Fine-Grained Static Detection of Obfuscation Transforms Using Ensemble-Learning and Semantic Reasoning

Ramtine Tofighi-Shirazi, Irina Mariuca Asavoae, Philippe Elbaz-Vincent

To cite this version:

HAL Id: hal-02355528
https://hal.archives-ouvertes.fr/hal-02355528
Submitted on 12 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Fine-Grained Static Detection of Obfuscation Transforms Using Ensemble-Learning and Semantic Reasoning

Ramtine Tofghi-Shirazi  
Univ. Grenoble Alpes, CNRS, Institut Fourier  
Trusted Labs, Thales Group  
Meudon, France  
ramtine.tofghishirazi@thalesgroup.com

Irina Măriuca Asăvoae  
Trusted Labs, Thales Group  
Meudon, France  
mariuca.asavoaie@thalesgroup.com

Philippe Elbaz-Vincent  
Univ. Grenoble Alpes, CNRS, Institut Fourier  
F-38000 Grenoble, France  
philippe.elbaz-vincent@univ-grenoble-alpes.fr

ABSTRACT
The ability to efficiently detect the software protections used is at a prime to facilitate the selection and application of adequate deobfuscation techniques. We present a novel approach that combines semantic reasoning techniques with ensemble learning classification for the purpose of providing a static detection framework for obfuscation transformations. By contrast to existing work, we provide a methodology that can detect multiple layers of obfuscation, without depending on knowledge of the underlying functionality of the training-set used. We also extend our work to detect constructions of obfuscation transformations, thus providing a fine-grained methodology. To that end, we provide several studies for the best practices of the use of machine learning techniques for a scalable and efficient model. According to our experimental results and evaluations on obfuscators such as Tigrress and OLLVM, our models have up to 91% accuracy on state-of-the-art obfuscation transformations. Our overall accuracies for their constructions are up to 100%.

KEYWORDS
machine learning, ensemble learning, deobfuscation, obfuscation, reverse engineering, symbolic execution

1 INTRODUCTION
Code obfuscation is a widely used software protection technique to mitigate the risks of reverse-engineering. It aims at protecting intellectual property by hiding the logic and data of a code. The use of code obfuscation transformations depends on the sensitivity of the application. Its applications are mainly digital right management, software licensing code or white-box cryptography, among others. Malicious codes also use extensively code obfuscation to hide their intent, evade detection and hinder analyses.

In order to properly evaluate obfuscation transformations, or to efficiently analyze malwares, many deobfuscation techniques have emerged. Their goal is to remove the protection layers applied on the code. The deobfuscation process can be seen as different strategies such as reverting, simplifying, or gathering information about the obfuscated code. In this paper we mainly focus on information gathering, particularly the static detection of obfuscation transformations. We also study an extension to the transformations constructions, namely the different methods employed for a specific obfuscation transformation to be achieved (e.g. dispatch-methods for control-flow flattening or code virtualization). This approach is previously known as metadata recovery attacks [53].

State-of-the-art deobfuscation techniques are often specific to obfuscation transformations. For example, the work of Udupa et al. [64] targets control-flow transformations, whereas others [6, 42, 47, 62] aim at removing opaque predicates. Generic deobfuscation techniques, however, make no assumption about the applied protections [54, 71]. These techniques are based on dynamic symbolic execution and may lack in code coverage and scalability.

Though obfuscation transformations are semantic-preserving, they may introduce side effects to the code [14]. Each transformation has its own construction methodology, thus specific patterns. Recent works try to tackle the detection of software protections using machine learning or deep learning techniques. Ugarte-Pedrero et al. [65] propose a semi-supervised learning approach in order to classify packed and unpacked binaries. Sun et al. [59], and more recently Biondi et al. [7], aim at detecting and identifying packers using machine learning techniques. Tofghi-Shirazi et al. [61] propose a deobfuscation methodology for invariant opaque predicates based on machine learning techniques.

From the variety of obfuscation techniques, as well as deobfuscation methodologies, the ability to efficiently detect the software protections used is at a prime. To that end, the recent work of Salem et al. [53] focuses on the detection of obfuscation transformations. Their goal is to facilitate the selection and application of adequate deobfuscation techniques. To the best of our knowledge, their work is the first to tackle code obfuscation detection using machine learning. However, their methodology is also prone to some limitations as explained next.

Current limitations. Existing detection technique for code obfuscation [53] based on machine learning techniques comes with the following limitations:

1. **Code dependency**: machine learning and syntax-reasoning used for the detection of obfuscation transformations can lead to code dependency. Namely, the trained model becomes dependent to the analyzed code used in the training set, thus lowering its accuracy.

2. **Multi-class problem**: the methodology used relies on multi-class problems for classification. Namely, they consider that one binary cannot be obfuscated with more than one obfuscation transformation. However, transformations can be combined, thus the necessity to be able to detect the several applied layers.

3. **Granularity**: the detection technique has a high-level of granularity. They may detect an obfuscation transformation, but they do not focus on their constructions. The latter is of
importance in order to decide which analysis to apply on obfuscated code. Many transformations constructions are made to prevent existing deobfuscation techniques.

Figure 1: Control-flow graph of a quick-sort function obfuscated using several Tigress transformations.

**Motivation.** When applying obfuscation transformations for software protections, stealth is sometimes not desired. Many applications aim for dissuasion in order to prevent reverse-engineering. In any case, the goal of our methodology is to provide a static and automated framework to help reverse-engineers. By detecting obfuscation transformations, and more specifically their constructions, an analyst will gain an important amount of time. The selection of the deobfuscation process to apply requires such knowledge beforehand. A motivating example is illustrated in Figure 1. It represents the obfuscated control-flow graph of a quick-sort function. Based on the previously introduced problems, our goal is to answer the following questions:

- **Complexity:** can we detect all applied layers of obfuscation transformation?
- **Granularity:** can we detect the constructions of applied obfuscation transformations?
- **Efficiency:** can we create accurate and generic enough models for unknown data?

As previously discussed in [53], metadata recovery attacks are usually manual tasks, therefore a potential bottleneck in the reverse engineering process. Our methodology, which could be plugged-in a disassembler framework, provides all applied transformation and construction and allows reverse-engineers to setup automated deobfuscation strategies. As an example, several opaque predicates constructions prevent SMT-solver based deobfuscation techniques [70]. Other recent works prevent the application of dynamic symbolic execution techniques [5, 44]. Thus, knowing which transformations and constructions analysts are facing may prevent using unadapted techniques for the deobfuscation process.

**Contributions.** In order to face the above limitations and answer our motivating questions, we bring the following contributions:

1. A novel methodology that combines semantic reasoning with ensemble learning techniques applied for a multi-label and multi-output ensemble model. We believe that semantic reasoning will prevent our model from code dependency limitations, and provides us with the ability to detect several combined layers of obfuscation transformations.

2. An extension of our methodology for a fine-grained detection. Based on our main approach, a second classification model is used for the detection of the transformations constructions, based on a multi-class classification model (i.e. one unique label per instances).

3. Several studies and experiments that justify the constructions of our methodology. We compare different machine learning approaches and techniques in order to build efficient and scalable models. We also evaluate our methodology against state-of-the-art obfuscators such as Tigress [12] and Obfuscator-LLVM [29] (i.e. OLLVM).

Our paper is organized as follows. Section 2 presents the background information about code obfuscation and targeted transformations. We also introduce related work, as well as notions of supervised machine learning. Section 3 describes our methodology which combines semantic reasoning with ensemble learning. Section 4 contains our studies and experiments towards an efficient implementation of our methodology. Section 5 illustrates our evaluations on state-of-the-art and publicly available obfuscators. Section 6 briefly discuss the application of our methodology to setup deobfuscation strategies. Then, we discuss our design limitations in Section 7, as well as our perspectives in Section 8. Finally, Section 9 presents our conclusions.

**Limitations.** While our results illustrate the interest of the methodology, evaluating the exact gain of the different components of the approach and experimental comparison to related contributions are left as future work.

2 BACKGROUND

We briefly present code obfuscation and some of the employed transformations. Then we introduce several notions related to supervised machine learning and metadata recovery attacks introduced in [53].

2.1 Code obfuscation

Collberg et al. [14] define code obfuscation as follows:

Let $P \xrightarrow{T} P'$ be a transformation $T$ of a source program $P$ into a target program $P'$. We call $P \xrightarrow{T} P'$ an obfuscating transformation if $P$ and $P'$ have the same observable behavior, $P'$ is harder to analyze than $P$, and $P'$ is no more than polynomially slower than $P$. Consequently, the following conditions must be fulfilled for an obfuscating transformation: if $P$ fails to terminate, or terminates with an error condition, then $P'$ may or may not terminate; otherwise, $P'$ must terminate and produce the same output as $P$.

2.2 Obfuscation transformations

An obfuscation transformation $T$ can be classified into different categories such as data obfuscation, static code obfuscation, and dynamic code obfuscation. Early techniques are given by Collberg et al. [13, 14]. A classification of all these obfuscations, as well as known deobfuscation methods has been provided by S. Schrittwieser et al. [56]. The following paragraphs present a non-exhaustive list of obfuscation transformations.

2.2.1 Encodings. Static data within binaries, such as strings or constant values, contain useful information for an analyst. Encoding,
as an obfuscation transformation \( T \), converts data to a different representation. To this end, special encoding functions are employed to mitigate the need of storing the static data in clear text within the binary. During execution, the inverse function is used to decode the obfuscated data. To prevent pattern-matching attacks, the obfuscated representation must be parameterized in order to have a family of representations. In other words, each representation renders different-looking obfuscated variables. However, they are all based on the same obfuscating algorithm.

2.2.2 Instructions substitutions. Each program behavior can be implemented in multiple ways [68]. In other words, instructions or sequences of instructions can be replaced with syntactically different, yet semantically equivalent code. As an example, complex instruction substitution include the replacement of call instructions with a combination of push and ret instructions [35]. De Sutter et al [60] replaced infrequently used opcodes with blocks of more frequently used instructions in their work. This transformation reduced the total number of different opcodes used in the code and normalizes their frequency.

2.2.3 Opaque predicates. An opaque predicate [15] represents an obfuscated predicate with its outcome known at obfuscation time, but difficult to determine for a deobfuscator. Opaque predicates are used to make static reverse-engineering more complex. They introduce an analysis problem which is difficult to solve without running the program. There are two types of invariant opaque predicates and the two-ways opaque predicates. Collberg et al defined these predicates by, respectively, \( P_T \), \( P_F \) and \( P_O \) opaque predicates. Several works use two-ways opaque predicates constructs, either referred to as range-dividers [4], or as correlated opaque predicates [42, 69]. Moreover, regardless of their output, e.g. their type, there exists many different kinds of construction that produce the opaque predicates.

2.2.4 Control-flow flattening. This obfuscation transformation aims at obscuring links between basic-blocks by flattening the control-flow. Wang et al. [67] describe as chenxisication this transformation, which puts the basic-blocks of a program into a large switch-statement. A dispatcher decides then where to jump next. Control-flow flattening using a central dispatcher is also described by Chow et al. [11]. A similar concept by Lynn and Debray [38] uses what is called branch functions, which directs the control-flow to the actual target based on a call table. Further control-flow obfuscation constructions are described in [10, 17, 36, 46, 55].

2.2.5 Code virtualization. Code virtualization describes the concept of converting a program functionality into byte-code for a custom virtual machine interpreter that is bundled with the program [23, 31]. This obfuscation transformation can also be combined with polymorphism by implementing custom virtual machine interpreters and payloads for each instance of the program [2]. Other work [66] proposes the combination of fine-granular encryption and code virtualization to hide the virtual machine code from analysis. Collberg et al. [14] describe a variant of this concept under the term table interpretation. A similar concept by Monden et al. [43] uses a finite state machine-based interpreter to dynamically map between instructions and their semantics. Thus, code virtualization proposes many constructions, as for previous transformations.

2.2.6 Dynamic code modification. In this technique, similar functions are obfuscated by providing a general template in memory that is patched right before its execution [14]. Static analysis techniques fail to analyze the program, as its functionality is available at runtime only. Other concepts of dynamic code modification [30, 40] implement the idea of correcting intentionally erroneous code at runtime, right before execution.

Our goal in this paper is to evaluate our methodology against the previously presented obfuscation transformations. Beforehand, the next section will recall some notions about supervised machine learning techniques for classification.

2.3 Supervised machine learning

Supervised machine learning [26, 33] provides a dedicated methodology to produce general hypotheses from external supplied instances via a given algorithm. From these hypotheses, predictions about future instances are possible. The aim of a supervised machine learning is to build a classification model which will be used to assign labels to unknown instances. In other words, let \( X \) be an input (i.e. instance) and \( Y \) the output (i.e. predicted label). A supervised machine learning algorithm will be used to learn the mapping function \( f \) such that \( Y = f(X) \). The goal is to approximate \( f \) such that for any new instance \( X \) we can predict its label \( Y \). In our case the inputs are represented by \( n \)-dimensional vectors of numerical features for which the extraction is described in the following paragraph. The traditional single-label classification associates an instance \( X \) with a unique label \( Y' \) from a previously known finite set of labels \( L \). This approach is then considered a binary classification problem if \( |L| = 2 \), or a multi-class classification problem if \( |L| > 2 \). Other approaches exist, such as the multi-label classification. In this case, an instance \( X \) is associated with a set of labels \( S_Y \subset L \). Moreover, if the model is based on a mapping function \( f \) that can return a set of multiple labels, we have a multi-output classification model. In our work, we use all these classification problems as described in Section 3.

2.3.1 Feature extraction. Instances of a machine learning model are usually derived from what is called raw data, i.e. the data samples we want to classify or predict. These data samples cannot be directly given to a classification model and need to be processed beforehand. This processing step is called feature extraction [25] and consists in combining the raw data variables into numerical features. It allows to effectively reduce the amount of data that must be processed, while accurately describing the original dataset of raw data. In our case, raw data are text documents (e.g. disassembly code, symbolic execution state, etc.). Therefore, one practical use of feature extraction consists in extracting the words (i.e. the features) and classify them by frequency of use (i.e. weights). Different approaches exist for understanding what a word is and to compute its weight. In this paper we use the bag of words approach [41], which identifies terms with words using term frequency, in order to extract the features for our model. It is an efficient and simple approach which fits adequately our semantic reasoning approach.

2.3.2 Classification algorithms. The choice of which specific learning algorithm to use is a critical step. Many classification algorithms
exist [28], each of them having different mapping functions. Classification is a common application of machine learning. As such, there are many metrics that can be used to measure and evaluate our models. In order to compare these metrics, \( k \)-fold Cross-Validation [32] is a frequently used technique. The definition of \( k \)-fold cross-validation consists in reserving a particular set of samples on which the model does not train. The limited set of samples allows to estimate how the model is expected to perform on data not used during the training phase. The parameter \( k \) refers to the number of groups that a given dataset of samples is split into, in order to calculate the mean of our models accuracy as well as the F1-score based on the value of \( k \). While the accuracy of the model represents the ratio of correctly predicted labels to the total of labels, F1-score takes both false positives and negatives into account. In our experimentations and evaluations, the accuracies and F1-scores are calculated using \( k \)-fold cross-validation, with \( k = 10 \) for a better generalization of our model to unknown instances.

Another application of cross-validation, introduced in [53], consists in a functionality-based folding. In other words, the learning set and training set are divided based on the functionality of the samples from which the raw data are generated. The goal of such evaluation methodology is to measure if the model is dependent to the underlying code functionality, independently of the obfuscation transformation applied. The next paragraph introduce furthermore the work of Salem et al. [53], known as metadata recovery attack.

2.4 Related work: Metadata recovery attack

Salem et al. [53] introduce the use of machine learning techniques to evaluate the stealth of obfuscation transformations throughout their detection (known as metadata recovery attack). Their primary hypothesis is that machine learning techniques are capable of implementing these attacks by classifying obfuscated programs according to the transformations applied. Their experiments are based on two learning algorithms, namely Naive Bayes [22] and Decision trees [52]. Their raw data are based on static disassembly or dynamic instruction traces, either stripped or not. Thus, we refer to such raw data generation as syntax-reasoning. The evaluation of their models is made with two classification techniques. The first one is a traditional \( k \)-fold cross validation, with \( k = 10 \). The second one is more fine-tuned since it discriminates the training and test dataset on program functionality. In other words, the test dataset is excluded of any raw data that have been used in the training dataset, based on the functionality they implement. Such process is also repeated 10 times, to calculate the average accuracy for each fold. Their results are promising, showing up to 100% of accuracy for obfuscation transformations detection with decision trees, on dynamic traces. However, these results are obtained with the conventional cross-validation, whereas the second classification mode provides lower results (up to 61% of accuracy) with decision trees. This indicates that their model is dependent of the functionality implemented in their raw data. Moreover, their work is not implemented yet to cover several layers of obfuscation transformations, as it can be the case in most obfuscated programs. In our work, we also used both cross-validation approach to compare our results with their work. This gives an brief idea about the advantages of semantic reasoning over syntax-based approaches.

Our goal in this paper is to combine semantic reasoning and more advanced machine learning classification techniques in order to improve the accuracy. We want to have a static analysis tool, based on symbolic execution, in order to have a model that does not depend on the functionality of the program. The models are used to detect several layers of obfuscation transformations, thus having a multi-label and multi-output classification problem. Then, we extend our detection not only to the obfuscation transformations but also to their constructions. To this end, in the next section, we present our approach and methodology.

3 METHODOLOGY

In this section we present our methodology composed of several steps, as illustrated in Figure 2. I. In order to create our models, we generate obfuscated as well as clean samples. This generation is done using publicly available obfuscators, specifically Tigress and OLLVM. II. We employ then semantic reasoning via symbolic execution\(^1\) to extract our raw data, from the generated samples. This step is presented in Section 3.1. III. We create two different datasets for two different kinds of classifications. Using labeled raw data, we build our datasets for the detection of obfuscation transforms, including several combinations. Another dataset is made for the detection of specific constructions related to the transformations. These steps are introduced in Section 3.4. IV. The previous datasets are used to train our models. In order to select the most relevant approach and learning algorithms, several studies and experiments are provided in Section 4. V. The final step consists in their evaluation and their application on unknown instances, as presented in Section 5.

3.1 Semantic reasoning

Static symbolic execution is a binary analysis technique that captures the semantics (i.e. logic) of a program. An interpreter is used to trace the program, while assuming symbolic values for inputs rather than obtaining concrete values as a normal execution would. A symbolic state \( S \) is built and consists in a set of symbolic expressions \( \mathcal{S} \) for each variables (i.e. registers, memory, flags, etc.). Several techniques exist for symbolic execution [3]. In order to avoid path explosions in static symbolic execution, we use an intra-procedural and bloc-centric approach, as summarized next.

3.1.1 Bloc-centric intra-procedural symbolic execution. We use semantic reasoning for the generation of our raw data. The symbolic representation helps to efficiently detect obfuscation transformations and constructions. Raw data refers to the representation of data samples, containing noisy features, which need to be processed in order to extract the informative characteristics to train the models. For the detection of obfuscation transformations, we choose to work on disassembled functions of binary code. On these functions, we apply static symbolic execution to retrieve their semantic representation. In our work we use disassembled functions to collect the symbolic expressions from the code, as illustrated in Algorithm

\(^1\)In our work we consider semantic retrieval only. We are not interested in generating inputs for program exploration.
1. First, the semantic reasoning part of our methodology is given a disassembled function $F$ as input. For the learning phase of our methodology, $F$ needs to be labeled. In other words, we need to know which transformations are applied in order to properly train our model. However, in order to use our methodology as a static and automated detection framework, $F$ does not require to be labeled once the models are trained. Based on $F$, we iterate over each basic block $B$. We then collect the instructions of $B$, denoted by $I_B$, with the function $getInstructions()$. $I_B$ is translated into an intermediate language, denoted by $IR_B$, using $getIntermediateLanguage()$. Finally, $IR_B$ is being used for the block-centric symbolic execution function $symbolicExecution()$. The latter will return the symbolic state $S_B$, in order words, expressions of each modified variables in a static single assignment form, based on the intermediate representation $IR_B$ previously used. The generated semantics $S_B$ is then normalized using $normalizeSemantics()$ function. Finally, the normalized semantics $NS_B$ is added to the dictionary $L$ containing all normalized semantics for each processed basic block $B$. The content of $L$ will be used to generate our raw data as text file. Our normalization step has the crucial role of making the model scale to unknown data. Next, Section 3.2 describes this step, along with the content of our raw data.

3.2 Semantic-based raw data

Intermediate representations often use concrete values within their generated expressions. This causes raw data to depend on addresses that are specific to some binaries and prevents our models to scale on unknown data. Some intermediate representations also use identifiers in order to express modified registers or memory areas.

Listing 1: Symbolic state using Miasm2 intermediate language

This notation may further affect the scalability of our trained models. For the purpose of having a model that can scale to unknown data we use a normalization phase. The normalization consists in replacing all identifiers and concrete values by symbols, and non-alphanumerical characters by alphanumerical words. This is a necessary step for a complete features extraction phase that sometimes excludes non-alphanumerical characters when working on text-based raw data. In our methodology, we generate the raw data using the Miasm2 [18] intermediate language. This language is part of the symbolic execution engine that we use for the implementation of our methodology as IDA Pro plug-in. Additionally, the normalized Miasm2 intermediate language has also been successful for the application of machine learning techniques in order to de-obfuscate opaque predicates [61]. Listing 1 illustrates the symbolic representation of the disassembled function $F$ in Miasm2.
We can see that with the weaknesses of each algorithm. In order to select the best words to express the semantics of the basic blocks. For example, we base our core methodology on voting classifiers. The purpose is to create automated and efficient models for the classification, in order to return all the detected obfuscation layers, as illustrated in Figure 1. Note that the complete raw data will contain the symbolic states of each basic-blocks of the quick-sort function.

We use multi-label with multi-output classification. For example, if our set of labels are the applied transformations, namely control-flow flattening and code virtualization, then one binary can have both protections. In such case, our methodology needs to return all predicted labels. We then refer to such model as a multi-output classification.

Multi-label classification methods differ from binary or multi-class approaches. Tsoumakas et al. [63] group multi-label classification methods into two categories: problem transformation methods that transform the multi-label classification problem either into one or more single-label classification problems, and algorithm adaptation methods that extend specific learning algorithms in order to handle multi-label data directly. In our methodology we use classifier chains [49], where each model is an ensemble of learning algorithms, as presented in Section 3.3. We also study the binary relevance methodology [24] in Section 4. These two methodologies are briefly introduced in the following paragraphs.

3.3 Ensemble learning

In machine learning, ensemble methods [19] use multiple learning algorithms. They are mostly used to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [39, 51]. An ensemble, in this case, consists of a set of individually trained classifiers whose predictions are combined when processing novel instances. Different families of ensemble learning methods exist, e.g. Bagging [9], Boosting [20, 21] or Stacking [57]. Since every model has its strengths and weaknesses, ensemble models combine individual models to help cope with the weaknesses of each algorithms. In order to select the best possible predictions from our ensemble, we use a voting [58] algorithm. Hence, a model is selected to make the final prediction by a simple majority vote for accuracy. Our work aims to study the benefits of ensemble learning approach over individual models. Thus, we base our core methodology on voting classifiers. However, a more in-depth studies of other approaches could provide better insights into the reasons why/why ensemble models get consistently better results for this task.

3.4 Multi-label and multi-class classifications

Multi-label classification methods are increasingly required by modern applications [8, 37]. We use multi-label with multi-output classification, in order to return all the detected obfuscation layers, specially when combined. We also focus on multi-class classifications which play a key role in our methodology due to the following facts:

1. the detection of all the applied obfuscation transformations is a multi-label classification problem. For example, if our set of labels are the applied transformations, namely control-flow flattening and code virtualization, then one binary can have both protections. In such case, our methodology needs to return all predicted labels. We then refer to such model as a multi-output classification.

2. the fine-grained detection of the constructions is a multi-class classification problem. For example, if we know that control-flow flattening is applied on a code, then its constructions can only be one unique label (e.g. switch-based, ifnest-based, indirect, call-based, etc.).

4 EXPERIMENTS

In this section we present first the dataset used, common with previous related work [53, 61]. Our preliminary studies towards an efficient implementation of a fine-grained detection framework are...
also introduced. All our experiments and evaluations are done on a
Windows 7 laptop, using 16GB of RAM, and an Intel processor.

4.1 Datasets
Our experiments are made on several C code samples. We use the
scikit-learn API [45] for the implementation of the models. The
datasets contain various types of code, each of them having different
functionality in order to have models that do not fit to a specific
type of program. The used samples are listed below:

- GNU core utilities (i.e. core-utils) binaries [48] for normal
  predicate samples;
- Cryptographic binaries for obfuscated and non-obfuscated
  predicates [16];
- Samples from [4] containing basic algorithms (e.g. factorial,
sorting, etc.), non-cryptographic hash functions, small pro-
grams generated by Tigress;
- Samples involving the uses of structures and aliases [1, 27].

Our choice is motivated by the samples low ratio of dependencies
and their straightforward compilation. This makes their obfuscation
possible using tools such as Tigress and OLLVM without errors
during compilation. Furthermore, all datasets used for the studies and
evaluations are balanced and contain between 1000 to 5000 samples.

The obfuscation transformations applied are given in Appendix B
and A. The next section will present our studies based on these
datasets.

4.2 Preliminary studies
Our goal in this section is to provide some answers to the following
questions related to our methodology:

- **Study 1**: when only one obfuscation transformation is ap-
  plied, is a single model more effective than ensemble models
  for the detection?
- **Study 2**: when several obfuscation transformations are ap-
  plied, can the model from Study 1 be applied to the multi-
  label and multi-output classification problems?
- **Study 3**: when several obfuscation transformations are ap-
  plied, is a multi-label and multi-output model more efficient
  than one binary model for each transformation, i.e. classifier
  chains?
- **Study 4**: for the fine-grained detection of obfuscation con-
  structions, is a single model more efficient than ensemble
  models?

Our studies and evaluations present two different types of results
based on two different evaluations approaches. One is the tradi-
tional k-folds cross-validation with scores in black colored font.
The other is made with the functionality-based cross-validation
approach in red colored font, used in Salem et al. related work [53].
Besides, we use as a traditional single-model random-forest al-
gorithm throughout all our studies. As for the ensemble models,
we combined extra-tree and random-forest learning algorithms.
These algorithms were selected because they provided the best
scores in terms of accuracy. For simplicity, a preliminary evaluation
was made between several learning algorithms [34] (e.g. decision
trees, k-nearest neighbors, support vector machines, neural net-
work, naive Bayes, random forest, etc.). In order to select the best

<table>
<thead>
<tr>
<th>Obfuscation transformation</th>
<th>Mono-model</th>
<th>Ensemble-learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tigress transformations</td>
<td>Random-forest</td>
<td>Extra-tree &amp; Random-forest</td>
</tr>
<tr>
<td>EncA</td>
<td>93% / 98%</td>
<td>95% / 100%</td>
</tr>
<tr>
<td>EncL</td>
<td>100% / 97%</td>
<td>100% / 100%</td>
</tr>
<tr>
<td>EncD</td>
<td>95% / 98%</td>
<td>95% / 100%</td>
</tr>
<tr>
<td>AddO</td>
<td>100% / 100%</td>
<td>98% / 100%</td>
</tr>
<tr>
<td>Flat</td>
<td>97% / 100%</td>
<td>97% / 100%</td>
</tr>
<tr>
<td>Virt</td>
<td>100% / 100%</td>
<td>100% / 100%</td>
</tr>
<tr>
<td>Jit</td>
<td>100% / 100%</td>
<td>100% / 100%</td>
</tr>
<tr>
<td>clean</td>
<td>91% / 100%</td>
<td>91% / 100%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>97% / 99%</td>
<td>97% / 100%</td>
</tr>
</tbody>
</table>

Table 1: Multi-class accuracy and F1-scores per labels for the
detection of Tigress obfuscation transformations (1 layer).

Illustrates our results where we see that ensemble-learning provides
a similar accuracy to random-forest, up to 97%, with traditional
cross-validation. The illustrated F1-scores per labels, namely the
obfuscation transforms, also points out that most of them are pred-
icated similarly with both approaches. An exception is made for
arithmetic encoding, i.e. EncA, and opaque predicates, i.e. AddO.
With the functionality-based cross-validation approach however,
the results differs more as observed in red font. Ensemble-learning

technique provides 100% accuracy and F1-score for each classes,
whereas random-forest achieves slightly lower results, with an
average accuracy at 99%. Due to the semantic reasoning of our
methodology, the results are better with this approach when hav-
ing one layer of obfuscation. Yet, these results are not sufficient
to select traditional mono-models over ensemble-learning, or the
opposite way. Hence, the next study will experiment these two
approaches for multi-label and multi-output classification.

4.2.2 Study 2: In the following study, we combine all obfuscation
transformations. The goal of our model is to correctly predict all
the applied layers of obfuscation transformation. Thus, each sample
can have one or more labels. We aim to compare the random-forest
algorithm with the ensemble model based on random-forest and
extra-trees for multi-label and multi-output classification. Our re-

<table>
<thead>
<tr>
<th>Obfuscation transformation</th>
<th>Multi-label mono-model</th>
<th>Multi-label ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tigress transformations</td>
<td>Random-forest</td>
<td>Extra-tree &amp; Random-forest</td>
</tr>
<tr>
<td>EncA</td>
<td>95% / 93%</td>
<td>96% / 92%</td>
</tr>
<tr>
<td>EncL</td>
<td>96% / 78%</td>
<td>92% / 85%</td>
</tr>
<tr>
<td>EncD</td>
<td>95% / 93%</td>
<td>96% / 92%</td>
</tr>
<tr>
<td>AddO</td>
<td>96% / 88%</td>
<td>97% / 88%</td>
</tr>
<tr>
<td>Flat</td>
<td>98% / 97%</td>
<td>99% / 91%</td>
</tr>
<tr>
<td>Virt</td>
<td>99% / 98%</td>
<td>99% / 99%</td>
</tr>
<tr>
<td>Jit</td>
<td>100% / 95%</td>
<td>97% / 92%</td>
</tr>
<tr>
<td>clean</td>
<td>96% / 90%</td>
<td>91% / 87%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>96% / 88%</td>
<td>92% / 92%</td>
</tr>
</tbody>
</table>

Table 2: Multi-label accuracy and F1-scores per labels for the
detection of Tigress obfuscation transformations (several layers).

Results in Table 2 illustrate that traditional cross-validation provides
a higher overall accuracy for ensemble learning classifier as opposed to random forest. Our ensemble of models scores 92% as opposed to 90% for random-forest, with F1-scores per labels above 91%. The functionality-based cross-validation provides lower results, with an overall accuracy at 83% and at 82% for respectively random forest and ensemble models. Still, our result indicates that both approaches can efficiently detect several layers of obfuscation transforms. However, we may improve our results using problem transformations methods such as classifier chains.

The next study will experiment this hypothesis.

4.2.3 Study 3: As in the second study, we combine all obfuscation transformations but we use binary classification problem for multi-label and multi-output classification using classifier chains. Our results with standard cross-validation does not differ much from previous Study 2 as illustrated in Table 3. The functionality-based cross-validation provides improved overall accuracies and F1-scores per labels. Ensemble models used in classifier chains provide 90% of overall accuracy, compared to random-forest used in classifier chains that score 85% of overall accuracy. This study led us to select ensemble-learning techniques with classifier chains in our methodology since classifier chains allow us to create an efficient and accurate model for the detection of obfuscation transformations with one or more layers.

4.2.4 Study 4: For this final study, our goal is to evaluate the models for the fine-grained detection of an obfuscation transformation construction. We use in our dataset several Virtualized samples with Ti
gress for our experiment. Ti
gress allows the user to select different kinds of constructions, such as switch-based, ifnest-based, linear-based, interpolation-based for example. This experiment is equivalent to Study 1 in the sense that it is a multi-class classification problem. Namely, each sample has a unique label and the selected model will return one unique label per instance.

Our results in Table 4 show that both random-forest and ensemble models provide the same F1-scores per labels. Their overall accuracies with standard cross-validation are also with 100% accuracy. With functionality-based cross-validation, ensemble models are slightly more efficient with a 100% accuracy as opposed to 99% for mono-model based on random-forest. This led us to select ensemble models in our methodology also for the classification of constructions, as it allows a fined-grained detection capability.

5 EVALUATIONS
In this section we evaluate our models with respect to the following classification problems:

1. **Multi-label and multi-output evaluation:** can our model, based on a classifier chain of ensemble models, efficiently and accurately detect all obfuscation transformations when one or more layers are applied?

2. **Multi-class evaluation:** once the obfuscation transformation detected, can our ensemble model efficiently and accurately detect the construction of the latter?

We use both cross-validation evaluation schemes as detailed in Section 2.3.2. Our evaluations are made with publicly available obfuscators, specifically Ti
gress and OLLVM, in order to combine obfuscation transformations from different tools.

### 5.1 Transformations detection

Our goal is to evaluate the stealth of obfuscation transformation, either applied as unique layer or combined. We use our multi-label and multi-output model based on ensemble-models and classifier chain to detect all the transformations applied. To measure the efficiency of our model, we used both traditional and functionality-based cross-validation as explained in Section 2.3.2. A list of all combinations of the applied transformations used in our evaluations can be found in Appendices A and B. Additionally, command line options for Ti
gress and OLLVM are given in A.1 and B.1.

<table>
<thead>
<tr>
<th>Obfuscation transformation</th>
<th>Mono-model chain</th>
<th>Ensemble chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tigress transformations</td>
<td>Random-forest</td>
<td>Extra-tree &amp; Random-forest</td>
</tr>
<tr>
<td>EncA</td>
<td>95% / 92%</td>
<td>97% / 90%</td>
</tr>
<tr>
<td>EncL</td>
<td>90% / 80%</td>
<td>93% / 87%</td>
</tr>
<tr>
<td>EncD</td>
<td>95% / 92%</td>
<td>97% / 96%</td>
</tr>
<tr>
<td>AddI</td>
<td>96% / 92%</td>
<td>97% / 88%</td>
</tr>
<tr>
<td>Flat</td>
<td>97% / 97%</td>
<td>99% / 91%</td>
</tr>
<tr>
<td>Virt</td>
<td>99% / 98%</td>
<td>99% / 99%</td>
</tr>
<tr>
<td>Jit</td>
<td>100% / 90%</td>
<td>100% / 92%</td>
</tr>
<tr>
<td>clean</td>
<td>88% / 96%</td>
<td>92% / 96%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>90% / 85%</td>
<td>92% / 90%</td>
</tr>
</tbody>
</table>

Table 3: Classifier chain accuracy and F1-scores per labels for the detection of Ti
gress obfuscation transformations (several layers).

### 5.1.1 OLLVM

Our first evaluation uses OLLVM. It implements transformations such as opaque predicates (i.e. bogus control flow, bcf), instruction substitutions (i.e. sub) and control-flow flattening (i.e. fla). We built a dataset with several combinations of these transformations (c.f. Appendix B) in order to measure the efficiency of our model. Table 5 shows our results. Our model achieves an overall accuracy of 86% with traditional cross-validation and 89% with the functionality-based one. F1-scores for labels bcf, fla, and clean where no transformations are applied, are over 92% and up to 98% for bcf. However, the efficiency of our model to detect OLLVM instructions substitutions transformations, labeled as sub, achieves a low F1-score at 80%. Further evaluations indicate that sub is often considered clean by our model. Thus, when combined with other transformations, sub transformation is often undetected.

<table>
<thead>
<tr>
<th>Obfuscation transformation</th>
<th>Classifier Chain</th>
<th>Ensemble model</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLLVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bcf</td>
<td>98% / 98%</td>
<td></td>
</tr>
<tr>
<td>fla</td>
<td>92% / 95%</td>
<td></td>
</tr>
<tr>
<td>sub</td>
<td>82% / 80%</td>
<td></td>
</tr>
<tr>
<td>clean</td>
<td>94% / 93%</td>
<td></td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>86% / 89%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Evaluated accuracy and F1-scores per labels for the detection combined OLLVM transformations.
5.1.2 **Tigress.** Our second evaluation is made with the Tigress obfuscator. Tigress can generate state-of-the-art transformations such as dynamic-code generation (i.e. Jit), code-virtualization (i.e. Virt), control-flow flattening (i.e. Flat), opaque predicates (i.e. AddO) and several encoding (i.e. Arithmetics, Literals and Data, respectively EncA, EncL and EncD), among others. As illustrated in Table 6, our model accuracy is up to 90% with standard cross-validation. With functionality-based cross-validation, the overall accuracy is at 91%. F1-scores for heavy transformation such as Virt and Jit are up to 99% and 100%. The lowest F1-score is for i.e. EncL which is sometimes considered as a clean sample by our model. Regardless, our evaluation underlines the accuracy and efficiency of our methodology against Tigress transformations.

5.1.3 **OLLVM and Tigress.** For this evaluation we combine both OLLVM and Tigress datasets. We aim to see if our model is able to detect common obfuscation transformations. Table 7 shows our results. F1-scores for heavy transformations such as Virt, Jit and Flat are high, averaging up to 100% for Jit as an example. Combined test samples between obfuscators such EncA-sub, AddO-bcf, and Flat-fla have high F1-scores, even when combined with other transformations. These heavy transformations introduce important side-effects to the code, allowing an efficient and accurate detection of our model. The ability to efficiently detect non-obfuscated samples is still low compared to the ability to detect all layers of obfuscation transformations. In that case, our model F1-scores are up to 83% and 80% depending on the cross-validation approach used. Still, our model is averaging an accuracy up to 88% and 86%. These overall accuracies illustrate our model efficiency regarding the detection of obfuscation transformations, even when combined, and between the two different obfuscators.

5.1.4 **OLLVM vs. Tigress.** Our final evaluation aims to compare the accuracies of our model depending on the learning dataset used. First, we use a learning dataset only based on OLLVM transforms. The model will be then evaluated against some similar obfuscation transformations generated by Tigress. Second, we do the opposite, namely train our model on Tigress samples to evaluate it on OLLVM raw data. The results are displayed in Table 8. As we can see, our model efficiently detects Tigress Flat transformation when training on 1000 samples of all OLLVM transforms, with 100% of accuracy. Results are lower when the training dataset is based on Tigress (4000 samples), against OLLVM fla transform, with an overall accuracy up to 95% with a standard cross-validation. Moreover, we can observe that our model cannot efficiently detect Tigress opaque predicates, i.e. AddO, when training only on OLLVM transforms. The results, in that case, indicate that our model efficiently detects the Flat transformation, but only few AddO ones. Finally, when our model is trained on Tigress, the overall accuracy is up to 82% against all OLLVM transforms (c.f. Appendix B). This result indicates that our methodology provides some genericity.

5.2 **Constructions detection**

In this section we evaluate our model for the detection of specific obfuscation transformations constructions. We use our multi-class model, based on ensemble-models, to provide a fine-grained detection technique. As for previous evaluations, we use traditional and functionality-based cross-validation techniques.

5.2.1 **Control-flow flattening.** As for code virtualization, control-flow flattening can also be constructed in several ways, as introduced in Section 2.2. Facing the same limitations as for code virtualization constructions, we evaluated two constructions namely switch-based from the Tigress obfuscator, and if-nest-based from OLLVM. The evaluation results are in Table 9. Our model averages high F1-scores and accuracy, the latter being at 98% with standard cross-validation evaluation.

5.2.2 **Opaque predicates.** Many opaque predicates constructions exists, some of them having as purpose preventing the usage of existing deobfuscation techniques based on dynamic-symbolic execution. For the detection of their constructions, we used Tigress,
OLLVM but also novel bi-opaque methods [70]. Our results in Table 10 show that our model is accurately detecting opaque predicates constructions. F1-scores are up to 100% with standard cross-validation. Bi-opaque constructions are however often undetected when combined with other transformations.

<table>
<thead>
<tr>
<th>Opaque predicates</th>
<th>Ensemble model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tigress and OLLVM</td>
<td>Extra-tree &amp; Random-forest</td>
</tr>
<tr>
<td>Floats</td>
<td>85% / 89%</td>
</tr>
<tr>
<td>Symbolic-memory</td>
<td>87% / 93%</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>100% / 100%</td>
</tr>
<tr>
<td>Aliasing</td>
<td>100% / 99%</td>
</tr>
<tr>
<td>Mixed-boolean and arithmetic</td>
<td>100% / 96%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>95% / 99%</td>
</tr>
</tbody>
</table>

Table 10: Evaluation accuracy and F1-scores per class for the detection of opaque predicates constructions.

Yet, the overall accuracy of our model is at 95% and 93% depending on the evaluation approach used. This illustrates the efficiency of our methodology towards the detection of obfuscation transformations constructions.

6 LIMITATIONS

One threat to the validity of our results is that we only use datasets of relatively small C programs, except for the core-utils binaries used for non-obfuscated samples. Nevertheless, the samples used in our dataset involve all common programming language constructions and various functionalities (e.g., hash functions, sorts, cryptographic algorithms, etc.). However, our future work will include the evaluation of our methodology on other obfuscators or programs, such as malwares. Our work shows that semantic reasoning combined with advanced machine learning present capabilities for a fine-grained detection of obfuscation transforms.

The capability of detecting unknown transformations or constructions represents another limitation of our methodology. If our model did not train on one specific transformation or constructions, it will not predict properly the unknown sample. This can lead to a loss of accuracy when unknown transformations are combined.

Dynamic transformations cause limitations to our model for the static detection of obfuscation transforms. Despite from the fact that we are able to accurately detect some of these transformations (i.e. JIT, Virt), when other obfuscation transformations are applied before them, our model is less efficient. Moreover, other transformations such as packing, or anti-symbolic execution techniques may lower the accuracy of our model. However, as pointed out in the next section, our methodology can scale to dynamically collected traces which allows to thwart some of these limitations.

7 PERSPECTIVES AND FUTURE WORK

First, more in-depth studies of aggregation approaches used in ensemble learning must be done in order to assess if ensemble learning are consistently more efficient for that task compared to mono-models. The hard voting scheme used is a simple approach, but may not achieve the effective benefit of using the ensemble learning approach.

As seen in [61], semantic reasoning and machine learning provides promising results for deobfuscation methodology. The evaluations shown in this paper illustrate that our model does not depend on the code functionality. A more accurate comparison must be made as future work.

To overcome the dynamic transformations limitations, we can adapt our methodology to dynamically collected instructions traces. With a given instructions trace, we reconstruct each basic-blocks and apply our semantic reasoning approach in order to generate raw data. This step can be done either for the learning or the evaluation phase. Our future work consists in extending the implementation of our current framework and evaluating other combinations of obfuscation transformations based on dynamic traces.

Another issue we need to consider is the application of n layers of the same obfuscation transformations. Presently, our evaluations is done by combining several transformations, but using one time each of them. Our future study should consider the extension of our evaluations to the use of one transformation several times.

We also plan on extending our datasets of C programs with more complex real-world software libraries in the interest of strengthening our experiments.

8 CONCLUSIONS

In this paper we presented the efficiency of semantic reasoning combined with advanced machine learning techniques. This combination is motivated by the construction of a fine-grained detection framework of obfuscation transformations and constructions. By extending our approach to multi-label and multi-output classification, we enhanced metadata recovery attacks to the detection of multiple layers of obfuscation transformations. We proposed a new approach that combines a block-centric symbolic execution with machine learning ensemble model and classifier chains. We used our models to evaluate the stealth of both obfuscation transformations and constructions. Our results are promising, with overall accuracies up to 91% for the transformations and 100% for the constructions, showing slight improvements with respect to current mono-models machine learning. The use of static symbolic execution allows us to be dependent on the underlying functionality of the code samples used for the learning phase. Our empirical studies illustrate that our choices conduct towards the implementation of an efficient and accurate evaluation framework against state of the art obfuscators. However, there is still place for improvements with a more in-depth study of learning algorithms used and their parameters. Our work slightly improves metadata-recovery attacks, and paves the way towards the efficient use of advanced machine learning combined with semantic reasoning.

ACKNOWLEDGMENTS

We are much grateful to the referees comments and suggestions from the editorial board. This work is supported by the French National Research Agency in the framework of the Investissements d’Avenir program (ANR-15-IDEX-02).

A TIGRESS TRANSFORMATIONS

We list the combinations of obfuscation transformations used for our datasets, in their application order: AddOpaque (16 or 32 times); AddOpaque, EncodeLiterals; EncodeLiterals, AddOpaque, EncodeArithmetics;
We list the combinations of obfuscation transformations used for our datasets, in their application order: bcf; bcf, sub; bcf, sub, fla; bcf, fla, sub; sub; sub, fla; bcf, fla; bcf, sub; sub, fla; bcf, sub, fla; bcf, fla; bcf, sub, fla.

A.1 Commands options

1. **AddOpaque options**
   - `tigress -t Transform_InitEntropy --Transform=InitOpaque --Transform=InitOpakeStruct` list array var = Function::main -- Transform=AddOpaque -- Functions="{3[1] AddOpaqueCount=5 [HEM] AddOpaqueKinds=call, fake, true}
   - `flatten`
   - `tigress -t Transform=flatten --Transform=flattenDispatch=switch goto` -- Functions="{3[1] JitEncoding=hard}
   - `functions="{3[1] InitEntropy=linear"}
   - `enableLiteral`
   - `enableData`
   - `tigress -t Transform=EncodeData --Functions="{5[3] EncodeLiteralKinds=integer, string}"
   - `encodeArithmetics`
   - `encodeLiterals`

Listing 3: Tigress commands for sample generation

B OLLVM TRANSFORMATIONS

We list the combinations of obfuscation transformations used for our datasets, in their application order: bcf; bcf, sub; bcf, sub, fla; bcf, fla, sub; sub; sub, bcf; sub, fla; fla; fla; fla; sub; bcf, fla; bcf, sub; sub, fla; bcf; bcf, fla; sub; bcf.

B.1 Commands options

1. **Augus control-flow**
   - `clang -I {1} -o {1} -mllvm -bcf -mllvm -bcf_prob=50`
   - `clang -I {1} -o {1} -mllvm -bcf -mllvm -bcf_prob=100`
   - `control-flow flattening`
   - `clang -I {1} -o {1} -mllvm -fla`
   - `clang -I {1} -o {1} -mllvm -fla -mllvm -split`
   - `instruction substitution`
   - `clang -I {1} -o {1} -mllvm -sub`

Listing 4: OLLVM commands for sample generation

REFERENCES
