

# A Match Made in Heaven? Matching Test Cases and Vulnerabilities With the VUTECO Approach

Emanuele Iannone

*Institute of Software Security*

*Hamburg University of Technology*

Hamburg, Germany

emanuele.iannone@tuhh.de

Quang-Cuong Bui

*Institute of Software Security*

*Hamburg University of Technology*

Hamburg, Germany

cuong.bui@tuhh.de

Riccardo Scandariato

*Institute of Software Security*

*Hamburg University of Technology*

Hamburg, Germany

riccardo.scandariato@tuhh.de

**Abstract**—Software vulnerabilities are commonly detected via static analysis, penetration testing, and fuzzing. They can also be found by running unit tests—so-called *vulnerability-witnessing tests*—that stimulate the security-sensitive behavior with crafted inputs. Developing such tests is difficult and time-consuming; thus, automated data-driven approaches could help developers intercept vulnerabilities earlier. However, training and validating such approaches require a lot of data, which is currently scarce.

This paper introduces VUTECO, a deep learning-based approach for collecting instances of vulnerability-witnessing tests from JAVA repositories. VUTECO carries out two tasks: (1) the “*Finding*” task to determine whether a test case is security-related, and (2) the “*Matching*” task to relate a test case to the exact vulnerability it is witnessing. VUTECO successfully addresses the *Finding* task, achieving perfect precision and 0.83 F0.5 score on validated test cases in VUL4J and returning 102 out of 145 (70%) correct security-related test cases from 244 open-source JAVA projects. Despite showing sufficiently good performance for the *Matching* task—i.e., 0.86 precision and 0.68 F0.5 score—VUTECO failed to retrieve any valid match in the wild. Nevertheless, we observed that in almost all of the matches, the test case was still security-related despite being matched to the wrong vulnerability. In the end, VUTECO can help find vulnerability-witnessing tests, though the matching with the right vulnerability is yet to be solved; the findings obtained lay the stepping stone for future research on the matter.

**Index Terms**—Mining Software Repositories, Vulnerability-witnessing Tests, Security Testing, Unit Testing, Language Models

## I. INTRODUCTION

Software vulnerabilities are flaws violating certain security requirements [1], [2], commonly caused by how the flawed code components handle their inputs [3]. Differently from traditional bugs, vulnerabilities are commonly detected via reactive mechanisms, such as Static Application Security Testing (SAST) tools, penetration testing, or fuzzing [4]–[7], which are often considered enough to detect most security issues [8], [9].

The “shift-left” principle encourages the adoption of a *test-first approach*, in which vulnerabilities are detected proactively via automated testing before the code is committed—in a similar fashion to how bugs are intercepted [2], [10]. Despite having different characteristics [11], [12], we argue that vulnerabilities should be treated like traditional bugs in this sense: Employ *unit tests* to intercept security issues in earlier stages and reduce the risk of shipping them into production. Via

automated tests, vulnerabilities can be found by stimulating a certain behavior by supplying the application with crafted inputs [9], [13], [14]. For instance, a test case for a method affected by a Cross-site Scripting vulnerability would craft exploitative strings, like `<script>alert(1)</script>`, and pass them to the method parameters. Then, it would leverage assertions to ensure the method behaves correctly, i.e., the HTTP response object does not contain the input reflected as-is. A failed assertion would mean that the vulnerability was present. A test case behaving like this is called **vulnerability-witnessing test** [15], [16] (a.k.a. Proof of Vulnerability, PoV [17], [18]), or simply “*witnessing test*” in this context.

Listing 1 shows the test witnessing a path traversal vulnerability (CWE-22) affecting APACHE JSPWIKI, disclosed via CVE-2019-0225. Suppose this test is executed on the vulnerable version (i.e., the one including the `request.getPathInfo()` in the `return` statement highlighted as red). In that case, it will fail due to not returning to the main wiki page—as intended behavior defined by the developers. Interestingly, the complete version of the witnessing test was only added in a later commit, not in the vulnerability-fixing commit.

VUL4J is the primary reference collection of witnessing tests [18], made of manually-validated 108 unit tests matched with 79 vulnerabilities affecting 51 JAVA projects. VUL4J allows reproducing such tests in isolated environments where the project has successfully been built on its vulnerable and patched versions. To find them, the test suite after the patch had been applied was run on both versions; any test methods passing on the former and failing on the latter were manually inspected and confirmed to be witnessing a vulnerability. Despite similar datasets with vulnerability data of other programming languages exist [19]–[22], VUL4J is the only one having working witnessing tests to date.

According to what has been described in the VUL4J original paper [18], the process of reproducing a vulnerability through its witnessing tests was: (1) Select a vulnerability having access to the commit that patched it; (2) checkout and build the project onto that commit, (3) run the test suite found at that revision, (4) checkout and build the commit before representing the vulnerable version, and (5) run the same test suite from before on this vulnerable revision. Such a

```

1 // Fix to CVE-2019-0225
2 public String getForwardPage(HttpServletRequest request) {
3     - return request.getPathInfo();
4     + return "Wiki.jsp";
5 }
6
7 // Vulnerability-witnessing test
8 @Test
9 public void testNastyDoPost() throws Exception {
10    MockHttpServletRequest req =
11        new MockHttpServletRequest("/JSPWiki", "/wiki/Edit.jsp");
12    WikiServlet wikiServlet = new WikiServlet();
13    MockServletConfig config = new MockServletConfig();
14    config.setServletContext(new MockServletContext("/JSPWiki"));
15    wikiServlet.init(config);
16    wikiServlet.doPost(req, new MockHttpServletResponse());
17    wikiServlet.destroy();
18    Assertions.assertEquals("/Wiki.jsp?page>Main&edit", req.getForwardUrl());
19 }

```

Listing 1: Documented fix for CVE-2019-0225 and its related witnessing test in APACHE JSPWIKI.

process had to face several challenges. For instance, not all vulnerabilities can be reproduced, as the building must succeed in both the vulnerable and patched versions. It is also required to observe at least one test case passing on the patched version but failing on the vulnerable one. Therefore, it was required that (1) the witnessing tests exist in the test suite at the patched version and (2) the fix commit must fully resolve the vulnerability (otherwise, the witnessing test might fail in that version as well). The whole process worked for only 79 out of 899 (8.78%) vulnerabilities inspected.

In other words, the manual search of witnessing tests is likely **unsuccessful**—like *finding a needle in a haystack*—and **effort-consuming**. Thus, an automated solution could reduce the burden of finding real-world examples of witnessing tests in the wild. In fact, the security researcher would benefit from an automatic tool that *finds* more examples of vulnerability-related tests—which are currently scarce—to be used for developing novel AI-based techniques for generating security tests [15], [16] or improving the automated repair process [18], [23]–[25]. At the same time, software engineers also would benefit from an automated *matching* mechanism that relates test cases in their projects with the historical vulnerabilities that affected it, likely because they lost—or never had—track of such relationships explicitly.

Therefore, we present VUTECO (**V**ULnerability **T****E**st **C**ollector), a fully-static approach collecting vulnerability-witnessing tests in JAVA test suites. VUTECO addresses two tasks: (1) the “*Finding*” task to determine whether a test case is security-related and (2) the “*Matching*” task to relate a test case to the exact vulnerability it is witnessing. The former occurs through a deep-learning model based on a pre-trained CodeBERT [26] and a logistic classifier to return binary predictions, called *Finder*. The latter task, instead, requires the joint use of the *Finder* and another CodeBERT-based model, called the *Linker*, to validate the match between a test case and a description of a vulnerability with binary predictions. VUTECO has been trained for both tasks using the data in VUL4J, accounting for 62,635 distinct test cases from the 51 projects, of which 108 witnessed 79 vulnerabilities.

After finding the best configurations, VUTECO successfully addressed the *Finding* task with perfect precision and 0.83 F0.5 score, while addressing the *Matching* with satisfactory results, i.e., 0.86 precision and 0.68 F0.5 score. Afterward, VUTECO was employed on a set of 244 open-source JAVA projects affected by 640 vulnerabilities, finding a total of 102 out of 145 (70%) truly security-related test cases. Unfortunately, the match between the test cases and the vulnerabilities did not bring the expected results, failing to match any test case with the exact vulnerability. After a deeper inspection of the matching results, we found that in almost all cases, the test case was still related to security aspects; in some cases, the matched vulnerability was similar to the exact one. Hence, we conclude that VUTECO can benefit researchers and practitioners in finding vulnerability-witnessing tests, though the matching remains a challenging task requiring more attention.

In summary, this paper:

- Introduces VUTECO, the *first ever* approach to find vulnerability-witnessing tests in JAVA repositories and match them with the related vulnerabilities.
- Validates the performance of VUTECO for the *Finding* and *Matching* tasks on datasets extracted from VUL4J, accounting for 62,635 test cases, of which 108 witnessed 79 vulnerabilities in 51 JAVA projects.
- Employs VUTECO to find security-related tests and match them with vulnerabilities in 244 open-source JAVA projects, succeeding in the former case and failing in the latter.
- Releases a replication package containing all the scripts and data connected to the experiments [27].

## II. THE VUTECO APPROACH

### A. Overview

VUTECO faces two distinct tasks. The *Finding* task accepts a JUNIT test case as input and tells whether it is security-related (i.e., “Security”) or not (i.e., “Not-Security”). The *Matching* task accepts a JUNIT test case and a description (in natural language) of a known historical vulnerability (related to the project) and tells whether the test case is witnessing that vulnerability (i.e., “Matched”) or not (i.e., “Not-Matched”).

The VUTECO approach has been wrapped into a tool (having the same name, VUTECO) that automates both tasks. The tool accepts (i) a JAVA project repository, (ii) the project revision (i.e., commit) to inspect, and (iii) a list of descriptions (in natural language) of known historical vulnerabilities related to the project—taken from CVE (Common Vulnerabilities and Exposures). Any GIT-based repository is valid as long as there are JUNIT test cases to process.

Figure 1 depicts the general functioning of VUTECO. Once checking out the project repository to the selected revision, VUTECO parses all JAVA files and marks any class method having the following properties as a test case:

- 1) it is annotated with `@Test` (JUNIT 4 and 5) or its class extends `TESTCASE` or any subclass of it (for JUNIT 3);
- 2) it is not overriding a method defined in class `TESTCASE`, like `run()` or `getName()` (for JUNIT 3);

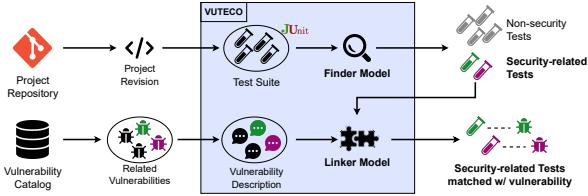


Fig. 1. Graphical overview of the functioning of VUTECO.

- 3) it is not a “lifecycle method”, i.e., annotated with `@BeforeAll`, `@AfterAll`, `@BeforeEach`, or `@AfterEach`;
- 4) it returns `void` if not annotated with `@TestFactory`;
- 5) it is not abstract, static or private;
- 6) its class is not abstract.

Such properties have been designed according to how the JUNIT guide describes a test case [28] and the official JAVADOC of JUNIT beyond version 3.

Each mined test case is sent to the *Finder* model, which is responsible for implementing the *Finding* task (i.e., determining whether the test case is security-related). Then, each test flagged as security-related is sent to the *Linker* model along with the descriptions of all vulnerabilities supplied via input to determine which of them is witnessed by the test case. Sections II-B and II-C describe how the *Finding* and *Matching* tasks have been carried out. The design choices behind the models used there are supported by the experimentation described in Section III-A.

### B. The Finding Task

The *Finding* task consists of determining if a test case is security-related, i.e., it is focused on some security properties of the project where it belongs. The *Finder* model addresses this task, trained to classify test cases into two classes, i.e., “Security” (the positive class) and “Not-Security” (the negative class). The upper part of Figure 2 depicts the architecture of the *Finder* model, which consists of a deep neural network built on top of a pre-trained CodeBERT [26]. The pre-trained CodeBERT starts from the checkpoint `microsoft/codebert-base` downloaded from HUGGINGFACE [29]. CodeBERT has been pre-trained on code examples of six programming languages, including JAVA, paired with natural language text. Hence, we deemed it suitable for understanding the content of JUNIT test methods.

As a preliminary action, VUTECO transforms the input test case by removing new line characters and consecutive white space characters (including tabs), reducing the whole method into a single line. Then, the resulting string is tokenized using the WordPiece algorithm [30] (whose vocabulary was fitted during the pre-training of CodeBERT), and each resulting token is replaced with a numeric identifier. Then, the resulting numeric vector is sent to the CodeBERT input layer (supporting up to 512 encoded tokens), which returns an embedded representation of each token. Besides, due to the underlying BERT-based architecture, an additional embedding

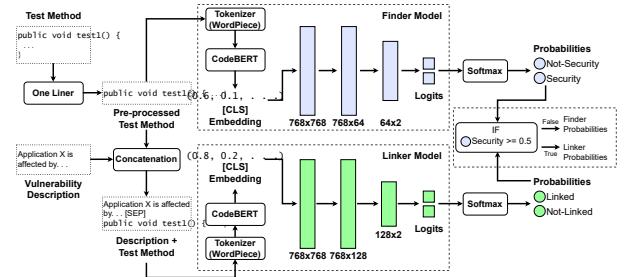


Fig. 2. Graphical depiction of the inner working of VUTECO.

for the special token `[CLS]` is returned, which has the goal of capturing the whole syntax and semantics of the entire sentence (here, the whole test case), making it suitable for sentence classification tasks. Considering the goal of the *Finding* task, we only selected the embedded representation of `[CLS]`, made of 768 values. On top of this, we added a deep neural network with an input layer of 768 neurons, matching the size of the sentence embedding resulting from the CodeBERT model; then, we added two linear hidden layers with 768 and 64 output neurons, respectively, calling GELU activation function [31]. After this, the final layer returns the two logits needed for the final classification, which is done through the Softmax function to return the probabilities for the two classes.

### C. The Matching Task

The *Matching* task consists of determining if a test case and a vulnerability are related, i.e., the former witnesses the latter. This task is carried out by the joint use of the *Finder* model (as in Section II-B) plus the *Linker* model. Indeed, the *Linker* model is focused on a *simplified* task that assumes the input test case is surely security-related. The motivation for the joint use of two models—rather than the direct use of the *Linker* model—is supported by the results of the in-vitro experimentation in Section IV.

The *Linker* model has been trained on its simplified task, i.e., trained to classify pairs of security-related tests and vulnerability descriptions into two classes, i.e., “Linked” (the positive class) and “Not-Linked” (the negative class). The lower part of Figure 2 depicts the architecture of the *Linker* model. Such a model follows a similar architecture to the *Finder* model, leveraging the same pre-trained CodeBERT model to create the input embeddings. However, instead of creating the sentence embedding of the sole test case code (still reduced to a single line), the *Linker* model also adds the vulnerability description as well. Therefore, the vulnerability description was *concatenated* before the test case with a special token `[SEP]` in between to indicate the two different sentences (as required by BERT-like models) and then tokenized in the same manner as the *Finder* model. The resulting sentence embedding is then sent to a deep neural network with an input layer of 768 neurons (to match the embedding size) and a linear hidden layer of 256 output neurons, calling the

GELU activation function. Like the *Finder* model, the final layer returns the logits for the classification, converted into probabilities of the two classes with the Softmax function.

To enable the joint training and use of the *Finder* and *Linker* models, VUTECO employs a *decision function* after the *Finder* made its prediction on the test case—depicted on the right-most side of Figure 2. Indeed, if the probability of the test case belonging to the “Security” class was at least 0.5, then the *Linker* model is invoked, and its probabilities are used for the “Matched” and “Not-Matched” classes. Otherwise, the probabilities of the *Finder* are used instead. Simply put, the *Linker*’s judgment is not considered if the *Finder* model did not flag the test case as “Security” as the *Linker* model expects security-related tests as input. Hereafter, the wording “integrated model” refers to the joint use of *Finder* and *Linker* to address the *Matching* task.

### III. EXPERIMENTAL DESIGN

Section II described VUTECO in its best setting, based on two evaluations aimed at understanding whether VUTECO successfully addressed the *Finding* and *Matching* tasks. The first one analyzed how VUTECO performed in an *in-vitro* setting, i.e., by testing it on a held-out set originating from VUL4J, the same source also used for its training and development. The second one is a complementary analysis that looks into its performance in an *in-vivo* scenario, i.e., by selecting the best model resulting from the *in-vitro* evaluation and running it on a set of projects outside VUL4J.

**Q RQ<sub>1</sub>.** How does VUTECO perform in *finding* security-related test cases?

**Q RQ<sub>2</sub>.** How does VUTECO perform in *matching* test cases and vulnerabilities?

#### A. In-vitro Evaluation Design

During the in-vitro evaluation, we explored the role of certain factors, such as training data augmentation, to observe how they affected the final performance, measured in the standard way for binary classification tasks. VUTECO’s architecture and training slightly vary depending on the task, as illustrated in Figure 2. For the *Finding* task, the *Finder* model was trained and tested using a collection of test cases extracted from the projects in VUL4J. On the other hand, the *Matching* task required three training sessions: one for the *Finder* model for its *Finding* task, one for the *Linker* model for its simplified task (see Section II-C), and an additional one integrating the trained *Finder* and *Linker* for the *Matching* task. We remark that evaluating the *Linker* model on its simplified task only aims at assessing its suitability for the integrated model, which is the one addressing the *Matching* task.

1) *Data Selection:* The primary source of data for the in-vitro evaluations was VUL4J [18], as it is the only known source having JUNIT test cases matched with the vulnerabilities they are witnessing. At the time of this paper writing, VUL4J had 51 JAVA projects, of which 108 JUNIT test cases

marked as witnessing tests. From such projects, we checked out their patched revision (i.e., the project version in which the vulnerability has been fixed and where the witnessing tests have been found) and mined 62,527 more test cases from the same projects using the heuristic in Section II-A to have examples of tests not witnessing any vulnerability. We fetched the metadata from the 79 vulnerabilities in VUL4J, such as the summary description and the weakness type, using the CVE Search API [32]. We ended up with 76 vulnerabilities with metadata, as three were reported via internal bug reports.

For the *Finding* task, we selected all the JUNIT test cases in VUL4J, ignoring any vulnerability metadata. Hence, all 108 tests have been labeled as “Security” class (the positive class), while the remaining 62,527 tests have been labeled as “Not-Security” class (the negative class). Then, we split this dataset using stratified sampling (i.e., keeping the same class distribution), creating a training set  $F_{TR}$  (70%), a development set  $F_{DE}$  (15%), and a test set  $F_{TE}$  (15%).

For the *Matching* task, two more datasets were needed, one for the *Linker* model and one for the integrated model. For the former, we created pairs of vulnerability-witnessing tests and the metadata of the linked vulnerability in VUL4J. Since three witnessing tests were matched to the three vulnerabilities without metadata, we ended up with 105 valid pairs, marked as “Linked”. Then, we paired the 105 tests with all the metadata of the vulnerabilities they do not witness, creating 7,665 invalid links to form the “Not-Linked” class. Then, we split this dataset with stratified sampling, creating a training set  $L_{TR}$  (70%), a development set  $L_{DE}$  (15%), and a test set  $L_{TE}$  (15%). The dataset for the integrated model followed a similar construction. The 105 valid pairs of the *Linker* model also formed the positive class “Matched”. However, in this case, the negative class “Not-Matched” was made by pairing all test cases from VUL4J projects (not just the witnessing tests) with the metadata of unrelated vulnerabilities affecting the same project, forming 85,305 invalid pairs. Again, we split this dataset with stratified sampling, creating a training set  $I_{TR}$  (70%), a development set  $I_{DE}$  (15%), and a test set  $I_{TE}$  (15%). When testing the integrated model for the *Matching* task, we removed from  $F_{TR}$  and  $L_{TR}$  any test case and vulnerability appearing in  $I_{TE}$  to avoid data leakages that could influence the evaluation results.

2) *Performance Indicators:* We relied on traditional metrics to measure the goodness of binary predictions, i.e., *precision* (Pr), *recall* (Re), *F1 score* [33], [34], the Area Under the Receiver Operator Characteristic curve (AUC-ROC) [35], and report the absolute number of positive classifications (i.e., True Positives and False Positives) to give them more context to the results. We argue that the main goal of VUTECO is to **maximize the number of correct findings and minimize the incorrect ones**, with the purpose of expanding the current knowledge base of witnessing tests to address the challenges described in Section I. Given this circumstance, we include the *F0.5 score* in the analysis, which is a varied version of the F1 score where twice as much weight is given to precision compared to recall [36]. During the analysis of the

TABLE I  
FACTORS FORMING THE EXPERIMENTED CONFIGURATIONS.

Factor	Meaning	Experimented Values
<b>FINDER (4 CONFIGURATIONS)</b>		
<i>F-Aug</i>	Data augmentation technique to apply on $I_{TR}$ .	JT, SPAT, BS, None
<b>LINKER (8 CONFIGURATIONS)</b>		
<i>L-CWE</i>	The presence of the CWE in the textual input, if available.	CWE-Yes, CWE-No
<i>L-Aug</i>	Data augmentation technique to apply on $I_{TR}$ .	JT, SPAT, BS, None
<b>INTEGRATED (13 CONFIGURATIONS)</b>		
<i>I-Mode</i>	The way the integration of <i>Finder</i> and <i>Linker</i> happens.	Linker-Only, FT-Only, PT-Only*, PT-FT
<i>I-Aug</i>	Data augmentation technique to apply on $I_{TR}$ .	JT, SPAT, BS, None

\* Ignores the value of *I-Aug*, as this mode does not use  $I_{TR}$ .

performance at test time, we give more attention to the F0.5 score as it captures the trade-off between precision and recall that aligns with the VUTECO’s purpose.

3) *Configurations Experimented*: A configuration of a model is made of one or more **factors**, i.e., aspects we hypothesized could notably affect the performance, e.g., the algorithm used to augment the training set. Every time we had to train a model (*Finder*, *Linker*, and both together), we did it with different configurations and compared their performance on the related test sets to find the most suitable one. The set of factors involved for each model type is in Table I.

All three models handled the class imbalance of their training sets by augmenting the data through *F-Aug*, *L-Aug*, and *I-Aug* factors. We experimented with three different mechanisms: (1) **JAVATRANSFORMER** (JT) [37], a JAVA command-line tool that transforms a JAVA method (including test methods) into semantically-equivalent clones by applying nine semantic-preserving transformation rules, like renaming variables or add random logging statements; (2) **SPAT** [38], similar to JAVATRANSFORMER, it creates semantically-equivalent clones of JAVA methods with 18 transformation rules; (3) **Bootstrapping** (BS), a resampling technique creating exact copies of instances of the minority class, a.k.a. random oversampling. We also experimented with a fourth case in which the training data were not augmented.

The *Linker* model accepts a vulnerability as input, whose metadata typically consists of two elements, i.e., the summary description and an assigned weakness type, identified using the CWE (Common Weakness Enumeration). The former conveys most of the information about the vulnerability and is present in all valid CVE records, while the latter might be missing or incorrectly assigned [39]. We experimented with two scenarios via the *L-CWE* factor, i.e., whether to prepend the CWE identifier and name (when available) before the summary description or not use it at all.

To keep the number of experimented configurations reasonable, the integrated model reused the best configurations of *Finder* and *Linker* models observed in their independent training and testing sessions. However, when used together for the *Matching* task, they were re-trained again on slightly

modified training sets (see Section III-A1). In a way, this can be seen as a *pre-training* session that we hypothesized could benefit the performance of the *Matching* task. In this regard, we experimented with several ways of integrating the two models via the *I-Mode* factor: (1) fine-tuning on  $I_{TR}$  using only the *Linker* model without prior pre-training (**Linker-Only**); (2) fine-tuning on  $I_{TR}$  using *Finder* and *Linker* together without prior pre-training (**FT-Only**); (3) using *Finder* and *Linker* together with prior pre-training but no fine-tuning on  $I_{TR}$  (**PT-Only**); (4) using *Finder* and *Linker* together with prior pre-training and fine-tuning on  $I_{TR}$  (**PT-FT**). In PT-Only mode, the factor *I-Aug* is ignored as it does not involve  $I_{TR}$ , resulting in just one configuration with this mode.

In summary, we tested four *Finder* configurations, eight *Linker* configurations, and 13 integration configurations, trained for eight, 10, and eight epochs, respectively. All training sessions used a dropout layer with 0.1 probability between every linear layer added after the CodeBERT to reduce the risk of overfitting [40]. The weights were updated using the AdamW optimizer [41], with  $10^{-5}$  learning rate decaying linearly, and the loss was measured using the cross-entropy function [42]. At the end of each epoch, we evaluated the model checkpoint on the related development set, returning the one with the highest AUC-ROC and lowest loss (in case of ties). The models have been implemented with PYTORCH and trained with the HUGGINGFACE API. The evaluation has been carried out on a server with an NVIDIA Tesla A100 Core GPU and an Intel Xeon Platinum 8352V CPU.

4) *Hyper-parameter Tuning*: In addition to the factors, we selected other secondary aspects that might have some impact on the performance, such as the size of the hidden layers of a model’s classification head. We considered such aspects as **hyper-parameters** that we optimized during the model’s training; namely, for each configuration, we searched the best hyper-parameter combination by comparing the models on the performance on the development set, selecting the one having the highest AUC-ROC and lowest loss. During the analysis of the in-vitro results for both tasks (Section IV), we only report the performance on the test set, leaving the full report in the paper’s online appendix [27].

All three models use training data augmentation, which can happen to different “extents”; we call this hyper-parameter *E*. JAVATRANSFORMER (JT) and SPAT can be run multiple times to generate further semantically equivalent clones as they are driven by certain randomness elements, e.g., to decide the new variable names. Bootstrapping (BS), instead, can generate instances of the minority class until a new imbalance ratio is reached; for example, 0.25 means that the minority instances should be 25% of the total. Hence, for JT and SPAT, we set  $E=\{5, 15, 25\}$  (discarding any duplicate instances), while for BS, we set  $E=\{0.25, 0.50, 0.75\}$ . When no data augmentation is done (*None*), this hyper-parameter is ignored. Besides, for the *Finder* and *Linker* models, we searched for the right size of the two hidden layers before the output layer, using  $H_1=\{768, 512, 256\}$  and  $H_2=\{256, 128, 64, 0\}$ , respectively.

5) *Baseline Approaches*: To the best of our knowledge, VUTECO is the first solution targeting the problem of matching tests and vulnerabilities; thus, no other approaches can be used for a direct comparison. Nevertheless, we developed some *baseline approaches* for each model tested, relying on mechanisms more lightweight than those used in VUTECO.

A baseline for the *Finder* model should extract facts from the test cases using a different principle than those used by the *Finder* model. We developed a heuristic algorithm that extracts *keywords* from the test case and checks (via case-insensitive match) if these appear in a vocabulary of keywords fitted on the test cases in the training set  $F_{TR}$ . If the number of keywords found in the vocabulary surpasses an arbitrary threshold  $N$ , the test case is flagged as “Security”. We extracted keywords in two ways: (i) using YAKE [43], which extracts the  $K$  most relevant keywords from a text in an unsupervised manner, and (ii) using a custom script to retrieve all the identifiers (i.e., variable and method names) in the test case, as they likely indicate the purpose of the test and are not mixed with other irrelevant programming keywords. We call these two flavors of the approach *YAKE-Vocab* and *Identifier-Vocab*, respectively. We experimented with  $N=[1..10]$  for both baselines; for *YAKE-Vocab* we experimented with  $K=\{5, 10, 15, 20, 25, 30\}$ ; while for *Identifier-Vocab* we experiment the honoring `camelCase` and `snake_case` notations, e.g., if identifiers like `user_Password` must be split into `user` and `password` or not. In total, 80 variants have been evaluated; in Section IV, we present only the best variants (in terms of F0.5 score) for both flavors.

A baseline for the *Linker* model should check whether there is a connection between the test case and the vulnerability description. Hence, we developed a heuristic algorithm that checks the *similarity* between the two textual inputs. If the similarity surpasses an arbitrary threshold  $S$ , the pair is flagged as “Linked”. We made two different implementations: (i) extracting the keyword sets from both text inputs using YAKE [43] and comparing them using the Jaccard index; (ii) extracting the embeddings from both text inputs using a pre-trained CodeBERT model [26] and comparing them using the cosine similarity. We call these two flavors of the approach *YAKE-Simil* and *CodeBERT-Simil*, respectively. For the former, we experimented with  $S=\{0.01, 0.02, 0.03, 0.04, 0.05, 0.1\}$  and  $K=\{5, 10, 15, 20, 25, 30\}$ , while for the latter we experimented with  $S=\{0.9, 0.95, 0.96, 0.97, 0.98, 0.99\}$ . In total, 42 variants have been evaluated; in Section IV, we present only the best variants (in terms of F0.5 score) for both flavors.

Regarding the baseline for the integrated model, we followed a different principle by employing an approach called *FixCommits*, which relies on the assumption that developers create unit tests alongside the patches to show that the vulnerability was fixed successfully. *FixCommits* does not look at the vulnerability description but rather inspects all fix commits of a vulnerability to gather the set of test cases  $TM$  added or modified and flags that vulnerability and all the tests in  $TM$  as “Matched” pairs. Any other pair is automatically considered as “Not-Matched”. We observe that there is no sufficient

evidence that its core assumption generally holds cases as developers might not create any test when fixing vulnerabilities. For example, in CVE-2010-0684 of ACTIVEMQ none of the three fix commits (2895197, fed39c3, and 9dc43f3) added any test. For this, *FixCommits* is to be considered as a heuristic approach rather than a ground truth.

### B. In-vivo Evaluation Design

During the in-vivo evaluation, we assessed the *precision* of the results obtained by VUTECO for both tasks due to the lack of ground truth—indeed, this evaluation aims at assessing its usefulness *in the wild*. Namely, the *Finding* task returns a list of test cases that VUTECO believes to be security-related, while the *Matching* task returns a list of pairs of test cases matched with their vulnerabilities. We prepared VUTECO for both tasks using the best configurations resulting from the respective in-vitro evaluation (partially revealed in Section II). This consisted of re-training the models used in each task by merging the content of the test sets in the related training sets (as there was no use for it anymore). The rest of the training followed the setting described in Section III-A3.

We selected open-source JAVA projects and their vulnerabilities from the latest version of PROJECTKB [22] available in its GITHUB repository [44]. First, we excluded the vulnerabilities in VUL4J, as they could have been used to train VUTECO—and so biasing the final results—and also since we were not interested in their witnessing tests. With this step, we discarded 60 vulnerabilities. We discarded 11 more vulnerabilities whose metadata could not be found in PROJECTKB or be retrieved with CVESEARCH API. Then, we assessed their projects’ accessibility, discarding those without a remote URL or no longer accessible, which determined the removal of 92 more vulnerabilities. Since we had no guidance on when the witnessing tests could have been added, we selected the latest revision of each project (i.e., the HEAD of the respective base branches) on June 13, 2024. We believe witnessing tests addressing past vulnerabilities can still be found in current test suites. Therefore, the input given to VUTECO for both tasks was a project’s repository (remote URL) and its latest revision. Besides, for the *Matching* task, VUTECO also requires the list of historical vulnerabilities that affected it (via PROJECTKB). At this point, we applied the test selection criteria in Section II-A to ensure the projects had valid test suites. This revealed 324 projects without any test that VUTECO could process; thus, we discarded the related 486 vulnerabilities. After all these steps, we ended up with 244 projects and 640 vulnerabilities. The 244 projects comprised a total of 823,529 test cases, all given to VUTECO during the *Finding* task as input. For the *Matching* task, instead, we first paired all the test cases with only the vulnerabilities affecting the related project, and then VUTECO processed each pair one by one.

To assess the validity of the results, we relied on two independent researchers with experience in software security and testing and familiarity with VUL4J. They autonomously inspected all the findings with a predicted probability of belonging to the positive classes (of the respective tasks) of

at least 0.5. They were instructed to mark the correct and incorrect ones, allowing the computation of true and false positive predictions and, therefore, the precision. We relied on Cohen’s Kappa score [45], [46] to measure agreement between the two raters; all cases of disagreement were jointly discussed and solved until a full agreement was reached.

#### IV. EXPERIMENTAL RESULTS

##### A. Finding Task Results

Table II reports the result of the four experimented configurations of the *Finder* model along with the best variants of the two flavors of *Vocab* technique. The best configuration achieved a 0.83 F0.5 score and perfect precision without augmenting the training set (highlighted in yellow). It also achieved the highest F1 score of 0.67 and a very high AUC-ROC—only surpassed by the configuration trained with SPAT [38] with 0.03 margin.

From a broader perspective, we observed that the data augmentation mechanism had a negative effect on increasing the number of false positives—while keeping the number of true positives unaffected. Indeed, improper data augmentation (e.g., bootstrapping) can decrease precision by up to 47%, introducing several false positive classifications for at most one extra true positive. At the same time, we observed that all configurations could not score a recall higher than 0.56—achieved by the least precise configuration. The best configuration, indeed, could reach 0.50. This is the only metric in which a baseline technique, i.e., *Identifier-Vocab*, (slightly) outperforms the best configuration. However, this was only due to the one extra true positive classification—just as in the bootstrapping-augmented variant—at the cost of 288 false positives, making this baseline unsuitable for this task.

The high F0.5 score (due to the perfect precision) observed in the *None* variant allows the *Finder* model to be the right tool for finding candidate vulnerability-witnessing tests. Thus, we re-trained the *Finder* model with this configuration and ran it on the set of JAVA projects in the wild to find security-relevant test cases. The model returned 145 test cases from 40 projects (16% of the analyzed ones). The inspectors then validated all those findings (spending two minutes per entry on average), agreeing on 134 (92%) cases, obtaining 0.82 Kappa score, indicating *quasi-perfect* inter-rater agreement. Afterward, they jointly inspected the 11 remaining cases where they had conflicting judgments until reaching a common decision. At the end of the inspection process, the number of valid test cases was 102 out of 145 (70%).

This caused VUTECO to score 0.7 precision in the wild, which is not too far from the perfect precision achieved during the in-vitro validation. In fact, this result must be contextualized with the much greater set of test cases VUTECO analyzed, which was more than 18 times bigger than the held-out set extracted from VUL4J (823,529 tests vs. 43,844), inevitably leading to a higher chance of false positives. Zooming into the incorrect findings, we observed that the *Finder* model was fooled by test cases containing certain keywords related to vulnerabilities but did not actually test any security

TABLE II  
PERFORMANCE OF THE FOUR *Finder*’S CONFIGURATIONS (OPTIMAL HYPER-PARAMETERS) AND *Vocab* BASELINES (BEST VARIANTS).

Configuration		Performance						
	F-Aug	Pr	Re	F1	F0.5	AUC	TP	FP
<i>Finder</i>	JT	0.67	0.50	0.57	0.63	0.93	8	4
	SPAT	0.73	0.50	0.59	0.67	<b>0.96</b>	8	3
	BS	0.53	<b>0.56</b>	0.55	0.54	0.92	9	8
	None	<b>1.00</b>	0.50	<b>0.67</b>	<b>0.83</b>	0.93	8	0
<i>YAKE-Vocab</i>	0.11	0.44	0.18	0.13	N/A	4	199	
<i>Identifier-Vocab</i>	0.03	<b>0.56</b>	0.06	0.04	N/A	9	288	

issues. For example, in SONARQUBE, VUTECO flagged the test case `write_cve` as security-related [47]; however, the test checks whether the scanner writes certain information about a CVE in the report, which is definitely not a security issue affecting SONARQUBE itself. Hence, the *Finder* model failed to interpret the context in which the security-related terms, like `cve` or `DoS`, have been used. This problem could be addressed by introducing fine-grained semantic analyses and better keyword matching, which would greatly reduce the number of false positives.

**Answer to RQ<sub>1</sub>.** In the *Finding* task, VUTECO achieved perfect precision and 0.83 F0.5 score when tested on a held-out set from VUL4J without the need for augmenting its training data. The *Finder* model is generally conservative, favoring precision more easily than recall; an improper data augmentation gives up too much precision for a marginal, often null, increase in recall. VUTECO found 102 security-related test cases in the wild, i.e., 70% of the total. Most of the false positives were induced by some security-related terms appearing in the test code.

##### B. Matching Task Results

Table III reports the result of the eight experimented configurations of the *Linker* model along with the best variants of the two flavors of *Simil* technique. The best configuration was the one that augmented the training set with JAVATRANSFORMER [37] and did not use the CWE information to describe the vulnerability, but only the summary description, achieving a 0.57 F0.5 score and 0.71 precision (highlighted in yellow). Despite having a lower F0.5 score than the highest achieved by the *Finder* model in the *Finding* task, the precision value was still in an acceptable range. The factor that lowered the F0.5 was the recall of 0.31. Despite this, this configuration achieved a very high AUC-ROC score of 0.89—a trait shared with all the other configurations except one.

Similarly to what happened with the *Finder* model, the training data augmentation had noticeable effects in terms of precision, observing that JAVATRANSFORMER always led to an increase. On the other hand, the presence of CWE did not cause noteworthy differences in all metrics, making the two groups comparable; yet, we cannot conclude the presence of absence is either beneficial or detrimental. In any case, we observed that the recall was difficult to increase, just like

TABLE III

PERFORMANCE OF THE EIGHT *Linker*'S CONFIGURATIONS (OPTIMAL HYPER-PARAMETERS) AND *Simil* BASELINES (BEST VARIANTS).

Configuration		Performance							
		Pr	Re	F1	F0.5	AUC	TP	FP	
	L-CWE	L-Aug							
<i>Linker</i>	CWE-Yes	JT	<b>0.67</b>	0.25	0.36	0.50	0.84	4	2
		SPAT	0.63	0.31	<b>0.42</b>	<b>0.52</b>	0.55	5	3
		BS	0.33	0.31	0.32	0.33	0.88	5	10
		None	0.44	0.25	0.32	0.39	0.82	4	5
	CWE-No	JT	<b>0.63</b>	<b>0.31</b>	<b>0.42</b>	<b>0.52</b>	<b>0.89</b>	5	3
		SPAT	0.50	0.25	0.33	0.42	0.77	4	4
		BS	0.50	0.25	0.33	0.42	0.82	4	4
		None	0.60	0.19	0.29	0.42	0.81	3	2
<i>YAKE-Simil</i>		0.25	0.19	0.21	0.23	N/A	3	9	
<i>CodeBERT-Simil</i>		0.01	<b>1.00</b>	0.03	0.02	N/A	16	1,150	

the *Finder* model. Between the two flavors of *Simil*, only *YAKE-Simil* managed to have some true positives without making many false positive classifications. Still, all *Linker*'s configurations behaved better than both baselines. On the contrary, the best variant of the *CodeBERT-Simil* approach ended up behaving like a constant classifier, flagging class "Linked" in almost all the cases.

The lower performance of the *Linker* could be ascribed to the implications of its simplified task, which limited the diversity in its training set (especially when compared to the *Finder*'s training set, seen in Section III-A). Considering the difficulty of learning relationships between two textual inputs, the *Linker* model could only marginally understand the task. Nevertheless, the sufficient precision and the high AUC-ROC score let us consider the *Linker* model ready for integration with the *Finder* model to address the *Matching* task.

Table IV reports the result of the 13 experimented configurations of the integrated model along with the *FixCommits* technique. The best configuration was the one that ran a pre-training of both *Finder* and *Linker* models in their specific tasks (using their best configurations seen in Tables II and III) and then ran a fine-tuning for the *Matching* task without augmentation (highlighted in yellow in Table IV). This configuration achieved a high precision of 0.86, a good F0.5 score of 0.68, and a greatly high AUC-ROC score of 0.91—surpassed by two other configurations by a small margin.

The main factor that affected the performance of the integrated model was *I-Mode*, indicating how the *Matching* task was addressed through the *Finder* and *Linker* models. The setting that used the *Linker* model directly for the *Matching* task (i.e., without the *Finder*'s support) achieved 0.70 precision and 0.63 F0.5 score without augmented training data. Such results are better than the best *Linker* model on the simplified task. After switching to the joint use of *Finder* and *Linker* without pre-training (*FT-Only*), we observed an interesting fact. When the training data was augmented, there were no noteworthy differences; however, the configuration without data augmentation made no positive classifications at all. At this point, we employed the pre-training for both models independently but removed the fine-tuning (*PT-Only*), obtaining a noticeable improvement in precision, reaching 0.80, though with a drop in recall (0.25). Lastly, we introduced the fine-

TABLE IV

PERFORMANCE OF THE 13 INTEGRATION CONFIGURATIONS (OPTIMAL HYPER-PARAMETERS) AND *FixCommits* BASELINE.

Configuration		Performance							
		Pr	Re	F1	F0.5	AUC	TP	FP	
	I-Mode	I-Aug							
<i>Integrated</i>	Linker-Only	JT	0.26	0.38	0.31	0.28	0.90	6	17
		SPAT	0.31	0.31	0.31	0.31	0.86	5	11
		BS	0.24	0.44	0.31	0.27	0.91	7	22
		None	0.70	0.44	<b>0.54</b>	0.63	<b>0.94</b>	7	3
	FT-Only	JT	0.36	0.31	0.33	0.35	0.87	5	9
		SPAT	0.31	0.31	0.31	0.31	0.88	5	11
		BS	0.38	0.50	0.43	0.40	0.92	8	13
		None	N/A	0.00	N/A	0.00	0.64	0	0
	PT-Only	/	0.80	0.25	0.38	0.56	0.91	4	1
		JT	0.58	0.44	0.50	0.55	0.89	7	5
		SPAT	0.60	0.19	0.29	0.42	0.90	3	2
		BS	0.19	0.44	0.26	0.21	0.87	7	30
	PT-FT	None	<b>0.86</b>	0.38	0.52	<b>0.68</b>	0.91	6	1
		<i>FixCommits</i>	0.53	<b>0.63</b>	<b>0.57</b>	0.54	N/A	10	9

tuning back (*PT-FT*), achieving a 7.5% and 52% boost in precision and recall, respectively, when no augmentation is used. This confirms the benefit of employing two models for the *Matching* task with a two-stage training.

Just like the *Finder* and *Linker* models, the recall score was generally low; this further supports the fact that balancing precision and recall is challenging in all the tasks addressed in this work. The heuristic approach *FixCommits*, on the other hand, achieved a higher recall of 0.63. Nevertheless, its precision and F0.5 were still lower than those achieved by the best integrated model. Indeed, VUTECO achieved 62% and 26% more precision and F0.5, respectively, than *FixCommits*. Given our preference for precise solutions, the integrated model of VUTECO was preferred over the *FixCommits* approach.

At this point, we re-trained the integrated model with the best configuration and ran it on the set of JAVA projects in the wild to match the test cases with the vulnerabilities they witnessed. The model returned 159 matches from 10 projects (6% of the analyzed ones). The inspectors then validated all those findings (spending three minutes per entry on average), agreeing on 152 (96%) cases. Afterward, they jointly inspected the seven remaining cases where they had conflicting judgments until reaching a common decision. At the end of the inspection process, the inspector affirmed that none of the matches were valid—due to this, the inspectors never agreed on the presence of valid matches, making the resulting Kappa score insignificant despite the two inspectors agreeing in almost all cases.

The lack of valid matches did not meet the expectations set during the in-vitro validation. Unfortunately, VUTECO is not ready yet for matching test cases and vulnerabilities in the wild. Nevertheless, we recognized that this task is challenging, apparently much more than the one on VUL4J—despite both contexts comprising real-world JAVA projects. Hence, we shed more light on these results by asking the two inspectors to make a fine-grained inspection of all the matches to see whether the test cases were, at least, security tests and whether they were matched with a similar vulnerability. They found that all but one test cases were concerned with security issues. This accessory behavior provides an alternative

way to find security-related tests with a different principle than the one used by the *Finder* model alone. Besides, in *nine* cases, we found that the matched vulnerabilities are similar to the correct ones—in terms of vulnerability type or description. For instance, in DROPWIZARD, the test case `shouldNotBeVulnerableToCVE_2022_42889` was matched with CVE-2020-5245, which is an injection vulnerability that opens to remote executions akin to CVE-2022-42889 (the intended one) [48]. Hence, the integrated model demonstrated signs that the learning stage had some benefits, though insufficient. Better modeling, as well as more examples of real matches, are needed to improve the generalizability of the *Matching* task.

**Answer to RQ<sub>2</sub>.** In the *Matching* task, VUTECO achieved 0.86 precision and 0.68 F0.5 score when tested on a held-out set from VUL4J without the need for augmenting its training data. It benefited from the joint use of *Finder* and *Linker*, first pre-trained on their tasks and then further fine-tuned for the *Matching* task. Unfortunately, VUTECO failed to match test cases and vulnerabilities in the wild. Nevertheless, all tests involved in the matches were also security-related, and in some cases, the matched vulnerability was not too dissimilar to the exact one.

## V. DISCUSSION

**The Collection of Witnessing Tests.** The in-vivo analysis allowed us to observe how collecting vulnerability-witnessing tests is a *challenging task*. VUTECO achieved good results in the *Finding* task, though failing on the *Matching* task. We further reflected on the erroneous matches returned to find possible room for improvements. We believe the errors were due to the model *understanding the given vulnerability partially*, likely due to the limited context given by the summary description from CVE. Besides, the inspected test cases might contain *noisy elements* that are typical of the “security vocabulary,” despite them not being concerned about any security aspect (as seen in Section IV-A). A more curated pre-training and fine-tuning setting could be the initial action to mitigate such limitations. At the same time, we also reflected on the positive collateral behavior seen in the *Matching* task, where VUTECO successfully managed to exclude test cases not related to security. We hypothesize this could be an indirect effect of the longer training stage required to prepare it for the *Matching* task. Despite the in-vivo analysis being unable to indicate the true or false negative classifications, the extent of the results VUTECO returned—i.e., a few hundred test cases out of over 800,000—permit a confirmatory manual inspection to be carried out in a reasonable time. Thus, VUTECO can be a helpful support to reduce the search space greatly.

**The Usefulness of Witnessing Tests.** The release of the dataset VUL4J unlocked the research of numerous software security tasks. The ability of VUTECO to find security-related tests can help expand the known body of vulnerability-witnessing tests, enabling activities like the automated generation of security unit tests [15], [16]. Nevertheless, the

potential applications of vulnerability-witnessing tests extend far beyond this [49]. For example, the witnessing tests can act as *proofs-of-concept* supporting the automatic generation of realistic exploits, building on the advancements of security test generation models [15], [16]. Besides, they can support the automated vulnerability repair (AVR) process by localizing the vulnerable statements or assessing the plausibility of a generated patch [50]–[53]. We also envision the use of witnessing tests to support the retrieval of vulnerability-contributing commits [54], as they can be run to triangulate better the moment in which a vulnerability was introduced in a project; to the best of our knowledge, this task has not been investigated yet. Furthermore, software engineers can benefit from having many instances of witnessing tests and a tool like VUTECO supporting their retrieval. For instance, they can reuse known tests of past vulnerabilities to address similar issues affecting their projects. This scenario further motivates the need for VUTECO to perform better on this task.

**The Anatomy of Witnessing Tests.** The retrieval of witnessing tests is challenging mainly due to the lack of empirical knowledge on the matter. To date, no study outlined a clear profile of tests witnessing vulnerabilities or, more broadly, unit tests targeting security aspects. The *absence of characterization* of vulnerability tests entails a significant knowledge gap, especially when compared with traditional functional tests. For instance, we are unaware of what setup is needed before calling the vulnerable component, what the assertions should look at, or the number of tests required to “cover” all the relevant scenarios concerning a vulnerability type. The only known common point between the two types of tests is that they both aim to find undesired behaviors in the code that violate some requirements or properties. We believe this lack of knowledge might be attributable to the difficulty in formulating security requirements at the unit/component level (i.e., methods or classes) since they are often considered at the system level [13], [55]. Unfortunately, shedding light on these aspects still demands many examples of witnessing tests. In fact, this study was created to address this shortage, providing an approach to expanding the knowledge base of witnessing tests and drawing more attention to this topic. Once a line between vulnerability-witnessing tests and traditional tests is drawn, innovative solutions can be designed to support developers in writing more security tests.

## VI. THREATS TO VALIDITY

For the *Matching* task, we considered the performance of the *Finder* and *Linker* models in their tasks independently before integrating them. Yet, there was no guarantee that the selected configurations would work well once integrated. We mitigated this threat by running a new train-test session to assess the whole integrated model for the actual *Matching* task. The final answer to the in-vitro analysis on RQ<sub>2</sub> is only given by the results achieved by the integrated model.

Due to computational constraints, we could not thoroughly analyze the impact of many design aspects in the in-vitro

analyses. We treated some aspects as factors, like data augmentation, to be evaluated at test time, while others were treated as hyper-parameters to be optimized at development time.

For the *Matching* task, VUTECO accepts a natural language description of the vulnerability. Hence, VUTECO accepts any text describing the issue, also from other sources like bug reports [54], [56]–[59]. We opted to experiment with CVE, as it is the most comprehensive catalog for obtaining such information and can be easily queried.

VUTECO has been trained and tested on the JUNIT test cases in VUL4J. Therefore, all the results observed cannot be generalized to other programming languages, other testing frameworks (like TESTNG), or vulnerability types that are not found in JAVA code, like memory-related vulnerabilities. We chose JAVA mainly due to the availability of VUL4J containing examples of witnessing tests to train and validate VUTECO. At the same time, we remark that JAVA is still a relevant language to analyze from a security perspective as it keeps exhibiting vulnerabilities for a long time [60].

Despite being the only catalog of its kind, the limited number of witnessing tests in VUL4J does not allow models to learn many patterns that can be generalized to different contexts. It also lacks diversified examples, e.g., it has eight examples of tests witnessing CWE-835 (Infinite Loop), but just one witnessing CWE-78 (OS Command Injection) [18]. We partially handled this by trying to augment the training sets involved during the experimentation, though without the hoped effect. VUTECO was born to contribute to feeding such knowledge bases with new examples, which in turn could be used to re-train VUTECO and return more results.

VUTECO flags witnessing tests based on a static approach: No test cases are executed, and no projects are built. For this, VUTECO cannot provide a definite answer about whether the tests will trigger the matched vulnerabilities. The dynamic assessment has been taken outside the scope of this work due to technical difficulties in building numerous projects in a reasonable time, as encountered by the authors of VUL4J [18].

## VII. RELATED WORK

### A. Test Case Classification

Fatima et al. [61] presented FLAKIFY, a data-driven approach to detect flaky tests in JAVA projects. FLAKIFY leverages a pre-trained CodeBERT and a feed-forward neural network to predict whether a JUNIT test method had a flaky behavior. Such architecture resembles the VUTECO’s *Finder* model. FLAKIFY achieved 0.79 and 0.73 F1 scores on two different experiments on FLAKEFLAGGER dataset [62], while achieving 0.98 and 0.89 F1 score on IDOFT dataset [63], outperforming state of the art approaches. Somewhat similarly, FLAKYCAT exploits few-shot learning to predict the exact category of flakiness of JUNIT tests [64]. FLAKYCAT relies on a pre-trained CodeBERT to create the embeddings of test cases and a Siamese Network to project the embeddings into a space where tests of the same flakiness category appear similar—according to the cosine similarity function. This is enacted by the Triplet Loss function [65] during the training.

The design of VUTECO, particularly the *Finder* model, has been inspired by such approaches due to the similarity of their tasks, particularly those made in FLAKIFY [61]. Both approaches rely only on the test code for making the classification without requiring access to the production code, running it, or using any human-defined features.

### B. Vulnerability-witnessing Tests

Kang et al. [15] introduced TRANSFER to generate security test cases for JAVA projects affected by vulnerable library dependencies. TRANSFER builds on existing vulnerability-witnessing tests mined from the upstream library project and tries to generate a test case targeting the client project that recreates the same program state generated by the execution of the witnessing test in the original library. This approach leverages a genetic algorithm, whose goal is to find a client test that “mimics” the behavior of the library test. TRANSFER successfully generated security tests for 14 known library vulnerabilities in 23 client projects. Later, TRANSFER was extended by Chen et al. [66] by adding a migration step, which helps guarantee the similarity between generated tests from client projects and the original vulnerability tests. Their tool, VESTA, outperforms TRANSFER by 53.4% in terms of the effectiveness of generating tests on a dataset of 30 JAVA vulnerabilities. Zhang et al. [16] adopted a similar idea but relied on CHATGPT-4 rather than a genetic algorithm. The library witnessing tests have been used as examples in the prompt to nudge the AI to generate a similar test but targeting a different production method (living in the client code rather than the library). The proposed approach generated 24 working tests from 55 tasks, encompassing 30 known vulnerabilities and at least one affected client project for each.

Antunes et al. [67] proposed an approach to recycle the payloads from existing test cases of the File Transfer Protocol (FTP) protocol, which were then used to generate new vulnerability test cases for other protocols or new features. The proposed tool achieved better performance in terms of vulnerability coverage compared to two fuzzers and discovered 25 new vulnerabilities in FTP servers.

## VIII. CONCLUSION

In this work, we presented VUTECO, an automated approach to collect vulnerability-witnessing tests in JAVA projects. The in-vitro analyses showed the VUTECO approach is feasible if properly configured; the in-vivo analyses revealed that it works well for the *Finding* task but not for the *Matching* task. VUTECO is the first solution explicitly targeting this problem, laying the foundation for future research on the matter. After the analyses, we envisioned several ways to tackle this problem more efficiently. For instance, the input given to VUTECO can be enriched with *additional contextual information*, like the production code (e.g., the vulnerable components code) or a generated natural language summary of the test cases. VUTECO’s design could also include those in FLAKYCAT [64], i.e., exploiting a *similarity-based classification mechanism* that might perform well when having limited

examples, as well as converting the integration of *Finder* and *Linker* models into a *Mixture of Expert*. Lastly, VUTECO could include an *automated dynamic assessment* of the found tests to ensure they witness the vulnerability as expected.

## ACKNOWLEDGMENT

This work was partially supported by EU-funded project Sec4AI4Sec (grant no. 101120393).

## REFERENCES

- [1] S. Frei, D. Schatzmann, B. Plattner, and B. Trammell, “Modeling the security ecosystem - the dynamics of (in)security,” in *Economics of Information Security and Privacy* (T. Moore, D. Pym, and C. Ioannidis, eds.), (Boston, MA), pp. 79–106, Springer US, 2010.
- [2] G. McGraw, *Software Security: Building Security in*. Addison-Wesley professional computing series, Addison-Wesley, 2006.
- [3] F. Li and V. Paxson, “A large-scale empirical study of security patches,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’17, (New York, NY, USA), p. 2201–2215, Association for Computing Machinery, 2017.
- [4] S. Lipp, S. Banescu, and A. Pretschner, “An empirical study on the effectiveness of static c code analyzers for vulnerability detection,” in *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2022, (New York, NY, USA), p. 544–555, Association for Computing Machinery, 2022.
- [5] A. Austin, C. Holmgreen, and L. Williams, “A comparison of the efficiency and effectiveness of vulnerability discovery techniques,” *Information and Software Technology*, vol. 55, no. 7, pp. 1279–1288, 2013.
- [6] H. Shahriar and M. Zulkernine, “Mitigating program security vulnerabilities: Approaches and challenges,” *ACM Comput. Surv.*, vol. 44, jun 2012.
- [7] A. Kaur and R. Nayyar, “A comparative study of static code analysis tools for vulnerability detection in c/c++ and java source code,” *Procedia Computer Science*, vol. 171, pp. 2023–2029, 2020. Third International Conference on Computing and Network Communications (CoCoNet’19).
- [8] S. Elder, N. Zahan, R. Shu, M. Metro, V. Kozarev, T. Menzies, and L. Williams, “Do i really need all this work to find vulnerabilities? an empirical case study comparing vulnerability detection techniques on a java application,” *Empirical Software Engineering*, vol. 27, no. 6, p. 154, 2022.
- [9] D. S. Cruzes, M. Felderer, T. D. Oyetoyan, M. Gander, and I. Pekaric, “How is security testing done in agile teams? a cross-case analysis of four software teams,” in *Agile Processes in Software Engineering and Extreme Programming*, (Cham), pp. 201–216, Springer International Publishing, 2017.
- [10] D. Gonzalez, P. P. Perez, and M. Mirakhori, “Barriers to shift-left security: The unique pain points of writing automated tests involving security controls,” in *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, ESEM ’21, Association for Computing Machinery, 2021.
- [11] F. Camilo, A. Meneely, and M. Nagappan, “Do bugs foreshadow vulnerabilities? a study of the chromium project,” in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pp. 269–279, 2015.
- [12] G. Canfora, A. Di Sorbo, S. Forootani, A. Pirozzi, and C. A. Visaggio, “Investigating the vulnerability fixing process in oss projects: Peculiarities and challenges,” *Computers & Security*, vol. 99, p. 102067, 2020.
- [13] M. Felderer, M. Büchler, M. Johns, A. D. Brucker, R. Breu, and A. Pretschner, “Chapter one - security testing: A survey,” vol. 101 of *Advances in Computers*, pp. 1–51, Elsevier, 2016.
- [14] M. Mohammadi, B. Chu, H. R. Lipford, and E. Murphy-Hill, “Automatic web security unit testing: XSS vulnerability detection,” in *Proceedings of the 11th International Workshop on Automation of Software Test*, AST ’16, (New York, NY, USA), p. 78–84, Association for Computing Machinery, 2016.
- [15] H. J. Kang, T. G. Nguyen, B. Le, C. S. Păsăreanu, and D. Lo, “Test mimicry to assess the exploitability of library vulnerabilities,” in *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2022, (New York, NY, USA), p. 276–288, Association for Computing Machinery, 2022.
- [16] Y. Zhang, W. Song, Z. Ji, Danfeng, Yao, and N. Meng, “How well does llm generate security tests?,” 2023.
- [17] E. Pinconschi, R. Abreu, and P. Adão, “A comparative study of automatic program repair techniques for security vulnerabilities,” in *2021 IEEE 32nd international symposium on software reliability engineering (ISSRE)*, pp. 196–207, IEEE, 2021.
- [18] Q.-C. Bui, R. Scandariato, and N. E. D. Ferreyra, “Vul4j: a dataset of reproducible java vulnerabilities geared towards the study of program repair techniques,” in *Proceedings of the 19th International Conference on Mining Software Repositories*, MSR ’22, (New York, NY, USA), p. 464–468, Association for Computing Machinery, 2022.
- [19] G. Bhandari, A. Naseer, and L. Moonen, “Cvefixes: automated collection of vulnerabilities and their fixes from open-source software,” in *Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE 2021, (New York, NY, USA), p. 30–39, Association for Computing Machinery, 2021.
- [20] J. Fan, Y. Li, S. Wang, and T. N. Nguyen, “A c/c++ code vulnerability dataset with code changes and cve summaries,” in *Proceedings of the 17th International Conference on Mining Software Repositories*, MSR ’20, (New York, NY, USA), p. 508–512, Association for Computing Machinery, 2020.
- [21] G. Nikitopoulos, K. Dritsa, P. Louridas, and D. Mitropoulos, “Crossvul: a cross-language vulnerability dataset with commit data,” in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, (New York, NY, USA), p. 1565–1569, Association for Computing Machinery, 2021.
- [22] S. E. Ponta, H. Plate, A. Sabetta, M. Bezzu, and C. Dangremont, “A manually-curated dataset of fixes to vulnerabilities of open-source software,” in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pp. 383–387, 2019.
- [23] K. Liu, A. Koyuncu, D. Kim, and T. F. Bissyandé, “Tbar: Revisiting template-based automated program repair,” in *Proceedings of the 28th ACM SIGSOFT international symposium on software testing and analysis*, pp. 31–42, 2019.
- [24] M. Mohammadi, B. Chu, and H. R. Lipford, “Automated repair of cross-site scripting vulnerabilities through unit testing,” in *2019 IEEE International symposium on software reliability engineering workshops (ISSREW)*, pp. 370–377, IEEE, 2019.
- [25] X. Gao, B. Wang, G. J. Duck, R. Ji, Y. Xiong, and A. Roychoudhury, “Beyond tests: Program vulnerability repair via crash constraint extraction,” *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 30, no. 2, pp. 1–27, 2021.
- [26] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, and M. Zhou, “CodeBERT: A pre-trained model for programming and natural languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1536–1547, Association for Computational Linguistics, Nov. 2020.
- [27] Paper Authors, “Paper Online Appendix.” <https://figshare.com/s/44b014d85a5024358570>, 2024. Online.
- [28] Stefan Bechtold, Sam Brannen, Johannes Link, Matthias Merdes, Marc Philipp, Juliette de Rancourt, Christian Stein, “JUnit 5 User Guide.” <https://junit.org/junit5/docs/current/user-guide>, 2024. Online; accessed 9 July 2024.
- [29] Microsoft, “CodeBERT-based.” <https://huggingface.co/microsoft/codebert-base>, 2024. Online; accessed 9 July 2024.
- [30] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” 2016.
- [31] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” 2023.
- [32] Computer Incident Response Center Luxembourg (CIRCL), “CIRCL CVE Search.” <https://cve.circl.lu/>, 2024. Online; accessed 9 July 2024.
- [33] D. M. W. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [34] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press Books, ACM Press, 1999.
- [35] M. R. J. Junge and J. R. Dettori, “ROC solid: Receiver operator characteristic (ROC) curves as a foundation for better diagnostic tests,” *Global Spine J*, vol. 8, pp. 424–429, May 2018.

- [36] C. Van Rijsbergen, *Information Retrieval*. Butterworths, 1979.
- [37] M. R. I. Rabin, N. D. Bui, K. Wang, Y. Yu, L. Jiang, and M. A. Alipour, "On the generalizability of neural program models with respect to semantic-preserving program transformations," *Information and Software Technology*, vol. 135, p. 106552, 2021.
- [38] S. Yu, T. Wang, and J. Wang, "Data augmentation by program transformation," *Journal of Systems and Software*, vol. 190, p. 111304, 2022.
- [39] S. Pan, L. Bao, X. Xia, D. Lo, and S. Li, "Fine-grained commit-level vulnerability type prediction by cwe tree structure," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pp. 957–969, 2023.
- [40] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012.
- [41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.
- [42] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, (Red Hook, NY, USA), p. 8792–8802, Curran Associates Inc., 2018.
- [43] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.
- [44] SAP, "project-kb." <https://github.com/SAP/project-kb>, 2024. Online; accessed 9 July 2024.
- [45] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [46] M. McHugh, "Interrater reliability: The kappa statistic," *Biochimia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–82, Oct. 2012.
- [47] SonarQube, "Test case write\_cve in class ScannerReportWriterTest.java." <https://github.com/SonarSource/sonarqube/blob/ddd0e5a9/sonar-scanner-protocol/src/test/java/org/sonar/scanner/protocol/output/ScannerReportWriterTest.java#L149>, 2024. Online.
- [48] Dropwizard, "Test case shouldNotBeVulnerableToCVE\_2022\_42889 in class EnvironmentVariableSubstitutorTest.java." <https://github.com/dropwizard/dropwizard/blame/edaef8ce3de91cdada18c5767fa1e96cdcb2b04c/dropwizard-configuration/src/test/java/io/dropwizard/configuration/EnvironmentVariableSubstitutorTest.java#L73>, 2024. Online.
- [49] Z. Pan, X. Hu, X. Xia, X. Zhan, D. Lo, and X. Yang, "Ppt4j: Patch presence test for java binaries," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24*, (New York, NY, USA), Association for Computing Machinery, 2024.
- [50] M. Mohammadi, B. Chu, and H. Richter Lipford, "Automated repair of cross-site scripting vulnerabilities through unit testing," in *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pp. 370–377, 2019.
- [51] Z. Ságodi, G. Antal, B. Bogenfürst, M. Isztin, P. Hegedundefineds, and R. Ferenc, "Reality check: Assessing gpt-4 in fixing real-world software vulnerabilities," in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, EASE '24*, (New York, NY, USA), p. 252–261, Association for Computing Machinery, 2024.
- [52] X. Zhou, K. Kim, B. Xu, D. Han, and D. Lo, "Out of sight, out of mind: Better automatic vulnerability repair by broadening input ranges and sources," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24*, (New York, NY, USA), Association for Computing Machinery, 2024.
- [53] Q.-C. Bui, R. Paramitha, D.-L. Vu, F. Massacci, and R. Scandariato, "Apr4vul: an empirical study of automatic program repair techniques on real-world java vulnerabilities," *Empirical software engineering*, vol. 29, no. 1, p. 18, 2024.
- [54] L. Bao, X. Xia, A. E. Hassan, and X. Yang, "V-szz: automatic identification of version ranges affected by cve vulnerabilities," in *Proceedings of the 44th International Conference on Software Engineering*, pp. 2352–2364, 2022.
- [55] P. X. Mai, F. Pastore, A. Goknil, and L. C. Briand, "A natural language programming approach for requirements-based security testing," in *2018 IEEE 29th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 58–69, 2018.
- [56] Y. Zhou and A. Sharma, "Automated identification of security issues from commit messages and bug reports," in *Proceedings of the 2017 11th joint meeting on foundations of software engineering*, pp. 914–919, 2017.
- [57] T. H. M. Le, B. Sabir, and M. A. Babar, "Automated software vulnerability assessment with concept drift," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pp. 371–382, IEEE, 2019.
- [58] T. G. Nguyen, T. Le-Cong, H. J. Kang, R. Widjiasari, C. Yang, Z. Zhao, B. Xu, J. Zhou, X. Xia, A. E. Hassan, et al., "Multi-granularity detector for vulnerability fixes," *IEEE Transactions on Software Engineering*, vol. 49, no. 8, pp. 4035–4057, 2023.
- [59] E. Iannone, G. Sellitto, E. Iaccarino, F. Ferrucci, A. De Lucia, and F. Palomba, "Early and realistic exploitability prediction of just-disclosed software vulnerabilities: How reliable can it be?," *ACM Trans. Softw. Eng. Methodol.*, mar 2024. Just Accepted.
- [60] Veracode, "Annual Report on the State of Application Security." [https://info.veracode.com/rs/790-ZKW-291/images/Veracode\\_State\\_of\\_Software\\_Security\\_2023.pdf](https://info.veracode.com/rs/790-ZKW-291/images/Veracode_State_of_Software_Security_2023.pdf), 2023. Online; accessed 9 July 2024.
- [61] S. Fatima, T. A. Ghaleb, and L. Briand, "Flakify: A black-box, language model-based predictor for flaky tests," *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 1912–1927, 2023.
- [62] A. Alshammari, C. Morris, M. Hilton, and J. Bell, "Flakeflagger: Predicting flakiness without rerunning tests," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pp. 1572–1584, 2021.
- [63] W. Lam, R. Oei, A. Shi, D. Marinov, and T. Xie, "idflakes: A framework for detecting and partially classifying flaky tests," in *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*, pp. 312–322, 2019.
- [64] A. Akli, G. Haben, S. Habchi, M. Papadakis, and Y. Le Traon, "Flakycat: Predicting flaky tests categories using few-shot learning," in *2023 IEEE/ACM International Conference on Automation of Software Test (AST)*, pp. 140–151, 2023.
- [65] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2015.
- [66] Z. Chen, X. Hu, X. Xia, Y. Gao, T. Xu, D. Lo, and X. Yang, "Exploiting library vulnerability via migration based automating test generation," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24*, (New York, NY, USA), Association for Computing Machinery, 2024.
- [67] J. Antunes and N. Neves, "Recycling test cases to detect security vulnerabilities," in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*, pp. 231–240, 2012.