

Poster: Machine Learning for Vulnerability Detection as Target Oracle in Automated Fuzz Driver Generation

Gianpietro Castiglione*, Marcello Maugeri*, and Giampaolo Bella

University of Catania, Italy
 {gianpietro.castiglione,marcello.maugeri}@phd.unict.it
 giampaolo.bella@unict.it

Abstract. In vulnerability detection, machine learning has been used as an effective static analysis technique, although it suffers from a significant rate of false positives. Contextually, in vulnerability discovery, fuzzing has been used as an effective dynamic analysis technique, although it requires manually writing fuzz drivers. Fuzz drivers usually target a limited subset of functions in a library that must be chosen according to certain criteria, e.g., the depth of a function, the number of paths. These criteria are verified by components called *target oracles*. In this work, we propose an automated fuzz driver generation workflow composed of: (1) identifying a likely vulnerable function by leveraging a machine learning for vulnerability detection model as a target oracle, (2) automatically generating fuzz drivers, (3) fuzzing the target function to find bugs which could confirm the vulnerability inferred by the target oracle. We show our method on an existing vulnerability in LIBGD, with a plan for large-scale evaluation.

Keywords: Vulnerability Detection · Vulnerability Discovery · Fuzzing · Machine Learning · Automated Security Testing · Large Language Models

1 Introduction

In recent years, two automated techniques have emerged for finding 0-day vulnerabilities in functions: *Machine Learning for Vulnerability Detection (ML4VD)* [4] and *Fuzzing* for vulnerability discovery.

ML4VD models, when trained on large datasets, are employed to determine whether a given set of functions may exhibit specific vulnerabilities. However, the method involves static analysis, which cannot verify the vulnerability at runtime, may suffer from a significant false-positive rate [2], and ultimately, top-performing models may not be able to differentiate between vulnerable functions and patched functions [4].

On the other hand, fuzzing employs dynamic analysis, reducing false positives from static analysis. However, no push-button fuzzing technique exists, as

* These authors contributed equally

a *fuzzer* requires a *fuzz driver*, a test harness for parsing inputs and invoking the target function that, in turn, requires deep knowledge about the target function and the corresponding library and extensive manual work [1]. *Automated Fuzz Driver Generation (AFDG)*, despite its challenges, solves the burden, with OSS-FUZZ-GEN [3] standing on top due to the use of Large Language Models (LLMs). Contextually, a library could include several functions, and *target oracles* are employed to identify *interesting* functions [5], i.e. functions determined to be likely interesting targets. For example, OSS-FUZZ-GEN prioritises functions relying on FUZZ INTROSPECTOR’s¹ heuristics, which mainly consider the cyclomatic complexity of the target function or the simplicity to generate the fuzz driver. Nevertheless, these heuristics do not account for the likelihood of a function being vulnerable. Hence, we propose a combined method employing ML4VD models as a target oracle to prioritise relevant functions for the AFDG.

Considering these assumptions, this study is based on the following research questions:

- RQ1** Can machine learning for vulnerability detection be an effective target oracle for automated fuzz driver generation?
- RQ2** To what extent can such a combined method confirm the true positives, and/or reduce the number of false positives?

In this work, we present the design and workflow that combines the two techniques in Section 2. We validate the method by selecting a confirmed vulnerable function from the DIVERSEVUL dataset and successfully applying OSS-FUZZ-GEN to generate a fuzz driver that triggers the vulnerability. The target function originates from a project already included in the OSS-FUZZ infrastructure², ensuring compatibility with OSS-FUZZ-GEN, but is not currently covered by an existing fuzz target. This allows us to generate a novel fuzz driver and achieve previously unreachable code coverage. The complete experimental setup is detailed in Section 3. Subsequently, we examine the state-of-the-art, discussing insights or differences from our technique in Section 4. Finally, we discuss our practical contribution and propose a research plan for an in-depth evaluation in Section 5.

2 Design

To address the research questions, the proposed design employs two main techniques: a ML4VD model as the vulnerability detection component (static analysis of the target function code), and AFDG, for ultimately applying fuzzing as the discovery component (dynamic analysis that uncovers vulnerabilities during execution of the target function).

The overall workflow of the proposed method, illustrated in Figure 1, comprises three steps. It represents a generalised pipeline for AFDG and emphasises the main contribution of this work. Namely, the use of the ML4VD model as a target oracle.

¹ <https://github.com/ossf/fuzz-introspector>

² <https://github.com/google/oss-fuzz>

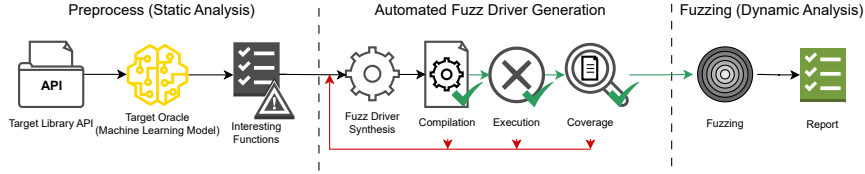


Fig. 1: Workflow of the proposed method

Preprocess (Static Analysis). Our method first identifies functions in the target library API that are likely to be vulnerable. The objective is accomplished by an ML4VD model, which performs a static analysis of each function and flags those that may present one or more weaknesses according to the *CWE* framework by MITRE. Ultimately, the flagged functions are considered potentially interesting functions for further analysis by fuzzing.

Automated Fuzz Driver Generation. Subsequently, the interesting functions are selected for the AFDG process to begin. This is an iterative process in which the fuzz driver synthesis produces a candidate fuzz driver. The candidate must (1) compile, (2) execute without immediate logical failure, and (3) gather sufficient coverage to ensure the input is correctly injected. Once these steps are met, the candidate proceeds to the next stage of the workflow.

Fuzzing (Dynamic Analysis). At this stage, the fuzzing process begins, leveraging the previously generated fuzz driver to inject the target function with a large volume of malformed inputs. At the expiration of the time budget, the fuzzer either discovers a crashing input that confirms the vulnerability or does not.

3 Case Study: LIBGD Library

Target Selection. To evaluate the design of the proposed method, we conducted initial experiments on the LIBGD³ library. In particular, we assumed to have already a ML4VD model trained on the DIVERSEVUL dataset [2], which is considered the best collection of vulnerable functions in C/C++.

Subsequently, we selected among the projects in DIVERSEVUL one already included in the OSS-FUZZ infrastructure and with FUZZ INTROSPECTOR reports available. We used such requirements to identify functions currently not covered by existing fuzz drivers in such a project. Consequently, we directly take an uncovered function labelled with CWEs as most likely classified as vulnerable, belonging to the dataset itself.

In particular, we chose the function *gdImageWebpPtr*, labelled as weak to **CWE-415: Double Free**⁴, which is confirmed by **CVE-2016-6912**⁵.

³ <https://github.com/libgd/libgd>

⁴ <https://cwe.mitre.org/data/definitions/415.html>

⁵ <https://nvd.nist.gov/vuln/detail/CVE-2016-6912>

Experimental Setting. Subsequently, we employed OSS-FUZZ-GEN to generate a fuzz driver. Initially, OSS-FUZZ-GEN relies on FUZZ INTROSPECTOR to retrieve the target function signature and its corresponding arguments, which are provided in YAML format, as illustrated in Figure 2. Since the target function is selected by the target oracle, we assume that our ML4VD model has flagged the *gdImageWebpPtr* function as potentially vulnerable.

Fig. 2: Function Signature of *gdImageWebpPtr*

```

1  functions:
2    - name: "gdImageWebpPtr"
3  params:
4    - name: "im"
5      type: "gdImagePtr"
6    - name: "size"
7      type: "int *"
8  return_type: "void *"
9  signature: "BGD_DECLARE(void *) gdImageWebpPtr (gdImagePtr im, int *size)"
10 language: "c++"
11 project: "libgd"
12 target_name: "gd_webp_fuzzer"
13 target_path: "gd_webp_fuzzer.cc"

```

After this minimal setup, the AFDG process starts with the creation of a prompt from a template to be provided to a Large Language Model (LLM). The prompt template is shown in Figure 3. In particular, we used the standard prompt template provided in the official repository of OSS-FUZZ-GEN; the only difference is the embedded information about the target’s weaknesses we added in the prompt, which is shown in red.

Fig. 3: OSS-Fuzz-Gen prompt

OSS-Fuzz-Gen Prompt
1 - System Prompt You are a security testing engineer who wants to write a C++ program to discover memory corruption vulnerabilities in a given function-under-test [...]
2 - C++ Specific Instructions Use <code><code>FuzzedDataProvider</code></code> to generate these inputs [...]
3 - Instructions and Examples [...] Do not create new variables with the same names as existing variables. WRONG: <pre> <code> int LLVMFuzzerTestOneInput(const uint8_t *data, size_t size) { void* data = Foo(); } </code> [...] </pre>

OSS-Fuzz-Gen Prompt (continue)**4 - Problem Statement**

Your goal is to write a fuzzing harness for the provided function-under-test signature [...] <function signature>

Note: The function is a candidate for the vulnerability CWE-415 (Double Free)

In our initial experiments, we employed *GPT-4* with a temperature setting of 0. Although the model successfully generated valid fuzz drivers on first attempts, the fuzz drivers struggled to find a bug and ultimately confirm the vulnerability. This is because the specific vulnerability can be detected whether *gdImageWebpPtr* is invoked by passing a sufficiently large image. Instead, the fuzz drivers presented a hard-coded limit cap on the size of the image, likely to prevent out-of-memory errors during the fuzzing campaign. To solve the issue, we instructed the model to allow large images without setting a cap. From the initial experiments, we can conclude that if a valid fuzz driver fails to find a bug, this does not rule out a vulnerability, and further instructions based on the expected vulnerability could be needed. Ultimately, our contribution aims to (1) prioritise functions likely to expose a weakness, (2) confirm a vulnerability whenever a critical bug is found.

4 Related Works

Risse *et al.* [4] have shown that top-performing ML4VD models are unable to distinguish between functions that contain a vulnerability and functions where the vulnerability is patched. Consequently, without a definitive solution, we expect an increase in false positives over time, which could be mitigated by our method. The state-of-the-art employs *Directed Greybox Fuzzing (DGF)* to steer the generation of inputs that can reach a specific program location. Zhu *et al.* [7], Yu *et al.* [6], already employed ML4VD as target oracle for DFG. However, DFG itself does not ensure that the target function can be reached. For example, a compiled binary could never call a library function. Our work employs automated fuzz driver generation to generate fuzz drivers which call the target function.

5 Considerations and Future Works

Contributions. FUZZ INTROSPECTOR uses two heuristics to determine functions likely to be interesting targets. However, the heuristics fall short of considering interesting only functions having high cyclomatic complexity and containing *parse* in their name (Heuristic 1), and accepting the same argument types as the fuzzing interface *LLVMFuzzerTestOneInput* (Heuristic 2). Our ultimate aim is to propose a third heuristic, in which an ML4VD model identifies potentially vulnerable functions.

Threats to Validity. As the case study involves a CVE from 2015, this may raise a question about whether the code to reproduce such vulnerability is memorised by the model from the training data. Future studies will focus on evaluating the effectiveness of novel candidate functions.

Future Work. Future work primarily focuses on applying the implemented method to a broader range of functions. Our plan encompasses the integration of an ML4VD model as a target oracle (third heuristic) into FUZZ INTROSPECTOR and evaluating its effectiveness on at least ten projects from both OSS-FUZZ and DIVERSEVUL, relying on OSS-FUZZ-GEN for AFDG.

In particular, to answer **RQ1**, we plan to evaluate our target oracle in novel interesting functions, i.e. not included in the dataset.

To answer **RQ2**, we will focus on functions in a test set, measuring the precision of both the target oracle and, ultimately, the AFDG process.

References

1. Babić, D., Bucur, S., Chen, Y., Ivančić, F., King, T., Kusano, M., Lemieux, C., Szekeres, L., Wang, W.: Fudge: fuzz driver generation at scale. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 975–985. Association for Computing Machinery (2019). <https://doi.org/10.1145/3338906.3340456>, <https://doi.org/10.1145/3338906.3340456>
2. Chen, Y., Ding, Z., Alowain, L., Chen, X., Wagner, D.: Diversevul: A new vulnerable source code dataset for deep learning based vulnerability detection. In: Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses. p. 654–668. RAID ’23, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3607199.3607242>
3. Liu, D., Chang, O., metzman, J., Sablotny, M., Maruseac, M.: OSS-Fuzz-Gen: Automated Fuzz Target Generation (May 2024), <https://github.com/google/oss-fuzz-gen>
4. Risse, N., Böhme, M.: Uncovering the limits of machine learning for automatic vulnerability detection. In: Proceedings of the 33rd USENIX Conference on Security Symposium. SEC ’24, USENIX Association, USA (2024)
5. Weissberg, F., Moler, J., Ganz, T., Imgrund, E., Pirch, L., Seidel, L., Schloegel, M., Eisenhofer, T., Rieck, K.: Sok: Where to fuzz? assessing target selection methods in directed fuzzing. In: Proceedings of the 19th ACM Asia Conference on Computer and Communications Security. p. 1539–1553. ASIA CCS ’24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3634737.3661141>, <https://doi.org/10.1145/3634737.3661141>
6. Yu, L., Lu, Y., Shen, Y., Li, Y., Pan, Z.: Vulnerability-oriented directed fuzzing for binary programs. Scientific Reports **12** (2022), <https://api.semanticscholar.org/CorpusID:247407573>
7. Zhu, X., Liu, S., Li, X., Wen, S., Zhang, J., Çamtepe, S.A., Xiang, Y.: Defuzz: Deep learning guided directed fuzzing. CoRR **abs/2010.12149** (2020), <https://arxiv.org/abs/2010.12149>