Classifying the Circuit De-obfuscation Runtime based on Learning Methods and K-means Clustering

Securesemi

Spring 2022

Contents

[1. Introduction 1](#_Toc112399876)

[2. Materials and Methods 4](#_Toc112399877)

[2.1. Dataset 4](#_Toc112399878)

[2.2. Proposed Framework 4](#_Toc112399879)

[2.2.1. IC labeling 5](#_Toc112399880)

[2.2.2. Feature extraction and selection 7](#_Toc112399881)

[2.2.3. Training, validating, and testing CNF-NET model 11](#_Toc112399882)

[2.3. Implementation details of CNF-NET model 14](#_Toc112399883)

[3. Results 15](#_Toc112399884)

[3.1. Evaluation and assessment 15](#_Toc112399885)

[3.2. Results of applying whole ICs to the proposed framework 16](#_Toc112399886)

[3.2.1. SFFS results 16](#_Toc112399887)

[3.2.2. 9-fold cross-validation results on the whole ICs 16](#_Toc112399888)

[3.3. Results of applying each IC to the proposed framework 17](#_Toc112399889)

[4. Conclusion and Discussion 20](#_Toc112399890)

[4.1. Discussion 20](#_Toc112399891)

[4.2. Conclusion 20](#_Toc112399892)

[References 21](#_Toc112399893)

# Introduction

Semiconductor fabrication is one of the most important steps in designing integrated circuits (ICs). However, the capital and operational costs of this step is considerably high. Due to it, many semiconductor companies outsource their designed ICs to off-shore foundries. Although this trend has different benefits such as low manufacturing costs and reduced design efforts, it could pose various risks and threads to the security and trustworthiness of ICs. The most considerable risks of this trend are IC counterfeiting and unauthorized overproduction, trojan insertion, IP stealing, and circuit piracy [1-4]. It was estimated that these threats could cause $169 billion financial loss per year. Unfortunately, these threats and risks could be posed during any stage of production and design of an IC. Also, it may be remain prevalent after the design release into the industry and market. The main risk and threat that arises from attackers is the invasive reverse engineering (RE) by fully recognizing its functionality stage by stage and extracting the unveiling gate-level netlist. To protect ICs and hardware systems from these threads, **IC obfuscation** techniques have been extensively proposed in recent years [5]. The general approach for IC obfuscation is to introduce the ambiguity in the functionality of the IC based on the obfuscation of some gates. In this scheme, the obfuscated gates cannot be determined by reverse engineering optically, while they preserve the same functionality as the original gates. In other words, such approaches select a part of the IC as the obfuscated part, whose functionality can be retrieved only if correct keys are provided at the additional input gates.

These approaches were effective against conventional attacks [6]. The conventional attacks are based on the fact that there are limited types of gates (e.g., AND, OR, XOR, NOT, MNOR) in an IC, so the attackers can just brute force all the possible combinations of types for all obfuscated gates to find out the one that functions identically to the targeted IC to be deobfuscated. However, advanced attack techniques have been proposed today that obfuscation methods are not very resistant and effective against them [6]. One branch of these techniques that have attracted enormous attention is the Boolean satisfiability problem (SAT)-based attacks [7]. SAT attacks could deobfuscate ICs in milliseconds to several days, years and months. In other words, the runtime of IC deobfuscation by SAT attacks could vary from milliseconds to days or years. This running time highly depends on different factors such as the location of the obfuscated gates, types of obfuscated gates, number of obfuscated gates, circuit topology, and key size. Therefore, it is necessary to obtain a successful obfuscation scheme to resist ICs against advanced attacks. In detail, a successful obfuscation scheme should increase the running time of reverse engineering by advanced attacks, so that this time reaches days, months and even years. However, the implementation of obfuscation schemes will have significant costs in finance, space, time, and power. As a result, such trade-off requires us to search for optimal positions instead of purely increasing their quantity.

In order to maximize the runtime for deobfuscating while incurring minimal overheads, the best set of gates should be selected for obfuscation scheme. Until present, this process is based on human heuristics or experience, which is seriously arbitrary and sub-optimal [8]. In other words, it is unable to “try and error” all the various approaches of obfuscation because it could be millions of combinations to try and the runtime of each attack could be extremely time-consuming. Moreover, the runtime of unknown attacks may be dramatically long and simulating such attacks could be impossible due to the limited resources.

To address the aforementioned issues, we have focused on implementing an approach for automatic recognizing vulnerability or resilience of the ICs against SAT attacks. This research topic is highly under-explored in the community due to its significant challenges listed here:

* **Difficulty in characterizing the algorithmic mechanism of attackers which are mostly sophisticated and hide.** Until now, several deobfuscation techniques have been introduced, which most of them include sophisticated theories, rules, and heuristics [8]. Consequently, conventional learning methods (e.g., support vector machine (SVM), linear regression (LR) [8]) could not characterize the behavior of such highly non-linear and strongly coupled systems.
* **Difficulty in extracting determinant features from discrete and dynamic graph-structured ICs.** The main input of the IC vulnerability or resilience recognizer is the IC topology with selected gates obfuscated. The structure of these data is highly variable. Hence, formulating and integrating this type of data as mathematical forms that could be input to conventional computational and machine learning models is highly challenging. As a result, traditional feature extraction approaches are not intuitive to be applied to such type of data without significant information loss.
* **Requirement for an automatic framework to recognize the resistance or vulnerability of ICs.** As mentioned, it is essential to obtain a successful obfuscation scheme to resist ICs against advanced attacks. To this end, we should select the best scheme based on the simulations of runtimes of the attackers. However, it could be extremely time-consuming and simulating of some attacks could not be possible until present.

In this work, we proposed an automatic framework based on deep learning methods to recognize the resistance or vulnerability of ICs. Recently, deep learning approaches have attained immense success in various classification and regression tasks such as image segmentation, object recognition, machine translation, and data generation, which motivates the generalization of them into graph-structured data [9, 10]. To this end, different architectures of graph neural networks (GNNs) have been proposed to solve classification and regression problems of the graph-structured data [10]. In the problem of the classification of the ICs based on their resistance to advanced attacks, we can formulate ICs as multi-attributed graphs and use GCN models as the predictive models. However, formulating ICs as multi-attributed graphs for the input of GCN models is a challenging task because of varying the ICs structures, which causes the inputs of GCN have different structures. To address the mentioned issues, we used the Conjunctive Normal Form (CNF) method to represent each IC into the graph model. After that, some scalar and vector features were extracted from the CNF representation of each circuit. Next, we used sequneial forward feature selection (SFFS) to select the best scalar features. Then, we used a dual-input GNN model to implement an automatic framework for recognizing the resistance or vulnerability of ICs. Also, we utilized k-means clustering and elbow technique to determine the class of each IC according to the runtime of the attacks. The major contributions of this study are listed as follows:

* Automatically feature extraction of the characteristics and the structure of the IC in the form of CNF representation.
* Selecting the best scalar extracted features from the CNF representation using SFFS algorithm.
* Formulate a dual-input GNN model that could process varying sizes of graph data by taking the best scalar features and vector features.
* Conducting comprehensive experimental assessments and analyses on real-world datasets.

# Materials and Methods

## Dataset

In this work, the datasets are obtained by running SAT algorithm on real-world ISCAS-85 benchmark (c1355, c2670, c3540, c5315, c6288, c7552). It is worth mentioning that each IC was stored by bench files as per the SAT-solver requirements. The obfuscated ICs, along with the original benchmarks, are given to the SAT-solver as input parameters, and SAT-solver tries to find the key. The time taken by the SAT-solver to find the correct key is the actual de-obfuscation time. The samples generated for implementing the proposed framework contains the obfuscation samples generated using varied obfuscation percentage and replacement policies. By this process, we collected the IC samples as bench files and the runtime of SAT-solver on each circuit.

## Proposed Framework

Figure 1 shows the main steps of the proposed framework for automatic recognition of IC resistance against SAT attacks. As shown in Figure 1, we first assigned labels to each IC based on the runtime of SAT attacks. To this end, elbow technique was employed to determine the best number of clusters of the runtimes of SAT attacks on the ICs. After that, k-means clustering with the best number of clusters applied to the runtime of SAT attacks on the ICs. Then, the labels were assigned to the ICs based on the clustering results and boundary between clusters. **In this study, the best number of clusters was equal 2 and the ICs were categorized into the SAT-resilient and SAT-vulnerable.** In the second step, the CNF representation was obtained from the bench file format of each IC. Next, the scalar and vector features were extracted from the CNF representation of each IC. Furthermore, the best scalar features were selected in this step by SFFS method. In the third step, a dual-input GNN model called CNF-NET was trained, validated, and tested, respectively by the scalar and vector features and the labels of the ICs. Finally, the trained and validated CNF-NET model was deployed to classify each IC into the SAT-resilient or SAT-vulnerable classes. Following, each step was described in detail.

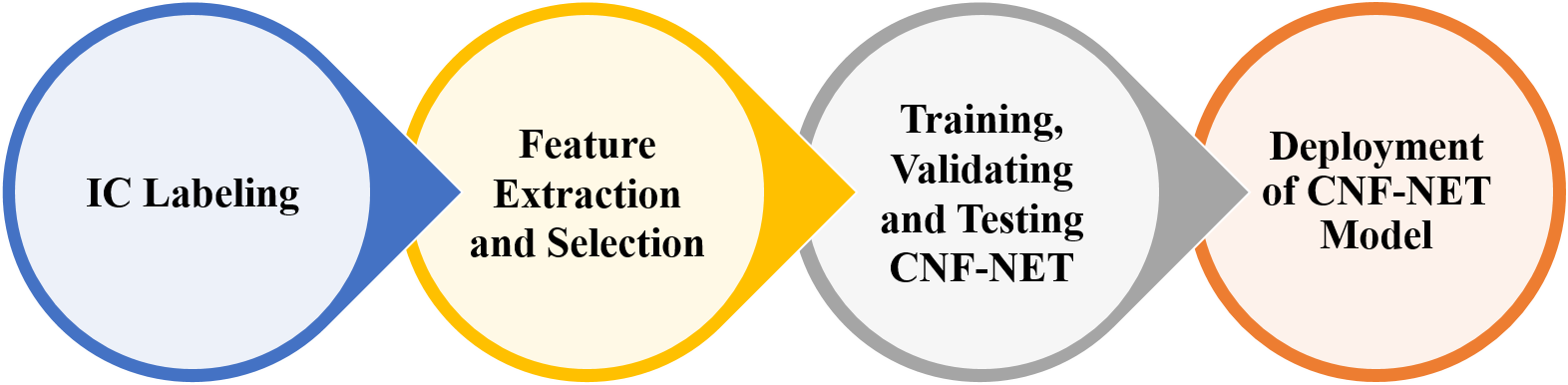


Figure 1: The main steps for implementing the presented framework for automatic recognition of IC resistance against SAT attacks

### IC labeling

For developing a supervised classification framework, some labels should be defined for each sample. In this study, the runtimes of SAT attacks on the ICs were available and we should perform a procedure to map these runtime into the class labels. To this end, an unsupervised process with the best number of clusters should be used to cluster each runtime into the separated clusters. In order to determine the best number of clusters, the elbow technique method was employed to determine the best number of clusters. This method selects the optimal number of clusters by fitting an unsupervised model with a range of values for number of clusters. In this work, the minimum and maximum number of clusters were chosen to 2 and 10, respectively. Also, the unsupervised model used in the elbow technique in this study was K-means clustering. It is worth mentioning that all SAT runtimes for all ICs were collected and applied to the elbow technique. Figure 2 shows the elbow technique results for determining the best number of clusters for the SAT runtimes based silhouette score. As illustrated in Figure 2, the highest silhouette score was obtained in the two number of clusters. Figure 3 shows the K-means clustering result by *K=2* on the SAT runtimes on

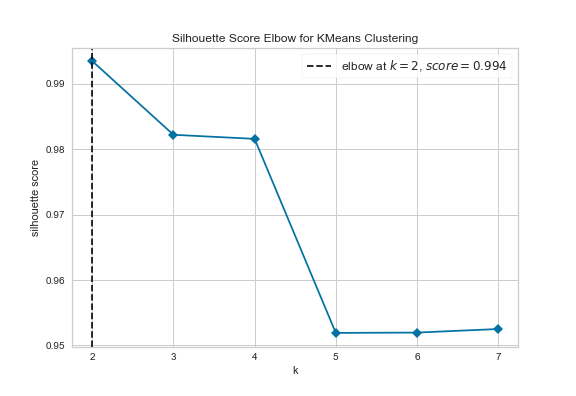


Figure 2: Elbow technique results for determining the best number of clusters for SAT runtimes on the ICs.

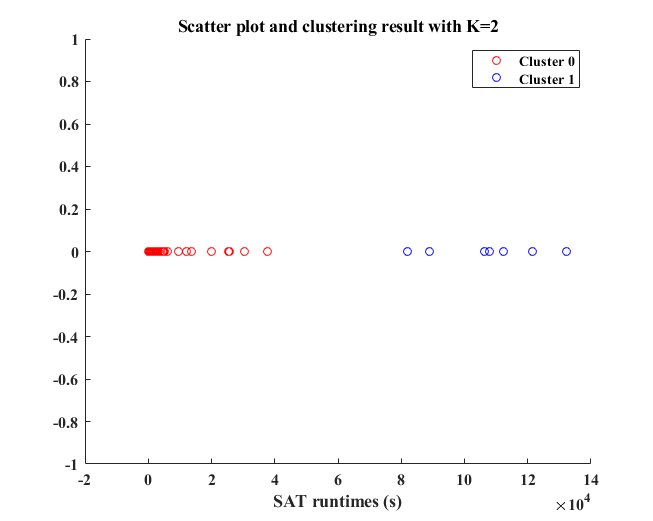


Figure 3: The K-means clustering result with K=2 on the SAT runtimes on the ICs.

the ICs. According to the K-means clustering results, the label of cluster 0 and cluster 1 were assigned to the SAT-vulnerable and SAT-resilient classes, respectively. Also, a mapping function was formulated based on the boundaries between cluster 0 and cluster 1, which is shown as follows:

|  |  |
| --- | --- |
|  | ‑ |

where *g(x)* is the mapping function, x is the SAT runtime on the IC in seconds, *y* is the label of the IC, and *m* is the threshold value between classes. In this function, ­the 0 and 1 values of *y* refer to the SAT-resilient and SAT-vulnerable cases, respectively. The value of *m* was obtained from the average of the highest value of cluster 0 and the lowest value of cluster 1. In our experiments, the value of ­*m* was obtained equal 59830 second.

### Feature extraction and selection

In this step, the CNF representation was obtained from bench file format of each IC. Next, some scalar and vector features were extracted from the CNF repression of each IC. After that, the best scalar features were selected using SFFS algorithm. In the following, the CNF representation and its derived features were explained in details. The blockdiagram of this step is shown in Figure 4.

#### CNF representation and its derived features

One of the main challenges in this study was the representing the obfuscated netlist in an intact and structured way for a machine leaning framework. Although CNF is typically as a sequence, it is mathematically not a sequence as the order among different clauses is meaningless. Moreover, one literal can appear in multiple clauses with or without their negation forms, which further complicates its representation. Also, there are very few existing machine learning methods that are designed for directly treating CNF representation as input without information loss. In order to solve this problem, the converted CNF of an obfuscated IC was modeled as an undirected and signed bipartite graph which uses one node type for clauses and the other for literals. Mathematically, the CNF bipartite graph could be defined as *G(E, Vliteral, Vclause)*, where *E, Vliteral,* and *Vclause* are the edges between nodes, the set of literal nodes, and the set of clause nodes, respectively. The value of *E* could be 0,+1, and -1. Consider a literal node as *l­­* and a clause node as ­*c*. If *l­* is in *c* and *l* is positive, the value of *E* between them is 1. If *l­* is in *c* and *l* is negative, the value of *E* between them is -1, and if there is no connection between *l* and *c*, the value of *E* is 0.

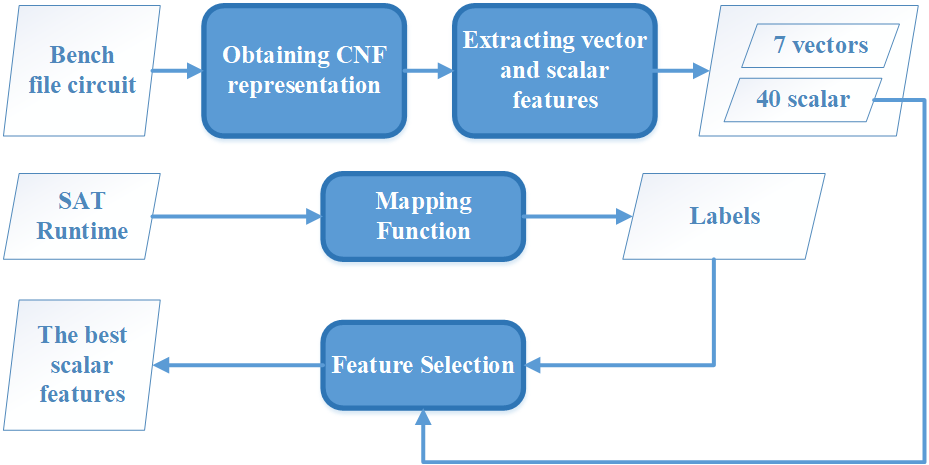


Figure 4: The blockdiagrm of feature extraction and selection of the proposed method

In other words, the adjacency matrix of CNF bipartite graph representation of an IC could be modeled as follows:

|  |  |
| --- | --- |
|  | ‑ |

where *IM* is the adjacency matrix, *M ­*is the number of clause nodes, and *N* is the number of literal nodes. Firstly, ­*IM* was obtained by the CNF transformation from each IC bench file. Next, all zero rows were eliminated from *IM* matrix. In the next step, literal and clause degrees were computed from ­­*IM* matrix using following equations:

|  |  |
| --- | --- |
|  | ‎2‑ |
|  | ‑ |

where |.|, *cd*, and *ld* are the absolute function, clause degree, and literal degree, respectively. Literal graph (LG) was also obtained from *IM* matrix, which could be modeled as *G1(E1,V1)*, where *V1*is the literal nodes and *E­1* is the edges of LG graph. In LG graph, *Ei,j* is defined as the count of clauses in which both *Vi ­*and *Vj* exist. After obtaining LG graph, the literal degree of it (*ldLG)* was also computed similar ‎2‑3 equation. Next, positive adjacency matrix (*pam*) and negative adjacency matrix (*nam*) were obtained based on the *IM* matrix. To obtain ­*pam*, all values of -1 of ­*IM* matrix was replaced with 0. For computing ­*nam*, all values of 1 and -1 of ­*IM* matrix was replaced with 0 and 1, respectively. Mathematically, *pam=[b­i,j]­M×N­* and *nam=[c­i,j]­M×N­* where each element of these matrixes are 0 or 1. Then, *pratio0*, *pratio1*, ­*nratio0*, and *nratio1* parameters were calculated as follows:

|  |  |
| --- | --- |
|  | ‑ |
|  | ‑ |
|  | ‑ |
|  | ‑ |

The vector features that are extracted from CNF representation and considered as the inputs of the machine learning model are *ld*, *cd, ldVG, pratio0*, *pratio1*, ­*nratio0*, and *nratio1*. After extracting vector features, the scalar features were obtained from the mentioned vector features and *IM* matrix. Three first scalar features are ­*M*, *N,* and *M/N*, which *M* and N denote the number of clause and literal nodes, respectively. Moreover, the statistical features of the mentioned vector features are extracted. These features are mean, variance, entropy, skewness, and kurtosis. Also, the ratios of binary and ternary clauses are calculated as another scalar features. Figure 4 illustrates the diagram of feature extraction of the CNF representation, which causes to obtain 40 scalar features and 7 vector features.

#### Feature Selection

The extracted features of a dataset in machine learning frameworks may contain redundant or irrelevant features. Moreover, redundant or irrelevant features could increase the computational load and lead to overfitting the model. The main purpose of feature selection is to find the best subset of features, which could reduce the dimension of feature space, and improve the performance of the machine learning framework. In this study, we applied the scalar features to SFFS algorithm to obtain the best subset of them. The steps of SFFS method in this work are summarized in Algorithm 1. The objective criterion (*J*) was defined as the mean of accuracies during 10-fold cross-validation. Based on the steps in Algorithm 1, the features are sequentially added to an empty candidate set until the addition of further features does not decrease ­*J* value. In this study, the classification model in the SFFS was k-nearest neighbor (KNN) model with K=2. In our experiments, SFFS returned 3 features form 40 features as the best subset of features. These features are ­*M*, mean of *cd*, and kurtosis of ­*pratio1*.

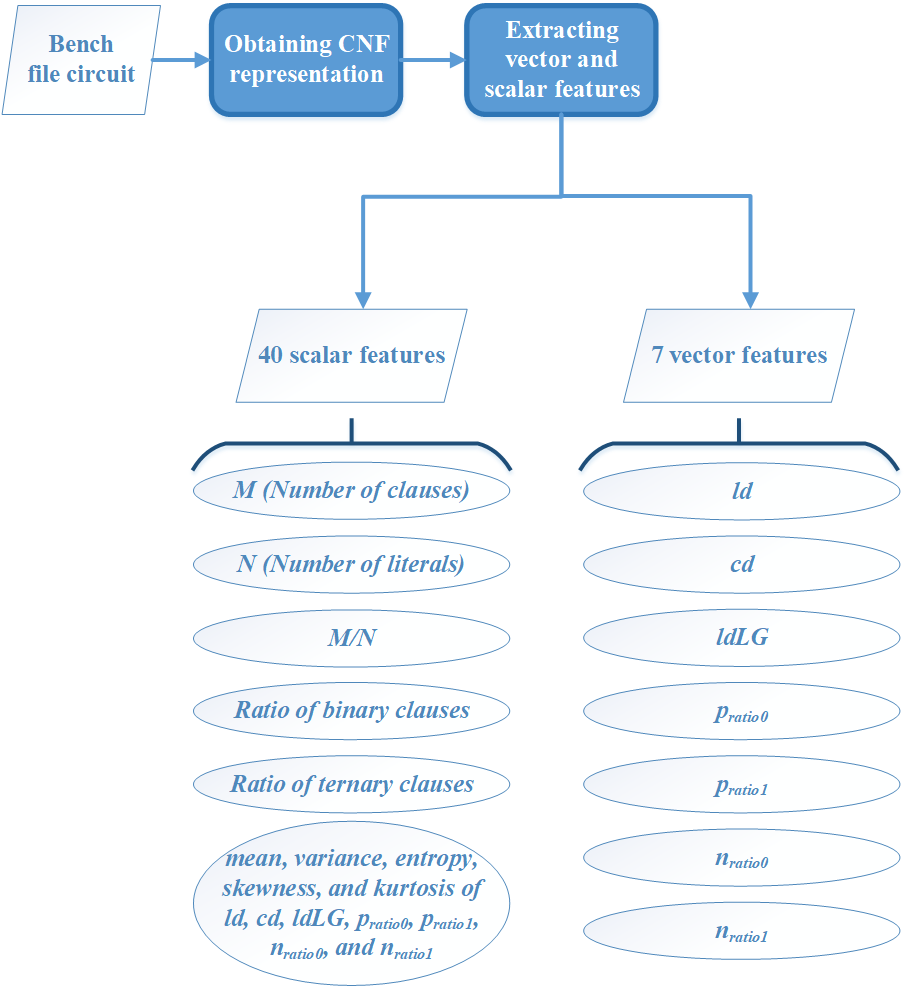


Figure 5: The diagram of feature extraction of the proposed framework

|  |
| --- |
| **Algorithm 1: The SFFS algorithm** |
| Input: The entire feature set, *Y = {y1, y2, ..., yd}*  Output: The selected feature subset, *S*k *= {Sj|j = 1, 2, ..., k; SjϵY }* where *k = 1, 2, ..., d*   1. Start with the empty set, *S0 = {}*, *k=0*. 2. *s∗ = argmax (J (Sk + s))*, where *s ϵ Y − Sk*. 3. Update *Sk+1 = Sk + s∗*; *k = k + 1*. 4. If *J(Sk) > J(Sk+1)*, go to step 6. 5. Else go to the step 2. 6. Stop |

### Training, validating, and testing CNF-NET model

After extracting scalar and vector features from CNF representation of each IC, the CNF-NET model was trained and validated to classify each IC into the SAT-resilient or SAT-vulnerable classes. The CNF-NET is a dual-input GNN model, which receives the vector, and scalar features as individual inputs. Figure 6 depicts the details of the structures of the CNF-NET model. As depicted in Figure 6, the main parts of the CNF-NET are incidence convolution, categorizing layer, and fully-connected (FC) layers. For each case, the vector features are applied to the incidence convolution part, and then the selected scalar features are categorized with incidence convolution’s output. Next, FC layers receives the categorization results and generates two probabilities for each case. The value of these probabilities indicates the probability that the input case of CNF-NET belongs to each of the SAT-vulnerable and SAT-resilient classes. Based on these probabilities, the input case is classified into the mentioned classes. In the following, each part of CNF-NET is explained in details.

#### Incidence convolution

The first part of CNF-NET, which receives the vector features as inputs, is incidence convolution. This part generates a vector with 7 elements as the output. In other words, it generates a scalar value for each input vector. In this part, consider *f­φ()* as a sequence of FC layers, which its input and output are scalar values. The details of the structure of *f­φ()* is given in Table 1. Also, suppose the core function of incidence convolution as *gα()* and each input vector as *Vi* which is applied to *gα()*. The main steps of *gα()* function is summarized in Algorithm 2. *gα()* generates a scaler value for each *Vi* *i=1, 2, …, 7­* and stores them in a vector. In the first step of *gα()*, ­*Vi* is normalized which results in *Vinorm*. Then, each element of ­*Vi* is applied to *f­φ()* and the outputs of it is stored in a vector

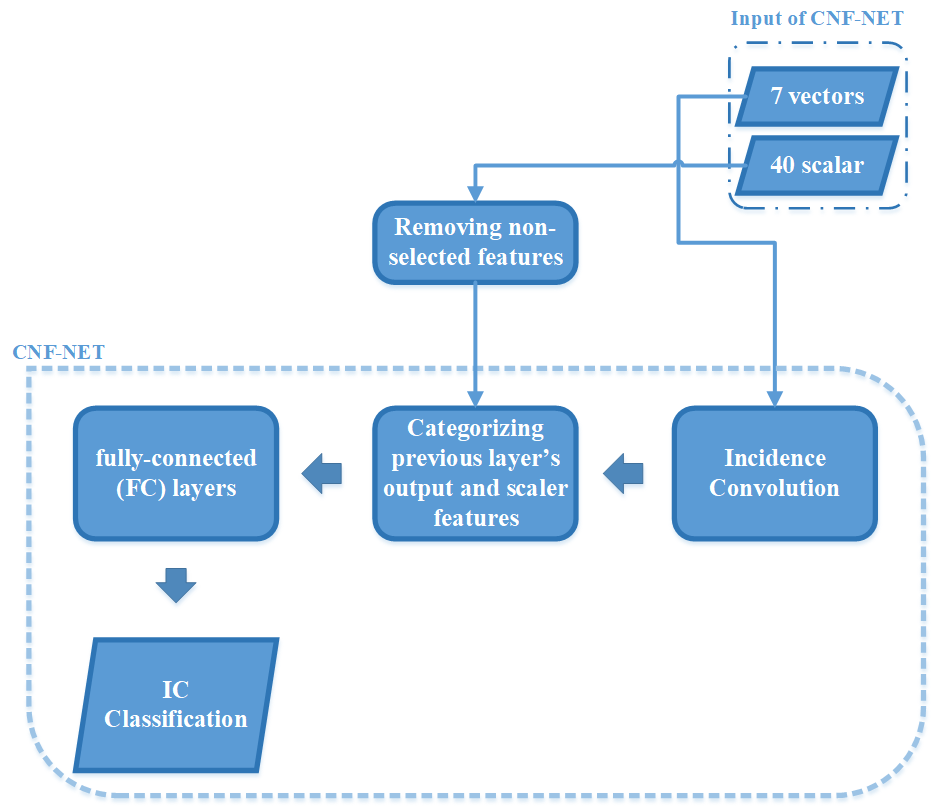


Figure 6: The diagram of the structures of CNF-NET model.

Table 1: The architecture and layers of fα() function

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer Number** | **Layer** | **Input** | **Output** |
| 1 | FC | 1 | 32 |
| 2 | FC | 32 | 32 |
| 3 | FC | 32 | 32 |
| 4 | FC | 32 | 1 |

called *Wi­*. After that, *Wi* is multiplied to *Vinorm* and then the summation of the output of this multiplication is computed. Finally, the summation result is stored in the output vector of incidence convolution.

|  |
| --- |
| **Algorithm 2: *gα() function*** |
| Input: Vector features, *V = {V1, V2, ..., V7}*  Output: A vector with 7 elements, *Vo={Vok}* where *k=1, 2, …, 7*   1. Initialization: *i=1, Wi={},* ­and *Vo={}* 2. Normalizing *Vi* using 3. Apply each element of *Vinorm* to *f­φ()* function and store the outputs in *Wi* 4. Computing and appending it to *Vo* 5. *i = i +1* 6. If *i≤7*, go to step 2 7. Else go to step 8 8. Stop |

#### Categorizing and FC parts

The next parts of CNF-NET model after incidence convolution are categorizing and FC layers parts, respectively. The categorizing layer receives the selected scalar features and the outputs of incidence convolution and categorizes them into a vector. Next, the categorization result is applied to FC part, which consists of multiple FC layers. Table 2 shows the architecture of the layers of FC part. The last layer of the FC part is a softmax layer, which generates the probabilities that indicate the input case of CNF-NET belongs to each of the SAT-vulnerable and SAT-resilient classes. Based on the values of these probabilities, the input case is classified into the mentioned classes.

Table 2: The architecture and layers of FC part

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer Number** | **Layer** | **Input** | **Output** |
| 1 | FC | 10 | 32 |
| 2 | FC | 32 | 32 |
| 3 | FC | 32 | 32 |
| 4 | FC | 32 | 2 |
| 5 | Softmax | 2 | 2 |

## Implementation details of CNF-NET model

The proposed framework for recognizing IC resistance against SAT attacks was implemented under Python 3.7 coding environment. The CNF representation of each IC was obtained using bench2cnf module. To extract features from CNF representation, we used numpy and scipy libraries. For implementing SFFS algorithm, we used mlxtend library in our study. For implementing k-means clustering and 10-fold cross-validation, scikit-learn framework was used. Also, elbow technique was employed in this work using Yellowbrick library. In order to implement the CNF-NET as the core of the proposed method, Pytorch framework was used. For training the model, the epochs, learning rate, batch size was set 50, 0.01, and 32 values, respectively. It is worth mentioning that that all simulations and implementations were conducted on a system with Intel® Xeon® Processor E5-2697 v2 CPU at 2 GHz and 8 GB memory.

# Results

In this section, the performance of the proposed framework for classifying each IC into the SAT-vulnerable and SAT-resilient classes is investigated. First, the evaluation of the proposed method and performance metrics are explained in details. After that, the obtained results of the assessment of the proposed framework are reported.

## Evaluation and assessment

To evaluate the proposed framework, 10-fold cross-validation technique was employed in this study. In 10-fold cross-validation technique, the dataset is initially divided into 10 folds, of which 9 folds are randomly used as a training set, and the remaining fold is employed as a testing set. This process is repeated 10 times until each fold is utilized as a testing set. During each iteration, the testing set is applied to the trained model, which results in 10 different evaluation metrics values. In this work, the 10% of training set is randomly categorized into the validation set in each iteration of 10-fold cross-validation. The main reason for using validation set during training phase is to prevent overfitting by early stopping the training phase. The performance metrics used in this study to investigate the classification performance of the proposed framework are accuracy (*AC*), sensitivity (*SE*), specificity (*SP*), and F1-score (*F1*), which are defined as follows:

|  |  |
| --- | --- |
|  | ‑ |
|  | ‑ |
|  | ‑ |
|  | ‑ |

where *TP* is the number of SAT-vulnerable cases that are correctly classified, *FN* is the number of SAT-vulnerable samples that are incorrectly classified as SAT-resilient samples, *FP* is the number of SAT-resilient samples that are incorrectly classified as SAT-vulnerable cases, and *TN* is the number of SAT-resilient cases that are correctly classified.

## Results of applying whole ICs to the proposed framework

### SFFS results

Figure 7 shows the SFFS result for selecting best subset of features where y-axis and x-axis are returned accuracy and number of features. In the experiments, SFFS returned a subset of features with 3 attributes as the selected features. By these three features, the returned accuracy by SFFS was 0.95, which it is the maximum accuracy in Figure 7. The selected features are ­*M*, mean of *cd*, and kurtosis of ­*pratio1*.

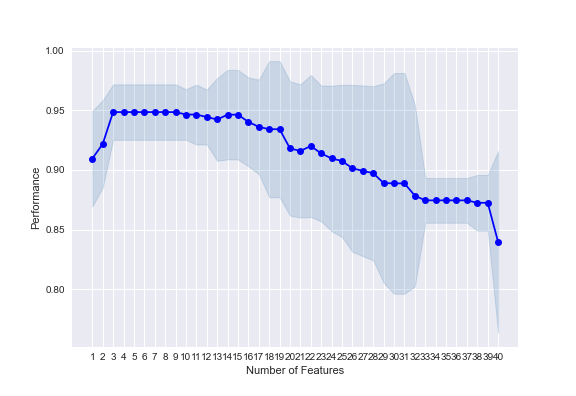


Figure 7: The returned accuracies by SFFS for each number of candidate features

### 9-fold cross-validation results on the whole ICs

Table 3 reports the obtained *AC*, *SE*, *SP­­*, and *F1* values by the proposed framework in each repletion of 10-fold cross-validation. The average of AC, SE, and F1 metrics are 88.54%, 88.32%, and 93.17%, respectively. As shown in Table 7, the distribution of data between positive and negative classes is so imbalance. This is due to lower number of negative cases against positive samples.

Table 3: The obtained values of the percentage (%) of AC, SE, SP, and F1 metrics of the proposed method in each iteration of 9-fold cross-validation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Iteration Number** | **Number of positive cases** | **Number of negative cases** | **AC (%)** | **SE (%)** | **SP (%)** | **F1 (%)** |
| 1 | 32 | 0 | 100 | 100 | nan | 100 |
| 2 | 32 | 0 | 100 | 100 | nan | 100 |
| 3 | 30 | 2 | 93.75 | 93.75 | nan | 96.77 |
| 4 | 23 | 9 | 81.25 | 79.31 | 100 | 88.46 |
| 5 | 30 | 2 | 93.75 | 93.75 | nan | 96.77 |
| 6 | 19 | 13 | 59.37 | 59.37 | nan | 74.50 |
| 7 | 31 | 1 | 96.87 | 96.87 | nan | 98.41 |
| 8 | 32 | 0 | 100 | 100 | nan | 100 |
| 9 | 23 | 9 | 71.87 | 71.87 | nan | 83.63 |

## Results of applying each IC to the proposed framework

In this step, the proposed framework was evaluated by each IC category individually using 10-fold cross-validation technique. In detail, it was assessed by C2670, C3540, and C6288 ICs, which were obfuscated using different approaches and keys. Tables 4, 5, and 6 report the obtained results of the proposed framework in 10-fold cross-validation for C2670, C3540, and C6288 ICs, respectively. For C2670 IC set, the average of AC, SE, and F1 metrics obtained by the proposed method were 96.88%, 98%, and 95.55%, respectively. The average of AC, SE, and F1 metrics obtained by the proposed method for C3540 IC set were 87.98%, 92%, and 86.72%, respectively. For C6288 IC set, the proposed method provided an average AC of 64.66%, SE of 72.44%, and F1 of 71.20%.

Table 4: The obtained values of the percentage (%) of AC, SE, SP, and F1 metrics of the proposed method in each iteration of 10-fold cross-validation for C2670 IC set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Iteration Number** | **Number of positive cases** | **Number of negative cases** | **AC (%)** | **SE (%)** | **SP (%)** | **F1 (%)** |
| 1 | 8 | 2 | 80 | 80 | nan | 88.88 |
| 2 | 10 | 0 | 100 | 100 | nan | 100 |
| 3 | 10 | 0 | 100 | 100 | nan | 100 |
| 4 | 9 | 0 | 100 | 100 | nan | 100 |
| 5 | 1 | 8 | 100 | 100 | 100 | 100 |
| 6 | 2 | 7 | 100 | 100 | 100 | 100 |
| 7 | 9 | 0 | 100 | 100 | nan | 100 |
| 8 | 9 | 0 | 100 | 100 | nan | 100 |
| 9 | 2 | 7 | 88.88 | 100 | 87.5 | 66.66 |
| 10 | 8 | 1 | 100 | 100 | 100 | 100 |

Table 5: The obtained values of the percentage (%) of AC, SE, SP, and F1 metrics of the proposed method in each iteration of 10-fold cross-validation for 3540 IC set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Iteration Number** | **Number of positive cases** | **Number of negative cases** | **AC (%)** | **SE (%)** | **SP (%)** | **F1 (%)** |
| 1 | 8 | 2 | 80 | 80 | nan | 88.88 |
| 2 | 10 | 0 | 100 | 100 | nan | 100 |
| 3 | 10 | 0 | 100 | 100 | nan | 100 |
| 4 | 9 | 0 | 100 | 100 | nan | 100 |
| 5 | 1 | 8 | 100 | 100 | 100 | 100 |
| 6 | 2 | 7 | 66.66 | 40 | 100 | 57.14 |
| 7 | 9 | 0 | 100 | 100 | nan | 100 |
| 8 | 9 | 0 | 100 | 100 | nan | 100 |
| 9 | 2 | 7 | 88.88 | 100 | 87.5 | 66.66 |
| 10 | 8 | 1 | 44.44 | 100 | 16.66 | 54.54 |

Table 6: The obtained values of the percentage (%) of AC, SE, SP, and F1 metrics of the proposed method in each iteration of 10-fold cross-validation for 6288 IC set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Iteration Number** | **Number of positive cases** | **Number of negative cases** | **AC (%)** | **SE (%)** | **SP (%)** | **F1 (%)** |
| 1 | 8 | 2 | 80 | 80 | nan | 88.88 |
| 2 | 10 | 0 | 100 | 100 | nan | 100 |
| 3 | 10 | 0 | 100 | 100 | nan | 100 |
| 4 | 9 | 0 | 100 | 100 | nan | 100 |
| 5 | 1 | 8 | 11.11 | 11.11 | nan | 20 |
| 6 | 2 | 7 | 22.22 | 22.22 | nan | 36.36 |
| 7 | 9 | 0 | 22.22 | 100 | 0 | 36.36 |
| 8 | 9 | 0 | 100 | 100 | nan | 100 |
| 9 | 2 | 7 | 22.22 | 22.22 | nan | 36.36 |
| 10 | 8 | 1 | 88.88 | 88.88 | nan | 94.11 |

# Conclusion and Discussion

## Discussion

As mentioned, the main purpose of this study is to implement an automatic machine-learning framework for classifying ICs to SAT-vulnerable and SAT-resilient classes. In other words, the proposed framework should recognize the IC resistance against SAT attacks. In terms of classification performance, the proposed method performed well in classifying SAT-vulnerable cases. However, it did not provide an acceptable performance in classifying SAT-resilient cases. This could be due to the imbalance distribution between SAT-vulnerable and SAT-resilient cases. In terms of time performance, the proposed framework could classify each IC in 236 seconds. This time is much lower than applying directly SAT attacks to ICs. Therefore, it could be used for automatic recognition of resistance of ICs against SAT attacks. As future works, such framework could be implemented to find the appropriate obfuscation method and key.

## Conclusion

In this work, an automatic framework for recognizing the IC resistance against SAT attacks was introduced. In the first step of this framework, the SAT runtimes of each IC was mapped to SAT-vulnerable and SAT-resilient classes to prepare the labels for training the CNF-NET model. It was performed by k-means clustering and elbow technique. Next, the CNF representation was extracted from IC bench file. After that, 7 vector features and 40 scalar features were computed from the CNF representation and then, the scalar features were applied to SFFS algorithm to select the best scalar features. Next, the CNF-NET model was trained, validated, and tested, respectively by the selected scalar and vector features and the labels of the ICs to classify each IC to the SAT-vulnerable and SAT-resilient classes. The model was validated using 9-fold cross-validation technique, which provided an average AC of 88.54%, SE of 88.32%, and F1 of 93.17%. Based on the obtained results, the proposed method had an acceptable performance in in recognizing SAT-vulnerable cases. However, it did not perform well in classifying SAT-resilient cases.

# References

[1] P. Subramanyan, S. Ray, and S. Malik, "Evaluating the security of logic encryption algorithms," in *2015 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2015, pp. 137-143: IEEE.

[2] K. Rosenfeld, "Trustworthy Hardware: Identifying and Classifying Hardware Trojans," 2010.

[3] U. Guin, D. Forte, and M. Tehranipoor, "Anti-counterfeit techniques: From design to resign," in *2013 14th International workshop on microprocessor test and verification*, 2013, pp. 89-94: IEEE.

[4] M. Rostami, F. Koushanfar, and R. Karri, "A primer on hardware security: Models, methods, and metrics," *Proceedings of the IEEE,* vol. 102, no. 8, pp. 1283-1295, 2014.

[5] M. Yasin, A. Sengupta, M. T. Nabeel, M. Ashraf, J. Rajendran, and O. Sinanoglu, "Provably-secure logic locking: From theory to practice," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1601-1618.

[6] M. El Massad, S. Garg, and M. V. Tripunitara, "Integrated Circuit (IC) Decamouflaging: Reverse Engineering Camouflaged ICs within Minutes," in *NDSS*, 2015, pp. 1-14.

[7] D. Liu, C. Yu, X. Zhang, and D. Holcomb, "Oracle-guided incremental SAT solving to reverse engineer camouflaged logic circuits," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2016, pp. 433-438: IEEE.

[8] S. Khaleghi and W. Rao, "Hardware obfuscation using strong pufs," in *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2018, pp. 321-326: IEEE.

[9] S. Wang, G. Ananthanarayanan, Y. Zeng, N. Goel, A. Pathania, and T. Mitra, "High-throughput CNN inference on embedded ARM Big. LITTLE multicore processors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 39, no. 10, pp. 2254-2267, 2019.

[10] M. Lechner, A. Jantsch, and S. M. P. Dinakarrao, "ResCoNN: Resource-efficient FPGA-accelerated CNN for traffic sign classification," in *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, 2019, pp. 1-6: IEEE.