_{E_K,L}}$, with time-complexity τ, memory complexity m, and making Q queries to the target implementation of the cryptographic algorithm. The attack outputs a guessing vector g of length o and is considered a success if g contains the correct key k^*. o is also known as the order of the success rate.

GE measures the average number of key candidates to test after the attack. The GE of the adversary $A_{E_k,L}$ against a key class variable S is defined as

$$GE_{A_{E_K,L}}(\tau, m, k^*) = \mathsf{E}[\mathrm{Exp}_{A_{E_K,L}}].$$

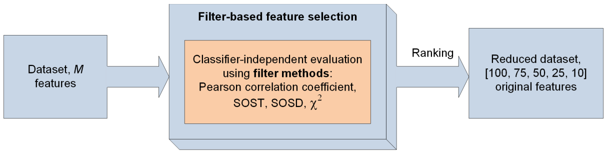

Filter-based feature selection
Dataset, *M* features
Classifier-independent evaluation using **filter methods**: Pearson correlation coefficient, SOST, SOSD, χ^2
Ranking
Reduced dataset, [100, 75, 50, 25, 10] original features

Fig. 1. Filter methods.

III. Feature Selection Techniques

A successful feature selection algorithm should output an optimal or near-optimal subset of features while ignoring the rest. Such algorithms can be classified into three broad classes of feature selection techniques: filter methods, wrapper methods, and hybrid methods [26]. The wrapper and hybrid classes of methods are known to either increase or retain the accuracy of the filter methods [26], [27].

Only the first three presented filter methods (Pearson correlation coefficient, SOSD, SOST) have been used as feature selection techniques for side-channel analysis in previous works, whereas the remaining methods, to the best of the authors' knowledge, have never been studied to find the most important features in SCA traces. We consider methods from all three classes of feature selection techniques in order to cover a wide set of feature selection cases. The choice of individual methods from these classes is based on our previous experience and the fact that all the methods are well-established in the field of feature selection, as noted in Sections III-A–III-d. We also consider in this section the PCA. Although PCA is, strictly speaking, dimensionality reduction and not feature selection technique, we compare it with the feature selection methods, because it is often used in SCA attacks.

A. Filter Selection Methods

The selection of features using filter methods is independent of the classifier method. Features are selected based on their scores obtained after running various types of statistical tests. We depict the filter methods principle in Fig. 1, with methods and numbers pertaining to our work.

Pearson Correlation Coefficient: It measures linear dependence between two variables, x and y, in the range $[-1, 1]$, where 1 is the total positive linear correlation, 0 is no linear correlation, and -1 is the total negative linear correlation. The Pearson correlation for a sample of the entire population is defined by [28]

$$\text{Pearson}(x, y) = \frac{\sum_{i=1}^{N}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \tag{3}$$

We calculate the Pearson correlation for the target class variables HW and intermediate value, which consists of categorical values that are interpreted as numerical values. The features are ranked in descending order of the coefficient.

SOSD: Gierlichs *et al.* [4] proposed the SOSD as a selection method, simply as

$$\text{SOSD}(x, y) = \sum_{i,j>i} (\bar{x}_{y_i} - \bar{x}_{y_j})^2 \tag{4}$$

where $\bar{x}_{y_i}$ is the mean of the traces where the model is equal to y_i. Because of the square, SOSD is always positive. Another advantage of using it is to emphasize big differences in means.

SOST: It is the normalized version of SOSD [4] and is thus equivalent to the pairwise student's t-test

$$\mathrm{SOST}(x, y) = \sum_{i,j>i} \left((\bar{x}_{y_i} - \bar{x}_{y_j}) \Big/ \sqrt{\frac{\sigma_{y_i}^2}{n_{y_i}} + \frac{\sigma_{y_j}^2}{n_{y_j}}} \right)^2 \tag{5}$$

with n_{y_i} and n_{y_j} being the number of traces where the model is equal to y_i and y_j, respectively.

Chi-Square: (χ^2) is a measure of dependence between two stochastic variables. It is a cumulative test statistic, which asymptotically approaches a χ^2 distribution. In the general case, χ^2 distribution may be obtained from the sum of squares of the set of k standard normal random variables, where k are the degrees of freedom. χ^2 test statistic for each feature-class pair may be calculated using the expression

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_{y_i} - E_{y_i})^2}{E_{y_i}}. \tag{6}$$

Here, n is the number of discrete categories, x_{y_i} is the observed value of category y_i, and E_{y_i} is the expected (theoretical) frequency of category y_i. Note that, for numerical features, the values need to be discretized to obtain categories before calculation of the statistic. By using the statistic, we proceed to remove the features that are the most likely to be independent of class attribute and therefore irrelevant for classification. Finally, since this measure works only for nonnegative values, before using it, we normalize the data into $[0, 1]$ range. The complexity of calculating the measure is $O(N \cdot D)$.

B. Wrapper Selection Methods

In wrapper methods, there is a feature selection algorithm implemented as a wrapper around a classifier [29]. The feature selection algorithm searches for a good subset by using a classifier algorithm as a part of the function evaluating feature subsets, as depicted in Fig. 2. Here, the classifier algorithm is considered as a black box and is run on the data set with different sets of features removed from the data. The subset of features with the highest evaluation is chosen as the final set on which to run the classifier [30]. Note that, since wrapper methods check many different subsets, the feature selection process is often treated as a high-dimensional problem. L1 regularization with linear SVM is used for wrapper-based feature selection in all the experiments, because the combination is sufficiently fast, accurate, and memory-undemanding. The other potential candidates that could have been used are naive Bayes, linear SVM, and k-nearest neighbors. However, although the sole use of these classifiers may be faster compared to L1 regularization with linear SVM, they may not be as accurate in estimating the accuracy of feature subsets. On the other hand, methods such as RF, neural network, nonlinear SVM, etc., are more complex and are not typically used as wrappers, since they exhibit nonlinear complexity dependence on the number of instances.

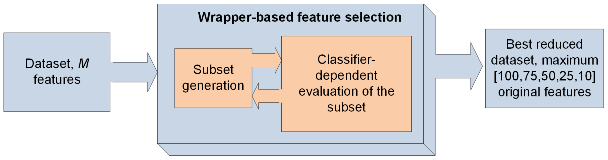

Wrapper-based feature selection
Dataset, *M* features
Subset generation
Classifier-dependent evaluation of the subset
Best reduced dataset, maximum [100,75,50,25,10] original features

Fig. 2. Wrapper methods.

L1-Based Feature Selection: In general, regularization encompasses methods that add a penalty term to the model, which then reduces the overfitting and improves generalizations. L1 regularization works by adding a regularization term $\alpha \cdot R(\theta)$, where θ represents the parameters of the model that is used to penalize large weights/parameters. For a D-dimensional input (i.e., the number of features equal to D), $R(\theta)$ is equal to $\sum_{i=1}^{D} |\theta_i|$. In the regularization term, α controls the tradeoff between fitting the data and having small parameters. By adding a penalty for each nonzero coefficient, the expression forces weak features to have zero as coefficients, where a zero value means that the feature is omitted from the set. The usage of L1 regularization as a tool for feature selection is well known, for example, the linear least-squares regression with L1 regularization (Lasso) algorithm [31]. There can be certain effects with L1 regularization when used for feature selection: most notably, out of a group of highly correlated features, L1 regularization will tend to select an individual feature [32].

C. Hybrid Selection Methods

Hybrid methods combine filter and wrapper techniques. First, a filter method is used in order to reduce the feature space dimension space. Then, a wrapper method is utilized to find the best candidate subset. Hybrid methods usually achieve high accuracy that is characteristic to wrappers and high-efficiency characteristic to filters. We depict a diagram for hybrid methods, as used in this article, in Fig. 3. In our experiments, we first use χ^2 to reduce the number of features to 250 in order to further reduce the runtime of hybrid selection techniques. Then, we apply either the linear SVM selection or the stability selection technique.

Linear SVM-Based Hybrid Selection: We use a recursive feature elimination approach with linear SVM wrapper to obtain the target reduced feature sets. The method was first described in [33]. Here, the "best-first" backward direction search method is used. This strategy uses greedy hill climbing, starting from the full feature subset and inspecting how the elimination of a feature or a set of features from the starting set influences the output of the classifier. The feature(s) whose removal influences the accuracy the least is eliminated from the set.

Stability Selection: It is a method based on subsampling in combination with some classification algorithm (that can work with high-dimensional data) [34]. The key concept of stability selection is the stability paths, which is the probability for each feature to be selected when randomly resampling from the data. In other words, a subsample of the data is fitted to the L1 regularization model, where the penalty of a random subset of coefficients has been scaled. By repeating this procedure n

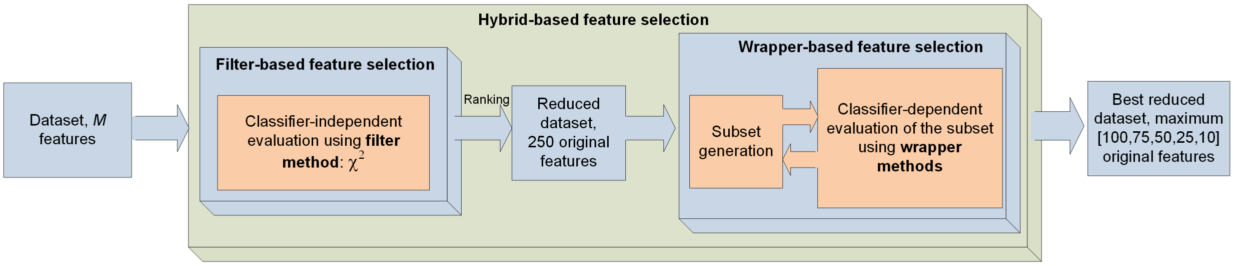

Hybrid-based feature selection
Filter-based feature selection
Wrapper-based feature selection
Dataset, *M* features
Classifier-independent evaluation using **filter method**: χ^2
Ranking
Reduced dataset, 250 original features
Subset generation
Classifier-dependent evaluation of the subset using **wrapper methods**
Best reduced dataset, maximum [100,75,50,25,10] original features

Fig. 3. Hybrid methods.

times, the method will assign high scores to the features that are repeatedly selected. We use multinomial logistic regression for this task and we set the number of randomized models n to 25. Multinomial logistic regression uses a linear predictor function $f(k, i)$ to predict the probability that observation i has the outcome k, of the form $f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \ldots + \beta_{M,k}x_{M,i}$ where $\beta_{M,k}x_{M,i}$ is a regression coefficient of the mth variable and the kth outcome. The β coefficients are estimated using the maximum likelihood estimation, which requires finding a set of parameters for which the probability of the observed data is the greatest.

D. Principal Component Analysis

PCA is a well-known linear dimensionality reduction method that may use singular value decomposition (SVD) of the data matrix to project it to a lower dimensional space [18]. PCA creates a new set of features (called principal components) that are linearly uncorrelated, orthogonal, and form a new coordinate system. The number of components is equal to the number of original features. The components are arranged in a way that the first component covers the largest variance by a projection of the original data and the subsequent components cover less and less of the remaining data variance. The number of kept components, designated with L, maximizes the variance in the original data and minimizes the reconstruction error of the data transformation. The Python implementation of PCA uses either the Linear Algebra Package (LAPACK) implementation of the full SVD or a randomized truncated SVD by the method of Halko *et al.* [35], depending on the shape of the input data and the number of components selected to extract. We experiment with L values in the range $[10, 25, 50, 75, 100]$.

TABLE I

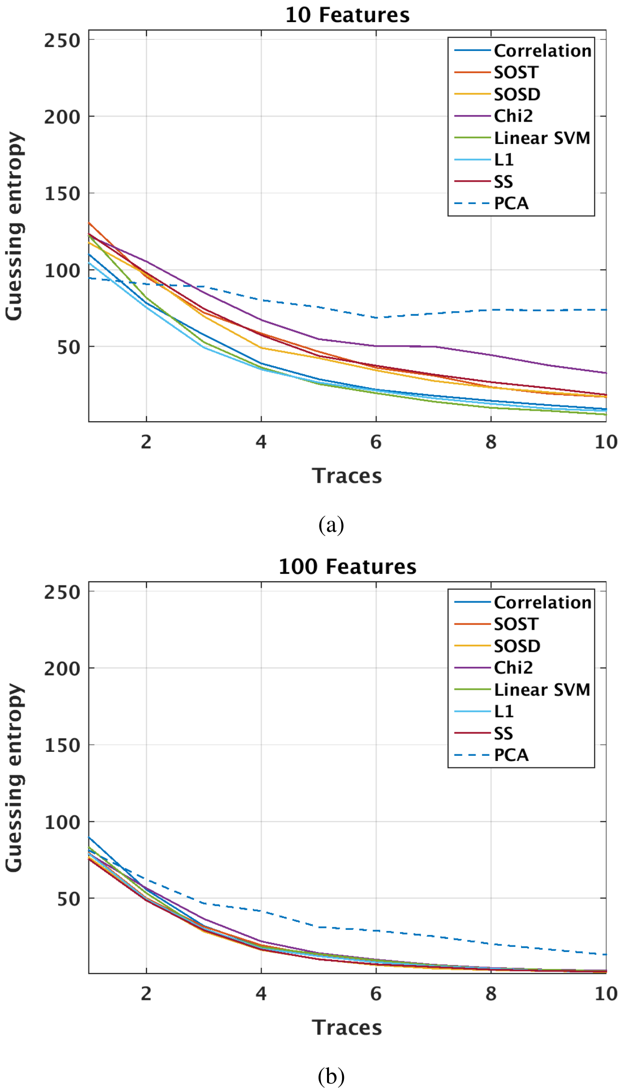
ACCURACY FOR DPACONTEST V4—HW MODEL

Pearson correlation					
Classifier	10	25	50	75	100
TA	69.8	72.5	0.3	2.8	41.8
TA (pooled)	68.5	71.7	80.9	81.7	91.4
RF	74.5	81.4	84.6	84.1	85.8
SOST					
Classifier	10	25	50	75	100
TA	71.5	73.9	0.5	20	0.2
TA (pooled)	69.4	73.4	80.6	86.6	91.6
RF	74.2	81.4	84.3	84.7	86
SOSD					
Classifier	10	25	50	75	100
TA	72.2	74.7	8.7	8.7	3.8
TA (pooled)	69.8	74.4	77.3	84.5	89.6
RF	75.9	81.6	82.5	83.7	84.4
χ^2					
Classifier	10	25	50	75	100
TA	72.2	74.3	36.9	30.5	0.5
TA (pooled)	69.8	74.3	81.1	84.9	91.6
RF	76.2	81.4	84.7	84.5	86.4
Linear SVM wrapper					
Classifier	10	25	50	75	100
TA	19.9	51.6	1.5	4.6	1.3
TA (pooled)	14	49.7	85.3	98.1	98.1
RF	89.7	91.9	92.3	91.6	91.2
L1 regularization					
Classifier	10	25	50	75	100
TA	8.4	32.1	1.3	91.5	11
TA (pooled)	9.6	27.4	90.1	97	97.3
RF	80.4	86.7	88.8	89.8	89.4
Stability selection					
Classifier	10	25	50	75	100
TA	20.7	31.3	0	86.2	30.3
TA (pooled)	16.3	28.9	92.2	97.4	98
RF	75.3	91.4	91.5	91.2	90.7

IV. EXPERIMENTAL EVALUATION

In our experiments, we are interested in supervised (profiled) problems that have a large number of features (sample points from power traces) D but where there could exist a small subset D' of features that is sufficient to classify the features X according to the classes Y. We use the previously described filter, wrapper, and hybrid methods to reduce the number of features found in the original data sets to the smaller subsets of sizes $[10, 25, 50, 75, 100]$. The investigated subset sizes are selected based on the usual number of features considered in related work (see Section I-B). We have also tried increasing the number of features, inspecting up to 200 features. The results were not better and the analysis was prolonged. Specifically, the features in the range of 101–200 lead to no improvement in accuracy or GE with respect to only the first 100 included features, for all methods.

Once the best feature subsets are selected, we run three profiled attacks: RF, TA, and TA pooled (TA$_p$) for each feature selection technique to evaluate its efficiency. We use multiple profiled attacks to avoid potential effects that a certain feature selection technique could have on a specific attack. We emphasize that the goal, here, is not to compare the efficiency of attacks and, consequently, we do not give such an analysis.



(a)

(b)

Fig. 4. GE, DPAcontest v4 data set. (a) 10 features, HW, RF. (b) 100 features, HW, RF.

TABLE II

ACCURACY FOR DPACONTEST V4—INTERMEDIATE VALUE MODEL

		Pearson correlation			
Classifier	10	25	50	75	100
TA	11.4	0.4	0.4	0.2	0.4
TA (pooled)	15.8	18.0	20.5	31.1	53
RF	13	20.4	25	29.8	36.2
		SOST			
Classifier	10	25	50	75	100
TA	11.5	0.1	0.1	0.1	0
TA (pooled)	16.2	32.6	51.7	62.3	64.3
RF	15.3	32.5	38.4	42.2	43.1
		SOSD			
Classifier	10	25	50	75	100
TA	17.9	0.3	0.1	0	0.1
TA (pooled)	23.2	38	56.1	64.4	65.7
RF	18	30.1	39.9	41.2	42.1
		χ^2			
Classifier	10	25	50	75	100
TA	1.6	0.3	0.1	0.2	0.2
TA (pooled)	1.6	3.7	28.4	57.1	69
RF	23.9	34.7	41.2	44	45.8
		Linear SVM wrapper			
Classifier	10	25	50	75	100
TA	26.8	20.2	0.	0.1	0
TA (pooled)	24.3	43.9	64.9	71	74.3
RF	24.6	44.5	70.8	74.2	75.5
		L1 regularization			
Classifier	10	25	50	75	100
TA	28.7	0	0.2	0	0
TA (pooled)	26.1	51.8	66.9	74	75.6
RF	28.8	53.1	73.9	74.4	75.3
		Stability selection			
Classifier	10	25	50	75	100
TA	25.2	0.2	0	0.3	0.1
TA (pooled)	21.4	47.4	64.3	73.1	75.7
RF	24.8	46.9	65.6	71	75.2

Finally, we note that for the wrapper methods, selecting the exact number of features can be difficult (since the methods can simply discard multiple features) and, consequently, subset sizes of [10, 25, 50, 75, 100] represent an upper bound on the number of actually selected features.

From the initial data sets, we randomly select 10 000 power traces for training and another 25 000 randomly selected traces for testing. We opted to have a larger test set to obtain meaningful results with GE. For evaluation on the training set, we conduct fivefold cross-validation and use the averaged results of individual folds to select the best classifier parameters. We report the results from testing phase only and we present them as the accuracy (%) of the classifier, where the accuracy is the number of correctly classified traces divided by the total number of traces. All experiments are done with MATLAB and Python (scikit-learn library) tools. For the L1 regularization with linear SVM wrapper, hybrid linear SVM, and hybrid stability selection, we tune the parameter C for each subset size. For linear SVM, we further select the step equal to 5 to remove features—in each iteration of the algorithm, we discard five least important features from the feature set. For RF, we experiment with $I = [10, 50, 100, 200, 500, 1000]$ trees in the tuning phase, with no limit to tree size. Based on the tuning phase, we select 500 trees for the HW model and 100 trees for the intermediate value model.

A. Results

We give results for test set accuracy in Tables I–VII and for GE in Figs. 4–6. Due to the lack of space, we do not show GE results for all tested scenarios, but only for a representative subset of them. For each size of the feature subset in Tables I–VI, we give the best-obtained solution in a cell with the gray background color. For Table VII, the gray background of a cell indicates a better result for PCA than for all feature selection methods.

1) DPAcontest v4 Data Set: Tables I and II display the results for DPAcontest v4 with the HW model and intermediate value model, respectively. For the HW model, we observe that

TABLE III
ACCURACY FOR AES_HD—HW MODEL

Pearson correlation

Classifier	10	25	50	75	100
TA	11.4	17.7	5.3	5	1.5
TA (pooled)	4.4	5.6	7.9	8.2	8.9
RF	23.8	24.7	25.2	25.3	25.5

SOST

Classifier	10	25	50	75	100
TA	10.5	17.7	5.4	13	11.1
TA (pooled)	4.3	5.7	7.9	8	9.5
RF	23.7	24.6	24.6	25	25.4

SOSD

Classifier	10	25	50	75	100
TA	10.5	18.4	1.4	0.7	0.6
TA (pooled)	4.3	6	8.2	8.8	9.5
RF	23.7	25.2	25.9	26.4	26.2

χ^2

Classifier	10	25	50	75	100
TA	11.4	18.2	4.1	2.3	1.4
TA (pooled)	4.4	6.4	7.8	9.1	9.5
RF	23.8	25	24.8	25.7	25.3

Linear SVM wrapper

Classifier	10	25	50	75	100
TA	11	16.3	1.9	10.2	4.3
TA (pooled)	5.1	6	8.2	8.9	9.9
RF	24.4	24.9	25.4	25.8	25.8

L1 regularization

Classifier	10	25	50	75	100
TA	10.1	16.3	7.1	3.3	7.9
TA (pooled)	5.5	7.5	8.1	9.3	9.9
RF	23.6	24.9	25.3	26	25.7

Stability selection

Classifier	10	25	50	75	100
TA	11	16.1	8.5	3.8	10.4
TA (pooled)	5.6	5.7	6.9	7.8	8.2
RF	24.7	25.8	25.7	25.8	26

TABLE IV
ACCURACY FOR AES_HD—INTERMEDIATE VALUE MODEL

Pearson correlation

Classifier	10	25	50	75	100
TA	0.3	0.4	0.4	0.3	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.4	0.4	0.3

SOST

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.4
TA (pooled)	0.4	0.5	0.4	0.5	0.5
RF	0.3	0.4	0.4	0.4	0.3

SOSD

Classifier	10	25	50	75	100
TA	0.3	0.4	0.3	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.3	0.4	0.4	0.4	0.5

χ^2

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.4	0.5	0.4

Linear SVM wrapper

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.5
RF	0.4	0.4	0.4	0.4	0.4

L1 regularization

Classifier	10	25	50	75	100
TA	0.3	0.4	0.4	0.5	0.5
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.4	0.4	0.4

Stability selection

Classifier	10	25	50	75	100
TA	0.4	0.4	0.5	0.4	0.4
TA (pooled)	0.4	0.4	0.3	0.4	0.5
RF	0.4	0.4	0.4	0.4	0.5

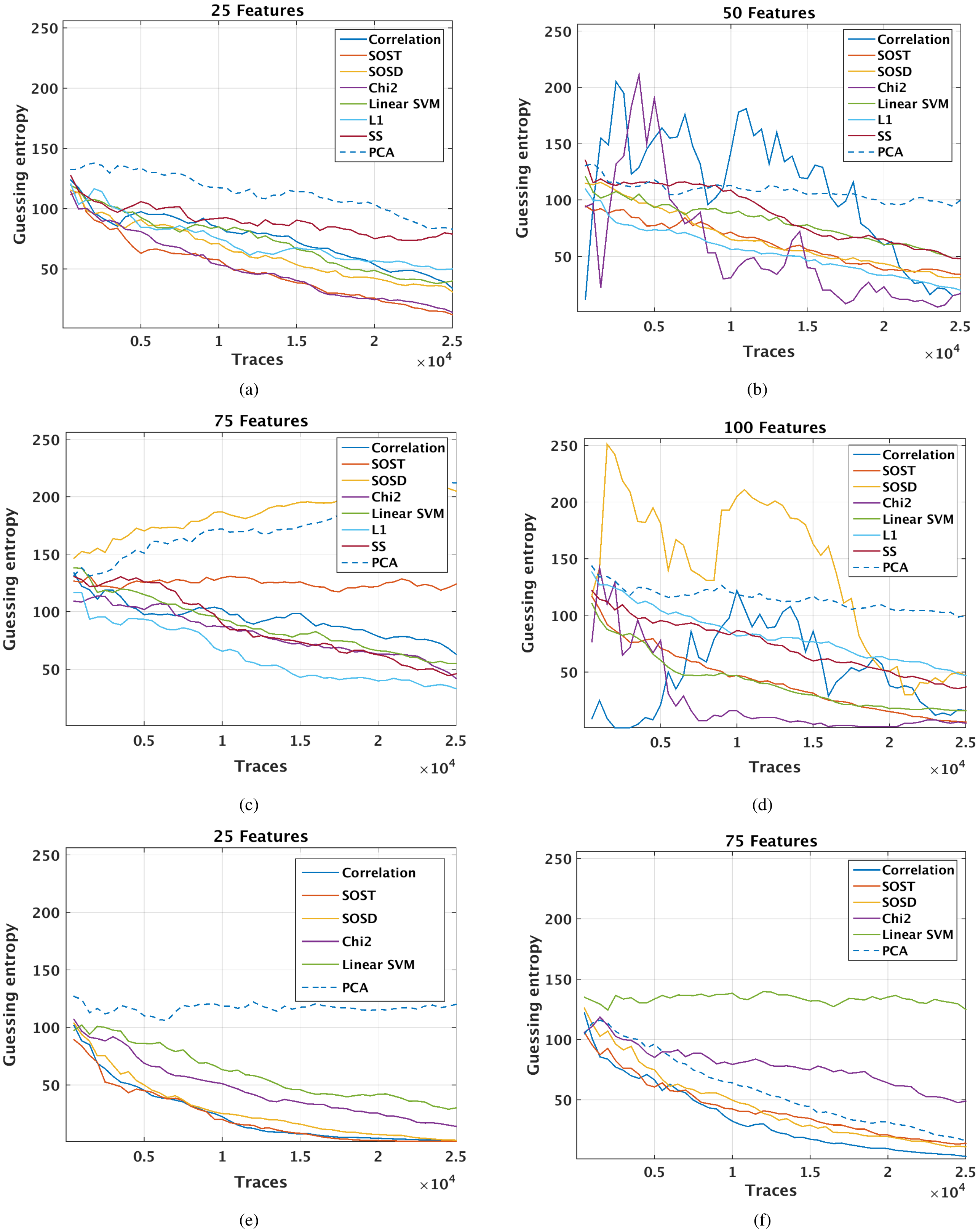
linear SVM hybrid method is, by far, the best performing feature selection method when considering accuracy, comparable, or outperformed only slightly by PCA for a larger number of features (see the first row of Table VII). Linear SVM works very well for the low-noise scenario and when the number of classes is rather low (nine for the HW model). Note that the results for linear SVM are comparable to the results for L1 and stability selection for the intermediate value model (256 classes), thus suggesting that the method is more appropriate for the smaller number of classes.

Fig. 4 shows that, for GE, the changes between different techniques are rather small with an advantage of linear SVM, L1, and correlation using ten features. When considering 100 features, all techniques perform almost equivalently, except for PCA, which performs the worst. Due to the low noise present in this scenario, all the feature selection methods have found highly similar features, see Fig. 7. Comparing the results for 100 and 10 features, it is shown that when the number of features is large (i.e., 100), there is a higher chance that most of the informative features are included by all methods than when the number of features is small (i.e., 10). For ten features, there is a larger difference between the methods, indicating that some important features are omitted by some methods.

When considering the intermediate value model (see Table II), we observe that the wrapper and hybrid methods have the highest accuracy, outperforming filters and PCA. Here, even accuracy for 100 features varies significantly.

For GE in the intermediate value model, we observe the same phenomena as for the HW model: all the techniques are differing only slightly when considering a low number of features and become closer when more features are considered.

2) AES_HD Data Set: For AES_HD data set, we give results in Tables III and IV for HW model and intermediate value model, respectively. For HW model, some observations made for DPAcontest v4 also apply for AES_HD. We see that having more features also, in general, results in higher accuracy. Still, in some scenarios, accuracy for the smaller feature set size is even higher than for larger feature set sizes but those differences are rather small. Differing from DPAcontest v4, for AES_HD, we do not observe a significant



(a)

(b)

(c)

(d)

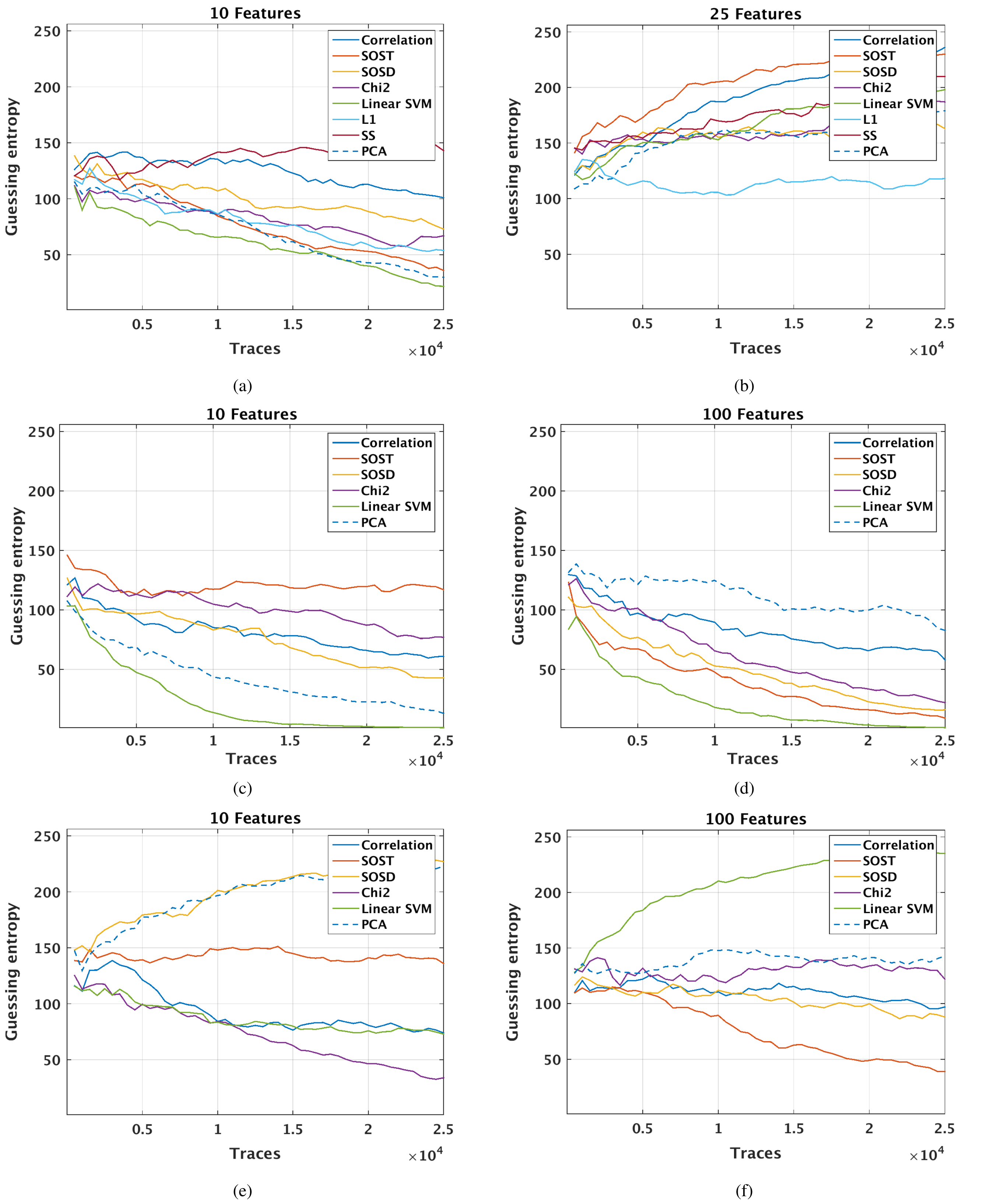
(e)

(f)

Fig. 5. GE, AES_HD. (a) 25 features, HW, TA. (b) 50 features, HW, RF. (c) 75 features, HW, TA. (d) 100 features, HW, RF. (e) 25 features, intermediate value, TA pooled. (f) 75 features, intermediate value, TA pooled.

drop in performance when using only ten features. PCA performs well for this case, slightly outperforming feature selection methods with respect to accuracy (see the third row of Table VII).

For the intermediate value model, the accuracy is very low and even looks like random guessing (1/256, see Table IV). The results show that there is no significant difference in behavior for any technique. This is expected, since there are



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6. GE, random-delay data set. (a) 10 features, HW, RF. (b) 25 features, HW, TA. (c) 10 features, intermediate value, RF. (d) 100 features, intermediate value, RF. (e) 10 features, intermediate value, TA pooled. (f) 100 features, intermediate value, TA pooled.

256 classes and only 10 000 measurements in the training phase, which is barely enough to have results better than random guessing when dealing with such difficult data sets.

We are able to reach a low GE (i.e., retrieve the secret key), as Fig. 5 clearly illustrates. More specifically, Fig. 5(a)–(f) depicts GE results for the AES_HD data set for HW

TABLE V

ACCURACY FOR RANDOM DELAY—HW MODEL

Pearson correlation

Classifier	10	25	50	75	100
TA	11.4	17.4	5.5	0.8	6.5
TA (pooled)	4.4	5.3	6.8	7.5	8.1
RF	25.3	26.1	26	26.2	26

SOST

Classifier	10	25	50	75	100
TA	9.4	15.7	5.5	5.7	4.5
TA (pooled)	5.9	6.2	8.5	9.3	9.8
RF	25.5	26	26.6	26.4	26.4

SOSD

Classifier	10	25	50	75	100
TA	10.1	17	0.5	8.2	14.6
TA (pooled)	6.6	7.9	8.8	9.5	9.9
RF	25.2	25.8	26.7	26.3	26.2

χ^2

Classifier	10	25	50	75	100
TA	9.6	16.6	2.5	8.5	10.6
TA (pooled)	5.9	6.9	8.3	9.1	9.5
RF	25	25.4	25.9	26	26.1

Linear SVM wrapper

Classifier	10	25	50	75	100
TA	7.2	15.5	1.7	1.9	7.6
TA (pooled)	4.7	5.9	7.0	7.5	8.3
RF	25.6	25.7	26.1	26.1	26.1

L1 regularization

Classifier	10	25	50	75	100
TA	10.7	15.9	1.3	7.3	6.3
TA (pooled)	6.3	6.5	7.8	8.7	9
RF	24.9	25.6	25.9	25.7	26.2

Stability selection

Classifier	10	25	50	75	100
TA	13.7	16.8	1.5	7.5	2.7
TA (pooled)	8	6.8	8.8	9.6	10
RF	24.8	25.5	25.8	25.8	26.1

TABLE VI

ACCURACY FOR RANDOM DELAY—INTERMEDIATE VALUE MODEL

Pearson correlation

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.3	0.5	0.4

SOST

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.3
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.5	0.4	0.4	0.3

SOSD

Classifier	10	25	50	75	100
TA	0.3	0.4	0.4	0.4	0.5
TA (pooled)	0.3	0.4	0.4	0.4	0.5
RF	0.4	0.4	0.4	0.4	0.5

χ^2

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.5
RF	0.4	0.4	0.4	0.4	0.4

Linear SVM wrapper

Classifier	10	25	50	75	100
TA	0.3	0.3	0.4	0.4	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.4	0.4	0.4

L1 regularization

Classifier	10	25	50	75	100
TA	0.3	0.4	0.4	0.4	0.5
TA (pooled)	0.4	0.4	0.5	0.4	0.4
RF	0.4	0.3	0.4	0.5	0.4

Stability selection

Classifier	10	25	50	75	100
TA	0.4	0.4	0.4	0.5	0.4
TA (pooled)	0.4	0.4	0.4	0.4	0.4
RF	0.4	0.4	0.4	0.4	0.4

TABLE VII

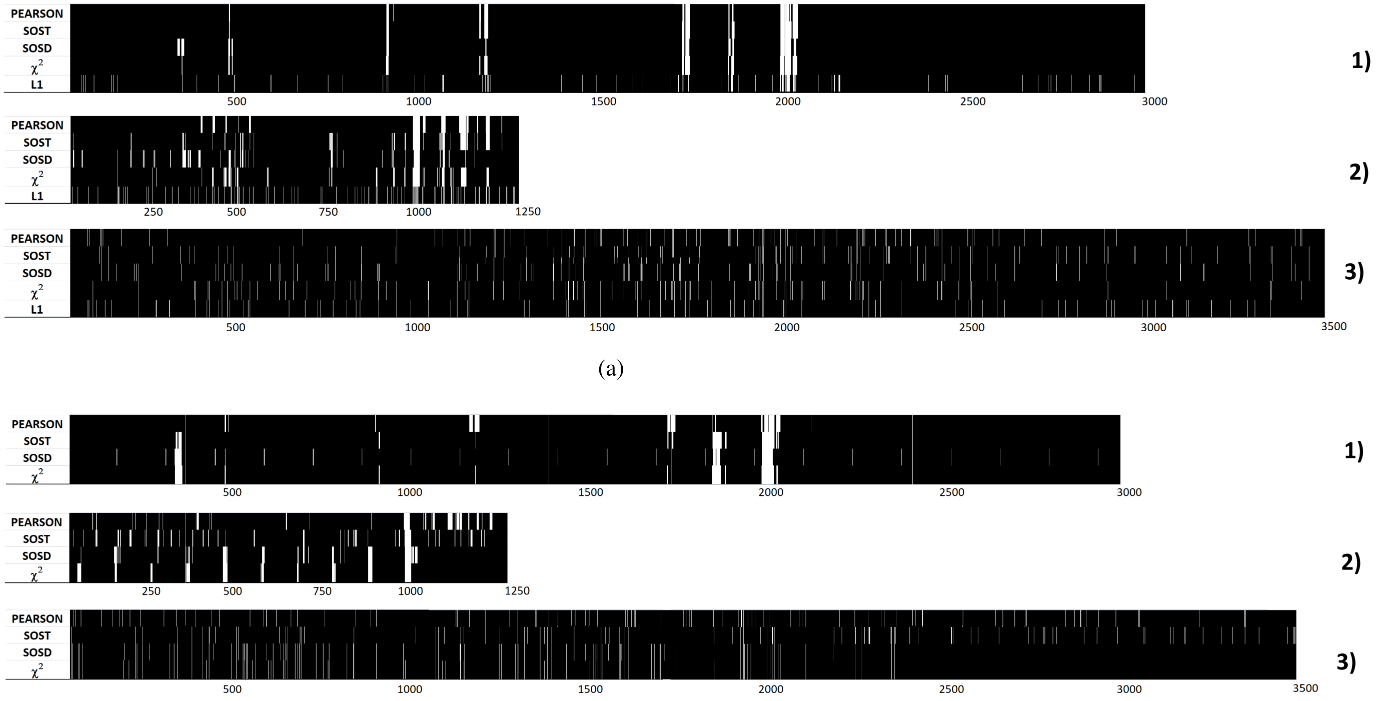
PCA CLASSIFICATION RESULTS

Accuracy, % (best classifier)

Dataset	10	25	50	75	100
DPA v4, HW	26.4(RF)	38.1(RF)	93.8(TA$_p$)	97.6(TA$_p$)	98.3(TA$_p$)
DPA v4, int.	0.7(RF)	4.0(RF)	40.1(TA$_p$)	61.3(TA$_p$)	74.7(TA$_p$)
AES_HD, HW	25.2(RF)	25.9(RF)	26.7(RF)	26.7(RF)	26.8(RF)
AES_HD, int.	0.4(all)	0.4(all)	0.5(RF)	0.4(all)	0.4(all)
Rand. D., HW	24.3(RF)	25.1(RF)	25.5(RF)	25.7(RF)	26.2(RF)
Rand. D., int.	0.4(all)	0.4(all)	0.4(all)	0.4(all)	0.4(all)

and intermediate value model ranging between 25 and 100 features. In this high-noise scenario, we observe a more distinct behavior for different techniques. Generally, PCA-based attack mostly performs comparably or worse than the feature selection techniques. In Fig. 5(b) and (d), one can observe that correlation for 50 features or correlation and SOSD for 100 features only become stable when using a large number of measurements in the attacking phase with RF. Fig. 5(e) and (f) shows that, despite approximately even accuracy for the intermediate model, there are marked differences among some methods with respect to GE. In these cases, when using TA pooled classifier, PCA, linear SVM, and Chi2 underperform with respect to other methods.

3) Random Delay Data Set: Finally, Tables V and VI give results for the random delay data set for HW and intermediate value model, respectively. For the HW model, the highest accuracies are spread among the feature selection methods. For example, for five scenarios, we have four different techniques reaching the highest accuracies. PCA performs slightly worse than feature selection methods for HW model. Fig. 6 shows that GE results are also widely spread. For HW, as well as for intermediate value model, linear SVM and L1 usually perform well; while in some rare cases, SOST also performs well, while linear SVM underperforms [Fig. 6(f)]. We can observe that, again, linear SVM is suitable when a small amount of features is selected [see Fig. 6(a) and (c)]. Comparing the results of RF and TA pooled classifiers for the intermediate value model, RF was shown to provide significantly more stable GE results. PCA-based attacks perform comparably to most feature selection methods on this data set.



(a)

(b)

Fig. 7. Hundred selected features for (a) HW model and (b) intermediate value model. 1) DPAcontest v4, 2) AES_HD, and 3) random delay.

4) Feature Illustration: In Fig. 7(a) and (b), we depict 100 selected features for all data sets, HW and intermediate value models, respectively. The visualization allows a more detailed inspection in the behavior of feature selection methods. For example, if different methods find similar features, then the selected features are probably globally more relevant than the others for the classification problem (assuming that not all the methods are wrong). If different methods find different features, while obtaining similarly good classification results, then this suggests that many features are informative enough to produce accurate models. If, however, different methods find different features, while obtaining different classification results (some better than others), then this suggests that some methods perform better selection than the others. For DPAcontest v4 and both considered models, a large part of the selected features for all techniques is the same. Consequently, the obtained results for both accuracy and GE are similar. This indicates that in a low-noise scenario, the choice among the feature selection methods is not crucial. For the AES_HD data set, we can observe that there are some regions where all the selection techniques find relevant features. Interestingly, for L1 regularization and HW model, the selected features are much less grouped when compared to the other selection techniques. The similarity in the selected features is reduced compared to the DPAcontest v4 data set. This indicates that, for the high-noise scenario, the choice of the methods is more important than for the low-noise one. Finally, for the random delay data set, all techniques select quite different features, which results in a significantly different performance, as seen in the GE results. This suggests that, for the high-noise with countermeasures scenarios, the choice of the feature selection method is very important; however, the overall results are still lower when compared to the less difficult scenarios.

B. General Observations

After presenting the results for different considered scenarios, we now concentrate on more general findings pertaining to feature selection in SCA.

1) Different feature selection techniques can result in a radically different classifier behavior, which is especially evident from the presented GE results. Consequently, one should devote the same amount of attention to feature selection as to classification. This is in line with the "No Free Lunch" theorem, which states that there is no single best algorithm for all problems [36].
2) It is important to conduct feature selection individually for each model considered. For instance, we show that, if feature selection is done for the HW model, then, in general, one should not use the same features when considering the intermediate value model.
3) We confirm that having a higher number of features than the number of traces per class results in TA becoming unstable, as also indicated by previous works [22], which is an observation that does not hold for ML techniques. In particular, for TA, we observe the effect of instability in the estimation of the covariance matrix when using the intermediate value model (= 256 classes) and if the number of features is > 10. The pooled version tries to circumvent instabilities by reducing the number of covariance matrices to be estimated to a single one, which may include information loss. We show an

alternative to increasing the number of traces or using only one pooled covariance matrix as suggested in [22]. More precisely, an alternative approach is to use one of the wrapper or hybrid techniques, which may result in improved performance of TA.

4) We show that even a very small subset of features, if selected properly, can result in better performance than a superset obtained with other selection techniques (that may contain redundant or incorrect features).
5) We show that it is possible to conduct feature selection even in the presence of a random delay countermeasure. There, although some important features are moved in the time domain, the amount of information obtained from traces is sufficient for a reliable feature selection, resulting in efficient attacks.
6) Data sets with large amounts of noise are difficult for classification as well as for feature selection. This is expected, especially for wrapper and hybrid methods, since there we use ML classifiers for feature selection.
7) When considering data sets with a large amount of noise or countermeasures, it is possible to conduct a successful attack even in extremely constrained scenarios where we have only ten features, if they are well chosen.

V. Conclusion

In this article, we addressed the following questions: how should we select the most informative features from raw data? and what is the influence of the feature selection step in the performance of the classification algorithm? Our results show that the proper selection of features has a tremendous impact on the final classification results. We notice that often with a small number of features when using a proper selection technique, one can achieve approximately the same results as some other method using a much larger number of features.

We demonstrated how state-of-the-art techniques for feature selection from the ML area behave for profiling in side-channel analysis. We observe that much more powerful techniques than those currently used in the SCA community are applicable and achieve higher accuracies. Unfortunately, our results do not reveal a single method as the best performing one. Still, this is to be expected, since the "No Free Lunch" theorem also holds for feature selection. We emphasize that the Pearson correlation is rarely the most successful technique for feature subset selection, which is a common choice for feature selection in the SCA community. When considering GE results, we emphasize the linear SVM hybrid method and L1 regularization wrapper that performed consistently well for all data sets. This is especially interesting, since L1 regularization did not perform the best when considering accuracy and the random delay and AES_HD data sets. Generally, feature selection in the case of "easy" scenarios (e.g., DPAcontest v4) is not the most important and effective task, but in scenarios with high noise and even countermeasures (random delay data set), our techniques may bring significant improvements.

The obtained accuracy results in most cases favor ML-based feature selection techniques when compared to PCA-based feature extraction. At the same time, when considering GE, we see that PCA is never the best technique. Future work may compare ML-based feature selection with other dimensionality reduction methods, e.g., SNR metrics [37], in detail and determine the superiority in specific contexts.

References

[1] A. Heuser and M. Zohner, "Intelligent machine homicide," in *Proc. COSADE*, in Lecture Notes in Computer Science, vol. 7275, W. Schindler and S. A. Huss, Eds. Berlin, Germany: Springer, 2012, pp. 249–264.

[2] L. Lerman, G. Bontempi, and O. Markowitch, "Power analysis attack: An approach based on machine learning," *Int. J. Appl. Cryptogr.*, vol. 3, no. 2, pp. 97–115, 2014.

[3] H. Maghrebi, T. Portigliatti, and E. Prouff, "Breaking cryptographic implementations using deep learning techniques," in *Security, Privacy, and Applied Cryptography Engineering* (Lecture Notes in Computer Science), vol. 10076, C. Carlet, M. Hasan, and V. Saraswat, Eds. Cham, Switzerland: Springer, 2016, pp. 3–26.

[4] B. Gierlichs, K. Lemke-Rust, and C. Paar, "Templates vs. stochastic methods," in *Cryptographic Hardware and Embedded Systems*, L. Goubin and M. Matsui, Eds. Berlin, Germany: Springer, 2006, pp. 15–29.

[5] S. Mangard, E. Oswald, and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards*. New York, NY, USA: Springer, 2006. [Online]. Available: http://www.dpabook.org/

[6] S. Picek, I. P. Samiotis, J. Kim, A. Heuser, S. Bhasin, and A. Legay, "On the performance of convolutional neural networks for side-channel analysis," in *Security, Privacy, and Applied Cryptography Engineering*, A. Chattopadhyay, C. Rebeiro, and Y. Yarom, Eds. Cham, Switzerland: Springer, 2018, pp. 157–176.

[7] S. Picek, A. Heuser, A. Jovic, S. Bhasin, and F. Regazzoni, "The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations," *IACR Trans. Cryptogr. Hardw. Embedded Syst.*, vol. 2019, no. 1, pp. 209–237, 2019.

[8] J. Kim, S. Picek, A. Heuser, S. Bhasin, and A. Hanjalic, "Make some noise. Unleashing the power of convolutional neural networks for profiled side-channel analysis," *IACR Trans. Cryptograph. Hardw. Embedded Syst.*, vol. 2019, no. 3, pp. 148–179, May 2019.

[9] E. Cagli, C. Dumas, and E. Prouff, "Convolutional neural networks with data augmentation against jitter-based countermeasures—Profiling attacks without pre-processing," in *Proc. 19th Int. Conf. Cryptograph. Hardw. Embedded Syst. (CHES)*, in Lecture Notes in Computer Science, vol. 10529, W. Fischer and N. Homma, Eds. Taipei, Taiwan: Springer, 2017, pp. 45–68.

[10] L. Lerman, G. Bontempi, and O. Markowitch, "A machine learning approach against a masked AES—Reaching the limit of side-channel attacks with a learning model," *J. Cryptograph. Eng.*, vol. 5, no. 2, pp. 123–139, 2015.

[11] S. Picek, A. Heuser, and S. Guilley, "Template attack versus Bayes classifier," *J. Cryptograph. Eng.*, vol. 7, no. 4, pp. 343–351, Nov. 2017.

[12] Y. Zheng, Y. Zhou, Z. Yu, C. Hu, and H. Zhang, "How to compare selections of points of interest for side-channel distinguishers in practice?" in *Proc. ICICS*, L. Hui, S. Qing, E. Shi, and S. Yiu, Eds. Cham, Switzerland: Springer, 2015, pp. 200–214.

[13] G. Becker *et al.*, "Test vector leakage assessment (TVLA) methodology in practice," in *Proc. Int. Cryptograph. Module Conf.*, Gaithersburg, MD, USA, 2013, p. 13.

[14] S. Chari, J. R. Rao, and P. Rohatgi, "Template attacks," in *Cryptographic Hardware and Embedded Systems—CHES* (Lecture Notes in Computer Science), vol. 2523, B. S. Kaliski, K. Koç, and C. Paar, Eds. Redwood City, CA, USA: Springer, 2002, pp. 13–28.

[15] L. Lerman, R. Poussier, G. Bontempi, O. Markowitch, and F.-X. Standaert, "Template attacks vs. machine learning revisited (and the curse of dimensionality in side-channel analysis)," in *Constructive Side-Channel Analysis and Secure Design* (Lecture Notes in Computer Science), vol. 9064, S. Mangard and A. Y. Poschmann, Eds. Berlin, Germany: Springer, 2015, pp. 20–33.

[16] O. Reparaz, B. Gierlichs, and I. Verbauwhede, "Selecting time samples for multivariate DPA attacks," in *Proc. 14th Int. Workshop Cryptograph. Hardw. Embedded Syst. (CHES)*, Leuven, Belgium, E. Prouff and P. Schaumont, Eds. Berlin, Germany: Springer, 2012, pp. 155–174.

[17] S. Picek *et al.*, "Side-channel analysis and machine learning: A practical perspective," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, May 2017, pp. 4095–4102.

[18] L. Bohy, M. Neve, D. Samyde, and J.-J. Quisquater, "Principal and independent component analysis for crypto-systems with hardware unmasked units," in *Proc. e-Smart*, Cannes, France, Jan. 2003, pp. 1–9.

[19] TELECOM ParisTech SEN Research Group. (2014). *DPA Contest (4^{th} Edition)*. [Online]. Available: http://www.DPAcontest.org/v4/

[20] J.-S. Coron and I. Kizhvatov, "An efficient method for random delay generation in embedded software," in *Proc. 11th Int. Workshop Cryptograph. Hardw. Embedded Syst. (CHES)*, in Lecture Notes in Computer Science, vol. 5747, C. Clavier and K. Gaj, Eds. Lausanne, Switzerland: Springer, 2009, pp. 156–170.

[21] M. Fernáandez-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, pp. 3133–3181, Oct. 2014.

[22] O. Choudary and M. G. Kuhn, "Efficient template attacks," in *Proc. 12th Int. Conf. Smart Card Res. Adv. Appl. (CARDIS)*, in Lecture Notes in Computer Science, vol. 8419, A. Francillon and P. Rohatgi, Eds. Lausanne, Switzerland: Springer, 2013, pp. 253–270.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[25] F.-X. Standaert, T. G. Malkin, and M. Yung, "A unified framework for the analysis of side-channel key recovery attacks," in *Proc. EUROCRYPT*, in Lecture Notes in Computer Science, vol. 5479. Cologne, Germany: Springer, 2009, pp. 443–461.

[26] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *Proc. 38th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2015, pp. 1200–1205.

[27] J. Suto, S. Oniga, and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," in *Proc. 6th Int. Conf. Comput. Commun. Control (ICCCC)*, May 2016, pp. 124–129.

[28] G. James, D. Witten, T. Hastie, and R. Tibsihrani, *An Introduction to Statistical Learning* (Springer Texts in Statistics). New York, NY, USA: Springer-Verlag, 2013.

[29] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf., Rutgers Univ.*. New Brunswick, NJ, USA, 1994, pp. 121–129.

[30] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B*, vol. 58, no. 1, pp. 267–288, 1994.

[32] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, Jul. 2004, pp. 78–85.

[33] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[34] N. Meinshausen and P. Bühlmann, "Stability selection," *J. Roy. Stat. Soc., B*, vol. 72, pp. 417–473, Sep. 2010.

[35] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.

[36] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.

[37] D. B. Roy, S. Bhasin, S. Guilley, A. Heuser, S. Patranabis, and D. Mukhopadhyay, "CC meets FIPS: A hybrid test methodology for first order side channel analysis," *IEEE Trans. Comput.*, vol. 68, no. 3, pp. 347–361, Mar. 2019.

Stjepan Picek (SM'19) received the Ph.D. degrees in computer science from Radboud University, Nijmegen, The Netherlands, and University of Zagreb, Zagreb, Croatia, in 2015.

He was with the COSIC Group, KU Leuven, Leuven, Belgium. He was a Postdoctoral Researcher with the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA. He is currently an Assistant Professor with the Cyber Security Research Group, TU Delft, Delft, The Netherlands. His current research interests include cryptography, machine learning, and evolutionary computation.

Annelie Heuser received the Ph.D. degree from TELECOM-ParisTech, Paris, France, in 2016.

She was an Associated Researcher with the Center for Advanced Security Research Darmstadt (CASED), Darmstadt, Germany. She was a Postdoctoral Researcher with TELECOM-ParisTech. She is currently a Researcher with the French National Center for Scientific Research (CNRS), IRISA, Rennes, France. Her current research interests include side-channel analysis, machine learning, hardware security, and malware detection/classification.

Alan Jovic (M'08) received the B.Sc. and Ph.D. degrees in computer science from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Zagreb, Croatia, in 2006 and 2012, respectively.

From 2006 to 2007, he was an Expert Associate with the Rudjer Boskovic Institute, Zagreb, Croatia. Since 2007, he has been with FER, University of Zagreb, where he is currently an Assistant Professor of computer science. He has authored or coauthored more than 50 refereed articles in international publications. His current research interests include machine learning with applications, biomedical engineering, and software engineering.

Dr. Jovic is a member of EMBS.

Lejla Batina (SM'18) received the Ph.D. degree in cryptography from Katholieke Universiteit Leuven, Leuven, Belgium, in 2005.

She has also studied with the Eindhoven University of Technology, Eindhoven, The Netherlands. From 2001 to 2003, she was a Cryptographer with SafeNet B.V., Rotterdam-Pernis, The Netherlands. She is currently a Full Professor with Radboud University, Nijmegen, The Netherlands. She has authored or coauthored more than 100 refereed articles. Her current research interests include implementations of cryptography and hardware security.

Dr. Batina is an Editorial Board Member of the IEEE TIFS. She served as a Program Co-Chair for CHES 2014.