

A Study on Using Explainable Retrieval for Assisting Vulnerability Patch Tracing

Abstract—In recent years, the rapid growth of security vulnerabilities poses great challenges to tracing and managing them. For example, it was reported that the NVD database experienced significant delays due to the shortage of maintainers [1]. The delays in updating the patch link information can pose significant security risks, as organizations may remain vulnerable to known exploits before the patch link is updated. Furthermore, the delay creates challenges for third-party security personnel (e.g., administrators) to trace the information related to the CVE.

To help security personnel trace a vulnerability patch, in this work, we build a retrieval system that automatically retrieves the patch in the repository. While existing work leverages deep learning models such as CodeBERT for tracing the vulnerability fix, one major challenge with deep learning models is the lack of transparency. Furthermore, since the retrieved code is usually very long, it can be challenging for the security maintainer to understand the model output.

Inspired by existing work on explainable machine learning, we ask the following research question: can explanations help security maintainers make decisions in patch tracing? First, we investigate using LIME [2] (a widely used explainable machine learning method) to highlight the rationale tokens in the commit message and code. However, we observe that LIME often selects non-informative words, especially for the code. Therefore we propose an explanation method called TfIdf-Highlight, which leverages the Tf-Idf statistics to select the most informative words in the repository and the dataset.

We evaluate the effectiveness of highlighting using two experiments. First, we compare LIME and TfIdf-Highlight using a faithfulness score (i.e., sufficiency and comprehensiveness) defined for ranking [3]. Faithfulness is a standard metric for evaluating explainable machine learning. We find that TfIdf-Highlight significantly outperforms LIME’s sufficiency scores by 15% and slightly outperforms the comprehensiveness scores. Second, we conduct a blind human labeling experiment by asking the annotators to guess the patch under 3 settings (TfIdf-Highlight, LIME, and no highlight). We find that the helpfulness score for TfIdf-Highlight is higher than LIME while the labeling accuracies of LIME and TfIdf-Highlight are similar. Nevertheless, highlighting does not improve the accuracy over non-highlighting. The code and data are released in the following anonymous repository: <https://github.com/anonyMSR2025/anonyMSR2025>.

Index Terms—Security Vulnerability; Traceability; Open-Source Software; Explainable Machine Learning; Information Retrieval.

I. INTRODUCTION

How to define the efforts of code review: [4], accuracy during precision recall at fixed review efforts

In recent years, Open-Source Software (OSS) has been widely adopted, deployed, and utilized across industries. One of the biggest challenges that OSS faces is the prevalence of security vulnerabilities. For example, in 2017, Equifax suffered from a data breach, compromising the personal information

🔍 CVE-2013-1814 Detail

Description

The users/get program in the User **RPC** API in Apache Rave 0.11 through 0.20 allows remote authenticated users to obtain sensitive information about all user accounts via the offset parameter, as demonstrated by discovering password hashes in the password field of a response.

Hyperlink	Resource
http://www.openwall.com/lists/oss-security/2023/05/22/1	Mailing List Third Party Advisory
https://lists.apache.org/thread/7vwxonzwb7k9hx9[3q33cmj7j97jo3]	Mailing List Vendor Advisory
https://security.gentoo.org/glsa/202305-37	Third Party Advisory
https://security.netapp.com/advisory/ntap-20230616-0004/	Third Party Advisory
https://www.debian.org/security/2023/dsa-5521	Third Party Advisory

(a) An example of patch missing (Nov 2024) in the NVD database

Applied refactoring patch
 committed on Mar 6, 2013

```
diff --git a/rave-components/rave-web/src/main/java/org/apache/rave/portal/web/api/rpc/
PersonApi.java b/rave-components/rave-web/src/main/java/org/apache/rave/portal/web/
api/rpc/PersonApi.java
index 7d866a60..f1459d8c 100644
--- a/rave-components/rave-web/src/main/java/org/apache/rave/portal/web/api/rpc/
PersonApi.java
+++ b/rave-components/rave-web/src/main/java/org/apache/rave/portal/web/api/rpc/
PersonApi.java
@@ -18,21 +18,19 @@
 */
package org.apache.rave.portal.web.api.rpc;
```

(b) The relevant patch of CVE-2013-1814

Fig. 1: An example of NVD’s missing patch link

of 147 million users, causing millions of dollars of loss, and exposing individuals to potential identity theft [5]. The attack was exploited because Equifax did not patch an Apache Struts vulnerability whose patch was already available. As a result, it is important for OSS users to be alerted of the updates of vulnerabilities and apply the patches in time. For the same reason, if a patch is available, it is important for security advisory databases to update the patch information in time. One of the most popular public vulnerability advisory databases for security vulnerabilities is the National Vulnerability Database (NVD) [6]. It is a common practice for users to rely on NVD for monitoring related information regarding the updates, patches, etc. However, in recent years, due to the rapid growth of CVE requests, NVD has experienced a drastic delay in processing the CVE submitters’ requests [1], [7], [8]. For example, Figure 1 shows the NVD page of CVE-2013-1814, where the patch link is still missing after 11 years. Nevertheless, this patch actually exists (created in March 2013) and has been identified by the GitHub Advisory database in

2023 [9]. The patch is shown in Figure 1, and its relevance with the CVE can be justified by the overlapping token *"rpc"*.

Besides posing security threats, another challenge caused by the delay of NVD in updating the patch is the information validation of CVE-related maintenance tasks. For example, the GitHub advisory database [10] relies on users to submit pull requests for the updates of meta-data information about a CVE (e.g., affected package name, affected version), then GitHub maintainers review and approve these pull requests before they are merged in the database. However, by reviewing tens of pull requests, we observe that the pull requests submitted by users often lack the information source, causing challenges for maintainers in understanding these pull requests and delays in approving these pull requests. One example is the request for changing the affected package name in CVE-2013-1814 [9]. Since the requester did not submit a link to the original patch, the GitHub maintainer has to manually search the patch by herself to verify the request. While security personnel can manually search the web/repository for the patches, this approach is time-consuming and may be prone to bias.

In this work, we identify that a critical step for assisting security personnel (e.g., database maintainers, and submitters) in maintaining CVE-related information is to build a retrieval system for the CVE-related code, especially for retrieving the patch. The closest work to our problem is Sun et al. [11], who build an automated retrieval system for tracing the relevant fix at the file level. Their method employs neural network models including CodeBERT [12] and RoBERTa [13]. The neural networks are used for retrieving the files most relevant to a CVE based on the CVE description as the query. However, one major challenge with the neural network approach is the lack of explainability [2], [14], [15]. That is, while CodeBERT can output a ranked list of files, it cannot explain why it makes the decision of one ranked list and not another. On the other hand, since code commits typically contain a large number of tokens (the median number of tokens in our dataset in Table II is 1,840), understanding the decision of a CodeBERT model is even more challenging.

To assist security personnel make informed decisions, this paper proposes to build an explainable model for retrieving the patch. Existing work on explainable machine learning (XML) has studied how to improve the informativeness of human decision-making when provided with the result of a model. While there are various ways to explain a machine-learning decision, one widely used technique is to select a few tokens that are the most important for the model's decision-making or the user's mental model. Previous work on explainable machine learning (XML) [2] report that by highlighting important tokens in the input (i.e., the explanation), it is easier for human users to accurately annotate the class label. Motivated by previous work's findings, this paper studies the following question: can explainable machine learning help humans retrieve the relevant patch of a CVE? In Figure 1, we show an envisioned interface for this purpose. While the system retrieves the relevant patch, it also highlights the matched token *"rpc"*.

To answer the proposed question, our study includes the following steps. First, since no existing datasets can be directly used for the setting of our problem, we construct a new dataset. We reuse the (CVE, patch) pairs from existing datasets, then we leverage the version information (e.g., CPE) provided by the security database to reduce the search space to a smaller number of commits. Second, following Sun et al. [11], we train a language model (e.g., CodeBERT) that retrieves commits (i.e., commit message and diff code) given the CVE description. Third, we implement two explainable machine-learning algorithms for the retrieval model. The first algorithm is LIME [2], a widely applied explainable machine learning algorithm. However, we observe that LIME sometimes selects non-informative words. Therefore we further introduce another explainable machine learning algorithm called TfIdf-Highlight. Fourth, we evaluate the effectiveness of explainable machine learning algorithms with the faithfulness score, a widely used metric for XML [3], [16] as well as human labeling experiments.

The findings of our experiments are summarized as follows. First, by comparing the faithfulness scores (sufficiency and comprehensiveness) of LIME vs TfIdf-Highlight, we find that TfIdf-Highlight significantly lowers the sufficiency scores of LIME by around 15% in almost all settings. That is, the highlighted words selected by Tf-Idf can better reflect the model's true decision-making process. Second, in the human labeling experiment, we ask three annotators to blindly select the patch among a group of commits. Each group either uses LIME, TfIdf-Highlight, or no highlight, and the groups are mixed so the annotator cannot tell between LIME and TfIdf-Highlight. The results of the human labeling experiments show two things. First, TfIdf-Highlight is more helpful than LIME Highlight. Second, highlighting does not result in a higher labeling accuracy. The reason is that while the highlighted tokens can help with eliminating token-level mismatches (e.g., file names, variable names), they cannot explain the semantic relatedness between the CVE description and the commit.

In this paper, we make the following contributions:

- We propose the first study on using an explainable retrieval system to assist vulnerability patch tracing;
- We collect a dataset for vulnerability patch tracing. The dataset will be made publicly available;
- We propose an explainable machine learning algorithm called TfIdf-Highlight that can highlight informative words in the CVE description and the patch. TfIdf-Highlight significantly outperforms LIME in the faithfulness score;

II. AN EMPIRICAL STUDY ON THE AVAILABILITY OF PATCH LINKS IN NVD

In this section, we conduct an empirical study on the availability of patch links in NVD. To the best of our knowledge, we are not aware of any statistics on CVEs that guarantee to have a patch. Rather than reporting the absolute proportions of missing links, we find it more meaningful to compare the availability of links on one platform with another.

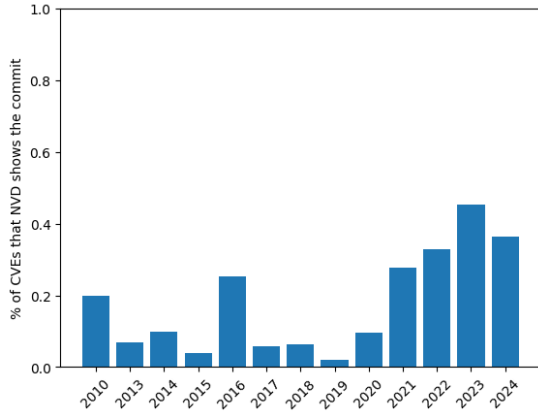


Fig. 2: The proportion of available patches in NVD compared to GitHub Advisory database (average: 0.3). That is, for all the CVEs in GitHub Advisory with at least 1 patch link, we examine their corresponding NVD pages where the patch link is not missing.

The reason is that not all CVEs have a patch: for some CVEs, a patch may not be available yet. To make sure the studied CVEs have at least one patch, we compare the patch availability of the NVD database with the reviewed CVEs in the GitHub Advisory Database [10]. The information of the reviewed CVEs (including the patch link) is verified by GitHub administrators, thus it guarantees the correctness of the metadata. All reviewed CVEs in the GitHub Advisory Database contain at least one patch link. More specifically, we investigate all the 2,000 reviewed CVEs under the Maven ecosystem [17]. For each CVE, we extract all the hyperlinks on the corresponding NVD webpage. Then we use regular expression to detect whether at least one hyperlink follows the format of a GitHub commit. We report this proportion (broken down by year) in Figure 2.

Results Analysis of Figure 2. In Figure 2, the average proportion of the available commit links in NVD is 0.3. That is, the link is missing in 70% CVEs. On the other hand, by observing the yearly trend, we can see that the proportion of NVD commits links has increased in the more recent years. For most of the years before 2020, the availability is lower than 10%. These results show that compared with GitHub Advisory, it is more difficult to trace the CVEs in the NVD database, especially for earlier years. Such statistics confirm that the shortage of manpower has caused significant delays in NVD data maintenance. It is imperative to investigate the automated approach for retrieving the relevant patch of a CVE.

III. PROBLEM FORMULATION

In this section, we formally define the problem of the explainable tracing of CVE patches in Open-Source software. Given all the available information about a CVE, the problem studies how to identify the commit that fixes the vulnerability. After identifying the repository name, e.g., `apache/tomcat`, the problem is reduced to ranking all the

commits from the repository. The query and candidates of the retrieval problem are explained below.

Query: CVE Information. For the query, we can leverage all the available information regarding a CVE, including the CVE description, the repository name, and the version of the patch. An example of CVE description can be found in Figure 1. The text description often mentions a natural language summary of the vulnerable code behavior, the affected software name, and the affected/patched software version, although not guaranteed to contain all fields [11]. The software name and version can help narrow down the candidates to the commits between the version ranges in the repository (We discuss this issue in Section IV-B). The software name and software version of a CVE can be automatically extracted from the CVE description using natural language processing techniques [18]–[20]. Meanwhile, most CVE databases [6], [10], [21] usually already provide the labeled software name and version. An example is the CPE information on an NVD page, e.g., “Up to (excluding) 8.5.7”.

The text of a CVE description usually contains two parts of information that can be used for matching the relevant candidate. First, a high-level summary of the vulnerability e.g., the description of CVE-2017-6056 [22] mentions “...It was discovered that a programming error...may result in denial of service via an infinite loop...”, which explains the vulnerability is *Infinite Loop* or CWE-79 [23]. Second, the descriptions often directly refer to the file name, function name, or variable name of the vulnerable code. For example, the description of CVE-2017-8085 [24] mentions the PHP file containing the vulnerability: “In Exponent CMS before 2.4.1 Patch 5, XSS in *elFinder* is possible in `framework/modules/file/connector/elfinder.php`”.

Candidate: Commit Information. To represent each commit, we leverage the commit message text and the Diff code. Some commit messages of the patch contain a summary of the vulnerability, e.g., the commit message of the patch for CVE-2017-6056 [22] is “...Fix potential infinite loop if `pos == 0...`”. The “infinite loop” in this commit message overlaps with the CVE description above. On the other hand, the diff code contains information about the changed file, changed function, changed lines-of-code, and variables, which helps with the matching if the CVE description directly refers to them.

Explainable Information Retrieval. Based on the above analysis, we propose to formulate the problem of patch tracing as an explainable information retrieval problem [25], [25]–[27]. Similar to explainable machine learning [2], explainable information retrieval highlights the important tokens in both the document and the query to help the searcher understand the ranking results of a retrieval model.

Given each query CVE q and a list of candidate commits d , our task is to first retrieve the ranked list of commits, then highlight the tokens in q and d to help with the reviewing. If the highlighting algorithm is effective, it should help the CVE maintainer quickly distinguish the patch commit from the non-patched commits.

IV. CONSTRUCTION OF A DATASET FOR SECURITY VULNERABILITY PATCH TRACING

In this section, we describe the data collection process.

A. Collecting CVEs and the Relevant Patch Commit Links

Dataset Requirements and limitations of the current dataset. To the best of our knowledge, no existing dataset can be directly used for our task of retrieving the relevant commit patch. The closest dataset is Sun et al. [11], however, the SHA of the patch commit in their dataset is missing, and it is non-trivial to recover this SHA based on their dataset.

Data Sources and Collecting Process. To collect more CVEs and more patch links, we use the combination of the following datasets: NVD [6], BigVul [28], patch_db [29], GitHub Security Advisory Database [10], and OSV Database [30]. For BigVul and patch_db, we use the labeled patch links. For NVD and GitHub advisory, we match GitHub commit links from the references as the patch commits for the current CVE. If multiple commits are extracted, the first one is selected. For OSV, we use its provided API to retrieve CVEs that are present in the databases but the GitHub commit links were not extracted in the previous databases. We gather the union set of (CVE, patch commit) pairs from five different data sources. In the case of conflicting patch commits, priority is given to the BigVul and patch_db datasets, as they provide labeled patch links. When neither BigVul nor patch_db contain the relevant information, we use the information in NVD and GitHub advisory.

While the patch can be from other websites, we focus on using the GitHub links, because it is easy to extract information using the GitHub API and related tools.

Among the 9,435 CVEs collected from the 5 sources, we successfully identified patch commits for 7,242 CVEs out of the 9,435 CVEs.

Removing Incorrect Commits. After collecting the (CVE, patch) pairs, we remove incorrect cases. The incorrect reasons are: (1) repositories that no longer exist or are inaccessible, (2) repository restructuring that results in the patch commit being excluded from the main tree, and (3) commit hash inconsistencies where the provided commit hash in the vulnerability databases does not match the actual patch commit. We further exclude repositories that are too large, e.g., `torvalds/linux`. After this filtering process, our final dataset contains 3,573 CVEs. This filtering ensures that our dataset consists of verifiable and accessible patch commits, which is crucial for both training and evaluation purposes.

B. Collecting the Candidate Commits

Since the number of candidate commits in a repository is large, we need to reduce the candidate size before retrieving the patch commit. We observe that most databases (e.g., NVD, GitHub Advisory, Snyk) provide a labeled patch version information. For example, for CVE-2017-8085 [24], the CPE information says the affected version is up to 2.4.0 (including), therefore the patch commit should locate between 2.4.0 and the next version. Technically, as long as this information is

correct, we can automatically filter the candidate commits by version. However, in practice, we find that version information in NVD can suffer from various data quality issues, including imprecise version records (e.g., using broader release versions instead of specific patch versions), incorrect version records, misalignment between repository tags and version numbers, and completely missing version information. These data quality issues will greatly affect the coverage of the method, so we end up constructing a dataset by assuming we already know the patch commit and using the nearby commits as the candidate commits. We report the findings of both methods below.

Method 1: Automatically Filtering Candidate Commits based on the NVD Version Tag. There exist two major challenges in mapping the labeled tag (e.g., 2.4.0) to the range of tags for filtering (e.g., 2.4.0 to 2.4.1). First, if the labeled tag is the lower bound, how to find the upper bound and vice versa? To address this challenge, we use Python’s `packaging` [31] library with regular expressions to normalize and compare version strings, which allows us to identify the next/previous version based on the current version information. Second, if the labeled tag (e.g., 2.4.0) has a different name than the tag in the original repository (e.g., `v2.4.0`), how to resolve the name difference? For this challenge, we apply `difflib`’s [32] `SequenceMatcher` algorithm.

Coverage Analysis of Method 1. Out of all the 3,573 CVEs identified in Section IV-A, only 2,784 contain the version tags. We manually inspected these 2,784 CVEs and found that only 1,805 contain the accurate patch version information in NVD. For the 1,805 CVEs, our tag matching using `difflib` has a 92.9% success rate, resulting in 1,688 CVEs. When the NVD tag is incorrect, we can increase the recall for covering the patch by increasing the candidate window to several versions before and after the NVD tag. We report this recall in Table I. Results in Table I indicate that while our fuzzy matching approach can effectively handle the version format differences, the actual coverage is significantly limited by version information quality in NVD. As a result, when preparing the candidate commits, we end up leveraging the true version tag. Although this information is not provided in the real application, we argue that our focus is to investigate the explainable tracing problem.

TABLE I: Statistics of distance range and recall

Range	Count	Total	Recall (%)
0	1,677	2,784	60.24
±1	1,875	2,784	67.35
±2	2,028	2,784	72.84
±3	2,083	2,784	74.82
±5	2,136	2,784	76.72
±10	2,263	2,784	81.29
±50	2,614	2,784	93.89

Method 2: Collecting Candidate Commits based on the True Tag (the Method we Actually Used). First, we find all tags that contain the patch commit. If we find multiple tags, we select the earliest one as the fixed version since it represents

the first version which includes the patch commit. Next, we retrieve the version that comes right before the fixed version, which is the last vulnerable version. Then, we collect all the commits between these two versions. These commits are our candidate pool of the potential patch commits because they represent all the changes between the last vulnerable version and the patched version.

Data Collection Results. For each collected candidate commit, we extract commit messages and metadata, modified file paths and file contents, and detailed code changes showing additions and deletions. By utilizing version tags to limit our search range, we significantly reduced the number of commits that need to be examined from 70,722,387 total repository commits to 1,310,068 candidate commits, representing a 98.1% reduction in the search space.

We report the statistics of our dataset in Table II. Table II shows that the positive and negative ratio is highly skewed. The diff code is usually a lot longer than the commit message.

TABLE II: Statistics of the dataset

Avg tokens (cve desc)	67.3
Avg tokens (commit msg)	43.9
Avg tokens (diff code)	1840.5
Median tokens (diff code)	645
3-rd Quartile tokens (diff code)	1846
# commits	114,523
# train	40,827
# validation	22,276
# test	51,420
Pos:neg	1:45

For token count we use the tokenizer of CodeBERT [12]

V. BUILDING AN EXPLAINABLE PATCH RETRIEVAL SYSTEM

In this section, we describe the investigated framework for the explainable tracing of the patch commits. First, in Section V-A, we describe how to train a multi-modal language model for the retrieval task (i.e., natural language and code). Second, in Section V-B, to identify the highlighted token, we first apply LIME [2], a highlighting token selection algorithm which is widely used for explainable machine learning [33], [34], including explainable information retrieval [25]. Third, we observe that some tokens highlighted by LIME are less informative, therefore we propose another highlighting method called TfIdf-Highlight (Section V-C). We describe each step as follows.

A. Training a Multi-Modal Language Model for Tracing Security Vulnerability Patches

Discussion on Retrieval Model Selection. Following existing work on similar tasks [11], [35], we train the retrieval model by fine-tuning language models, e.g., CodeBERT [12] and UnixCoder [36]¹. Notice that an alternative approach is to use vector-based retrieval models [37], [38]. That is,

¹In this paper, we focus on investigating smaller models. However, our framework can be generally applied to large language models such as Llama models.

first converting each of the CVE descriptions and commits into a vector, then training the retrieval model. We opt to not use the vector-based approach for the following reasons. First, compared with neural network, the vector-based models are harder to explain. Second, after reducing the candidate commits size by tag filtering, the computational cost is largely reduced. Therefore, we do not have to resort to faster solutions using the vector-based models.

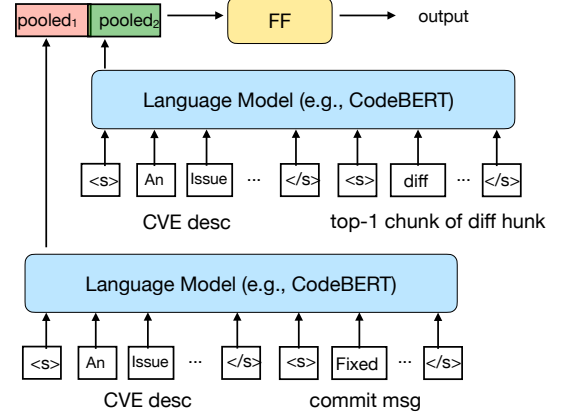


Fig. 3: The retrieval model for the multi-modal training

Removing Whitespace. We directly use the tokenizers of CodeBERT and UnixCoder. However, the raw tokenized outputs for code contain many whitespaces and a large number of tokens. To reduce the context length, we remove all the whitespaces (including line breaks and tabs). Although whitespaces are critical for defining indentations, removing them does not affect the token-level matching of the retrieval model.

Handling Long Contexts of Diff Code. Since CodeBERT and UnixCoder both have a context length of 512 tokens, it is difficult to handle long contexts of diff code without truncation. To handle the long context challenge, we split each diff into chunks of 64 tokens. We use a shorter length of chunks to: (1) reduce the chance of truncation for longer CVE descriptions; and (2) make it easier for LIME to select the rationale tokens from the chunk. Then we rank all chunks based on their Tf-Idf cosine similarity (explained in Section V-C) with the CVE description. We only use the top-1 chunk in the model.

Multi-Modal Training Framework. After the pre-processing, each commit is left with the commit message and one code chunk. We use the model in Figure 3. For each pair of (CVE, commit), we first use the model to obtain the pooled embedding vector for each of the text and code, then feed the concatenate pooled vector to the classifier on top. Since the data distribution is imbalanced, the weight of positive:negative examples in the training data is set to 10:1.

Multi-Modal Training: Handling Modality Bias. During the training of the model in Figure 3, we identify modality bias [39]. That is, when one modality (the code) is harder to train than another (the commit message), the model discovers

this shortcut and only learns the easier modality. In particular, we find that when tested with the code + an empty commit message, the model’s probability is nearly 0 (0.0269) for all positive cases. However, when tested with the commit message + an empty code, this problem does not happen. To force the model to learn from the code, we leverage data augmentation. For each positive example, we add 2 new examples: one with code only and one with commit message only. That is, if the commit is the patch, the model should recognize it even if only the code (commit message) is provided. We further add 10 times the single-modal negative examples to keep the positive:negative ratio consistent. After the data augmentation, the model’s positive probability for code + an empty commit increases to 0.312.

B. Explaining Patch Ranking Results using LIME

We first investigate using the LIME (locally interpretable model-agnostic explanation) algorithm [2] to select the highlighted tokens. Given the trained model, for each test example \mathbf{x} and a user-specified label (either 0 or 1), LIME can select k tokens in the test example to explain the reason why the model predicts that label. The way LIME selects the k tokens is by perturbing the test example into n alternative samples \mathbf{x}' , and using the model to obtain n probability scores p . Then LIME fits a ridge regression model (i.e., sparse linear regression) to p (if label = 1) or $1-p$ (if label = 0) by selecting k features.

One-Label Explanation Problem. Since our problem is a retrieval problem, i.e., for each CVE, rank all commits based on how likely they predict 1, we focus on always using LIME to explain 1 for all examples.

Commit/Code Only Highlighting. Since our problem is a retrieval problem, we focus on using LIME to explain the commit message and diff code only. That is, for each example, when perturbing $\mathbf{x} \rightarrow \mathbf{x}'$, we do not perturb the tokens in the CVE description. We set this constraint so we could control the number of tokens perturbed in the commit message and diff code. For the same reason, the tokens in the commit message and diff code are perturbed separately. That is, for each test example, we run LIME twice, each selecting k tokens from the commit message and diff code.

Dropping Non-informative Tokens. To prevent LIME from selecting non-informative tokens, we further restrict that the selected token must contain at least one letter or number.

C. TfIdf-Highlight

Although LIME can select k highlighted tokens, we observe that LIME sometimes selects non-informative tokens, e.g., meaningless words such as *are*, *its*, and *or*. It is a non-trivial problem to remove all the non-informative words (e.g., using rules or a stopword set). Therefore, we propose another approach for selecting the highlighted tokens.

Our approach, TfIdf-Highlight, is simply inspired by the Tf-Idf retrieval model. Given the CVE q , the commit message msg , and the diff code $diff$, for any word $w \in q$, we use the following TfIdf score to rank the words in msg and $diff$, where the top k words are selected as the rationale tokens:

$$\text{logtf}(w, d) = 0.5 \log_2\{\text{tf}(w, d_{msg})\} + 0.5 \log_2\{\text{tf}(w, d_{diff})\}$$

$$\text{idf}_{\text{CVE}}(w, q) = \max\{\log_2(\frac{N_q}{\text{df}_{\text{CVE}}(w, q)}) - \alpha, 0\}$$

$$\text{idf}_{\text{all}}(w) = \max\{\log_2(\frac{N}{\text{df}(w)}) - \beta, 0\}$$

$$\text{TfIdf}(w, d_{msg}) = \text{logtf}(w, d) \cdot \text{idf}_{\text{CVE}}(w, q) \cdot \text{idf}_{\text{all}}(w) \\ \text{if } w \in d_{msg} \cap q$$

$$\text{TfIdf}(w, d_{diff}) = \text{logtf}(w, d) \cdot \text{idf}_{\text{CVE}}(w, q) \cdot \text{idf}_{\text{all}}(w) \\ \text{if } w \in d_{diff} \cap q$$

Here, $\text{tf}(w, d_{msg})$ is the term frequency of word w in the commit message, $\text{tf}(w, d_{diff})$ is the term frequency of word w in the diff code, $\text{df}_{\text{CVE}}(w, q)$ is the number of commits under the CVE q which contain the word w , N_q is the total number of comments under the CVE q , and $\text{df}(w)$ is the number of commits in the entire dataset that contain the word w . Different from the standard Tf-Idf approach, here we introduce $\text{idf}_{\text{CVE}}(w, q)$ to penalize words that are less informative within the CVE. For example, words describing the repository name (e.g., “*rave*” in Figure 1) are non-informative within the repository but are informative outside the repository. Here we set $\alpha = 1$ and $\beta = 0.01$: if more than half of commits within the CVE contain the word, the word is not informative.

Tokenization using Snake/Camel Case. To match more tokens, we leverage regular expression to split all fields using the Snake/Camel case. We keep both the tokens before the split and after the split. For example, for the word *assertEquals*, 3 tokens are added: *assert*, *Equals*, and *assertEquals*.

After obtaining the highlighted tokens for each commit message and diff code, we map the highlighting result to the token space of the language model (e.g., CodeBERT and UnixCoder) using character indices. For each token by CodeBERT, the TfIdf-Highlight score is equal to the max of all overlapped original tokens.

VI. EVALUATION

In this section, we study the effectiveness of the proposed explainable tracing framework (Section V). Section VI-A evaluates the trained model’s retrieval performance. Section VI-B compares the rationale selection methods in Section VI-B. Finally, Section VI-C describes a blind human labeling experiment for comparing the rationale selection methods.

The research questions we investigate are the following:

- **RQ1:** How effective is the trained model for retrieving the patch commit?
- **RQ2:** What are the explanation performances of LIME and TfIdf-Highlight?
- **RQ3:** How do explanations affect human annotators’s decision-making process?

A. Evaluating the Retrieval Performance

Experiment Settings. We run all experiments on a cloud-based computing server (Ubuntu 22.04), with an RTX 4090 GPU (24GB).

Model Selection. In this paper, we study two language models for retrieving the patch commit: CodeBERT-base [12] and UnixCoder-base [36]. We choose CodeBERT because it has the best retrieval performance for the file-level patch-tracing task in Sun et al. [11]. Both CodeBERT and UnixCoder are frequently used for vulnerability-related tasks [40]–[42] including the retrieval of code [43]. While existing work shows that smaller language models (e.g., RNN, Bi-LSTM) are easier to explain, we skip these models since they usually have worse performance. We leave exploring large language models (e.g., Llama) for future work.

Hyperparameter Settings. For the model training, we use a learning rate = $2e-5$, batch size = 8, epoch = 5, and max_length = 256. Since the code is split into chunks of 64 tokens and we only use the top-1 chunk (Section V-A), only 4% text and 0.7% code are truncated into 256 tokens. This design is for keeping most tokens within the search space for highlighting.

TABLE III: Comparing the retrieval performance of different models

Metric		Tf-Idf	CodeBERT		UnixCoder	
		Test	Valid	Test	Valid	Test
Classify	Accuracy	-	0.979	0.984	0.976	0.984
	Prec	-	0.590	0.537	0.527	0.526
	Recall	-	0.651	0.657	0.646	0.674
	F1	-	0.619	0.591	0.580	0.591
	Auc	0.877	0.932	0.964	0.956	0.968
Rank	Prec@1	0.507	0.652	0.657	0.636	0.658
	Recall@1	0.507	0.652	0.657	0.636	0.658
	Recall@2	0.636	0.717	0.756	0.725	0.760
	Recall@5	0.754	0.780	0.825	0.787	0.833
	MAP	0.689	0.824	0.817	0.821	0.827

The pos:neg ratio is 1:59 in test and 1:39 in valid.

Baseline. We compare CodeBERT and UnixCoder with Tf-Idf [44]. For the tokenization of Tf-Idf, we leverage the same regular expression with Camel/snake case splitting as Section V-C. The weight for the commit message in Tf-Idf is 0.8 and for the diff code is 0.2.

Results Analysis. In Table III, we compare the ranking performance of different models. Overall, we find that UnixCoder and CodeBERT both outperform Tf-Idf. UnixCoder and CodeBERT have similar scores for classification and ranking. Notice that unlike the blackbox language models, the decision of Tf-Idf can be easily explained using the Tf and Idf scores. Since the explainable Tf-Idf model has a worse performance than the non-explainable CodeBERT and UnixCoder models, the goal of our work is meaningful. That is, we try to improve the explainability of the model without compromising on the ranking performance.

Summary of Findings for RQ1. We find that CodeBERT and UnixCoder achieve similar retrieval scores for patch retrieval.

B. Evaluating the Performance of Rationale Selection

In this section, we evaluate the explanation performance of TfIdf-Highlight by comparing it with LIME [2].

Defining Faithfulness in Ranking. Existing evaluation metrics for explainable machine learning often require comparison with ground truth tokens [3], [45]. However, since our dataset is large and manually labeling long code data is challenging (Table II shows the average length of the diff code is 1,840 tokens), we have not collected the ground truth labeling and will leave it for future work. One evaluation metric that does not require ground truth labeling is called *faithfulness*, which captures the extent to which the highlighted tokens result in the same model decision as the full input [16], [45]. For example, in Matthew et al. [45], faithfulness is defined as the distance between $proba[predicted_label]$ and $proba'[predicted_label]$, where $proba$ is the predicted probability using the original input, and $proba'$ is the predicted probability using only the highlighted tokens as the input (sufficiency), or excluding the highlighted tokens (comprehensiveness). However, their problem is a multi-class classification problem, which is different from our ranking problem. We define the faithfulness of ranking below:

$$faithfulness = \frac{1}{n} \sum_{cve} |Prec@1(cve, proba) - Prec@1(cve, proba')|$$

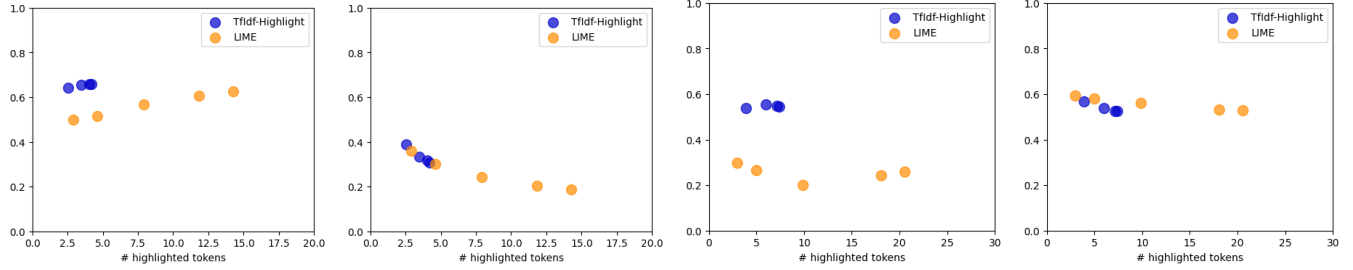
Here $Prec@1(cve, proba)$ is the precision@1 for each query cve using $proba$ as the probabilities. This $Prec@1$ can be replaced by other ranking metrics. Similar to Matthew et al. [45], we compare the ranking using the highlighted tokens only probabilities $proba'$ (or excluding the highlighted tokens) with the original probabilities $proba$. Since we only explain either the commit message or the diff code only, but not both, $proba'$ is obtained by keeping the code the same as the full input, while using only the highlighted tokens for the commit message, etc.

Controlling the Number of Highlighted Tokens. When comparing the faithfulness scores between LIME and TfIdf-Highlight, we need to keep the number of highlighted tokens the same for each example. This is because the higher the number of highlighted tokens, the closer to the original input, therefore the model will have a lower sufficiency score and a higher comprehensiveness score. While it is easier to control the number of highlighted tokens in LIME, TfIdf-Highlight will cut the highlighted tokens based on the Tf-Idf score, so the number of highlighted tokens is different for each example. To control the number of highlighted tokens, we first cut the highlighted tokens in Tf-Idf at k (therefore Tf-Idf may highlight 0- k tokens), then use LIME to highlight the same number of tokens as Tf-Idf. This setting guarantees a fair comparison between TfIdf-Highlight and LIME.

Result Analysis. We report the faithfulness scores in Table IV. For each setting, we report the average number of highlighted tokens in the positive examples, the average sufficiency and comprehensiveness scores, and the statistical

TABLE IV: Comparing the faithfulness of TfIdf-Highlight vs LIME [2] by controlling the #highlighted tokens

Model	Component	Fold	#tokens(pos)		Sufficiency↓		t	LIME	Comprehensiveness ↑		t
				LIME	TfIdf-Highlight	p-val			TfIdf-Highlight	p-val	
CodeBERT	Text only	Valid	3.16	0.398	0.286	4e-9	-5.9	0.089	0.116	3e-3	2.89
		Test	3.22	0.351	0.305	2e-3	-3.07	0.125	0.139	0.115	1.57
	Code only	Valid	4.15	0.494	0.403	3e-6	-4.68	0.057	0.081	0.016	2.4
		Test	4.14	0.484	0.413	2e-6	-4.7	0.061	0.075	0.05	1.94
UnixCoder	Text only	Valid	2.88	0.285	0.232	9e-4	-3.3	0.112	0.125	0.238	1.17
		Test	3.06	0.325	0.275	3e-4	-3.61	0.136	0.140	0.722	0.355
	Code only	Valid	3.70	0.432	0.344	5e-8	-4.57	0.088	0.088	1.0	0
		Test	3.79	0.424	0.366	3e-4	-3.5	0.083	0.083	0.884	0.145



(a) Pos prob w Highlight Text↑ (b) Pos prob wo Highlight Text↓ (c) Pos prob w Highlight Code↑ (d) Pos prob wo Highlight Code↓

Fig. 4: The trend of explainability score vs. the number of highlighted tokens. Model: CodeBERT, Fold: valid

significance test scores. From Table IV we can see that the sufficiency scores of LIME are worse than TfIdf-Highlight in almost all settings. The sufficiency metric measures the extent to which the highlighted tokens reflect the original model’s decision-making process. While the LIME tokens are identified based on the model, TfIdf-Highlight does not need to know any information about the model, thus this result is surprising.

By observing the values in Table IV, we find that the text-only mode usually has a lower sufficiency score and a higher comprehensiveness score. This result meets our expectations because the commit message is usually more informative than the code for the retrieval.

Explainability Score vs. #Highlighted Tokens. To understand the trend of TfIdf-Highlight vs LIME under different numbers of highlighted tokens, we further show the plot in Figure 4. The y-axis of Figure 4 is the average positive probabilities in *proba'* (i.e., the probability with/without only the highlighted tokens) of all the positive examples (i.e., patch). From Figure 4 we can see that even with a few times as many highlighted tokens, it is difficult for LIME to achieve the same positive probabilities as TfIdf-Highlight.

Summary of Findings for RQ2. We find that TfIdf-Highlight significantly outperforms BERT in the sufficiency score for all settings and slightly outperforms BERT in the comprehensiveness score.

C. Human Labeling of the Rationale Selection

While the explainability scores reflect a highlight’s faithfulness to a model’s decision-making, it does not reveal how helpful the tokens are for human reviewers to make a decision.

Therefore, we conduct a user study experiment to examine the helpfulness of the highlights.

	Content	label	Helpfulness
CVE desc	<s>SQL injection vulnerability in the checkEmailFormat function in plugins/jojo_core/classes/Jojo.php in Jojo before 1.2.2 allows remote attackers to execute arbitrary SQL commands via the X-Forwarded-For HTTP header to /articles/test/.</s>		
	<s>Core - Editor: Add field label to inline editor</s>		
commit	diff --git a/plugins/jojo_core/external/normalize b/plugins/jojo_core/external/normalize index 97e7867..0357529 160000 --- a/plugins/jojo_core/external/normalize +++ b/plugins/jojo_core/external/normalize @@ -1 +1 @@ -Subproject commit 97e78671617343e9ef8d8fc221200ecf1c18f260 +Subproject commit 03575299fc8c8ba752a51fcec7c35343085d6ed09		0
CVE desc	<s>SQL injection vulnerability in the checkEmailFormat function in plugins/jojo_core/classes/Jojo.php in Jojo before 1.2.2 allows remote attackers to execute arbitrary SQL commands via the X-Forwarded-For HTTP header to /articles/test/.</s>		
	<s>Fix SQL injection issue CVE-2013-3081</s>		
commit	diff --git a/plugins/jojo_core/classes/Jojo.php b/plugins/jojo_core/classes/Jojo.php index c61e1af..c47c5db 100755 --- a/plugins/jojo_core/classes/Jojo.php +++ b/plugins/jojo_core/classes/Jojo.php @@ -2461,14 +2461,18 @@ class Jojo { /* Gets the IP address of the visitor, bypassing proxies */ static function getip()	1	3

Fig. 5: An example of the human labeling interface

Annotator Information. Three authors separately provide the labels. Annotator 1 is a CS PhD student with knowledge on security vulnerabilities and frequent CWEs. Annotator 2 and Annotator 3 are CS undergraduate students with knowledge on security vulnerabilities.

First Blind Experiment (Easy). For the first experiment, we ask 2 annotators to guess the patch from 100 groups of (CVE, commit) pairs. All examples are highlighted, and an example of the interface of labeling is shown in Figure 5 (implemented using Google App Script [46]). For each group of CVEs, we display the original commit message and the diff

code, while highlighting the tokens selected by the highlighting algorithm.

Each group contains 1 CVE and 2 commits. One commit is the patch and the other commit is the top-2 ranked commit using TF-IDF. The choice of top-2 is for increasing the challenge of the labeling task. The 2 commits in each group are randomly shuffled. For each group, we randomly assign one highlighting method to annotator 1 and assign the other method to annotator 2. The annotators are not told which highlighting method is used in each group. The annotators are asked to provide 2 labels: choose the patch from the 2 commits, and rate the helpfulness from 1-3 Likert scale (1: not helpful at all; 2: the highlighted token can explain the match, but incomplete; 3: the highlighted tokens are perfect for explaining the match). The results in Table IV indicate that TfIdf-highlight receives higher average helpfulness ratings than LIME. However, the accuracies of labeling with the two highlighting methods are similar.

Second Blind Experiment (Hard). The second experiment is similar to the first except the following differences: (1) adding no highlighting as the third setting (as a result we add a third annotator); (2) if at least one commit is for changing .txt, .md, etc., we remove the entire group. This step is for eliminating cases that are too simple to label; (3) increasing the group size of commits from 2 to 4 to increase the challenge; (4) the number of groups is reduced to 50. (5) including two rounds: in the first round, the commit message is hidden, annotators need to select by reading the code only; in the second round, all information is revealed (including the first round label), they need to label whether to change the first round label. Same as the first experiment, the highlighted and non-highlighted groups are randomly shuffled and not revealed.

The result of the second experiment can be found in Table VI. We can make the following observations: first, the overall accuracy of with highlighting is actually lower than without highlighting. Second, by comparing TfIdf-Highlight with LIME, we find that when labeling with code only, TfIdf-Highlight has a worse accuracy than LIME, but after revealing the commit message, their accuracies are the same. This result shows that commits are critical for improving the accuracies. By comparing the three annotators, we find that the PhD annotator 1 who has knowledge on CWEs has a better accuracy than the two undergraduate annotators.

Case Study of Wrong Guesses. We inspect the reasons why the accuracy of without highlighting is higher than with highlighting. By interviewing the annotators, we find that the mental model they use for making decisions has 2 steps: first, they use the keywords in the CVE description and look for the exact match information (e.g., file names, variable names) in the commits to handle the easy cases; second, if the first step feels hard, they look for the changed lines of code for a deeper understanding of the vulnerability, and see if it matches the vulnerability statement in the CVE description. While highlighting (especially TfIdf-Highlight) shows the token-level matching, it cannot reflect a deeper

TABLE V: Results of the first user experiment (easy)

Labeler	Remove Changelog	Method	Accuracy	helpfulness
1	True	TfIdf-Highlight	0.92	2.30
		Lime	0.96	2.05
		Total	0.94	2.20
2	True	TfIdf-Highlight	0.93	2.46
		Lime	0.88	2.11
		Total	0.91	2.31
overall	True	TfIdf-Highlight	0.92	2.40
		Lime	0.93	2.08
		Total	0.92	2.24

Cohen’s Kappa [47]: 0.5665

TABLE VI: Results of the second user study experiment (hard)

Labeler	CommitMsg	Method			Total
		TfIdf-Highlight	Lime	None	
1	True	1.0	0.72	0.82	0.84
	False	0.8	0.77	0.88	0.82
2	True	0.43	0.91	0.91	0.76
	False	0.56	1.0	0.87	0.8
3	True	0.84	0.71	0.8	0.78
	False	0.63	0.57	0.7	0.62
Overall	True	0.76	0.76	0.86	0.79
	False	0.66	0.74	0.84	0.75

Fleiss’ Kappa [47]: 0.7244

understanding of the vulnerability. For example, CVE-2020-5295 is “... , an attacker can exploit this vulnerability to read local files of an October CMS server. ...”, and the patch commit includes words related to file operation such as *file*, *directory*, and *path*. Tf-Idf does not highlight the word *file* because it is a frequent word in the entire dataset, it also misses the other words due to the token mismatch. In general, a negative commit may have more matched tokens than the patch commits, thus highlighting can indeed be a double-sided sword in distracting the annotator from looking for the correct rationale tokens.

Summary of Findings for RQ3. We find that the reasoning process for selecting the patch requires a deeper understanding of the semantic relatedness between the CVE description and the commit. Highlighting based on only token-level matching using Tf-Idf may not be sufficient to support the patch tracing task.

VII. RELATED WORK

In this section, we summarize the related work.

Security Vulnerability Database Maintenance. To assist security personnel in detecting, maintaining, and tracing a security vulnerability, centralized databases are constructed to keep track of security vulnerabilities. The two major databases for security vulnerabilities are NVD [6] and MITRE [48], which maintain synchronized information about common vulnerabilities and exposures (CVE). One critical task for security vulnerability databases is to keep track of critical metadata information to help users be aware of the update, e.g., the

version number, the software name [18]–[20], and the fine-grained affected package name (e.g., Java) [35], [49]. For example, by identifying the affected package name, users who use the package can be alerted when there is an update of the CVE.

In recent years, due to the tremendous growth of CVEs, the manual maintenance of vulnerability databases has become more and more challenging [1], [7], [8]. To this end, existing work builds automated systems to accelerate the maintenance tasks. For example, existing work uses named-entity recognition (NER) to extract the software name and version information [20], [50]–[52]. Other work also models the problem as extreme multi-class classification (XML) where each software name is selected from a pre-defined list of packages [18], [19], [53]. However, the XML approach can only handle thousands of package names. To identify fine-grained software packages from an even larger number of candidates (e.g., Java packages), Chen et al. [35] train a retrieval model using the BERT model, which can identify the affected package with high accuracy when there are tens of thousands of candidates. Chen et al. [49] further leverage large language models to directly generate the affected package names.

Explainable Machine Learning. In recent years, deep neural network models have been widely applied to applications in all industries. However, one major challenge with deep learning models is the lack of transparency in the output [2], [15]. When deploying an automated machine learning system, it is critical to know that the model is making the right decision in order to build trust between humans and the model. Matthew et al. [45] uses explainable machine learning on hate speech detection tasks. They find that smaller models such as RNN and Bi-LSTM are easier to explain than larger models such as BERT. They use regularization to force the trained model’s attention to focus on the labeled tokens. They find the model’s explainability scores improve after training with regularization. Guo et al. [15] develop an explainable machine-learning learning model algorithm and apply it on security applications including binary code analysis and classifying PDF malware. Explainable machine learning has also been used for anomaly detection to capture an attacker’s strategy and help the user make better decisions [54].

Ribeiro et al. [2] find that the highlighted tokens can help non-expert human subjects make better decisions on classification tasks. Alufaisan et al. finds that users tend to follow the AI prediction when it is available [55]. Previous work finds that humans struggle with interpreting the explanation when the highlighted areas do not align with the human’s intuition [56], [57]. Besides highlighting the tokens (i.e., hard explanation), existing work also defines explanations as a set of soft rules [58].

VIII. DISCUSSIONS AND FUTURE WORK

Combining Explanations with Domain Knowledge. The user study results in Section VI-C show that highlighting may negatively affect the accuracy of labeling. To support users in

making the right decision, the highlight has to capture the semantic relatedness between the CVE description and the commit. To this end, one future work direction is to leverage the security domain knowledge using knowledgebase information such as the common weakness enumerations (CWE).

Using Large Language Models to Enhance Explanations.

Another format of explanation is to use natural language to explain the relation between the CVE description and the commit. We can leverage large language models to generate rationales for the match. For example, if the annotator does not understand why the highlighted tokens reflect a vulnerability, we can explain the general rules to bridge the knowledge gap of the annotator.

IX. THREATS TO VALIDITY

The statistics on NVD patch availability (Section II) are based on the comparison with 2000 CVEs on Maven in the GitHub Advisory database. When using other CVEs, this result may change.

Our study has excluded the CVEs and commits in large repositories such as `torvalds/linux` (Section IV) because their number of commits between versions greatly exceeds our processing capacity. We are capable of collecting enough data to validate our idea while keeping the computational demands under control using this selective approach.

Our user study focuses on asking three annotators to label 50-100 groups of commits. Since the number of labeled CVEs is relatively small, the conclusion may not stay true when the human subjects are a different group of people.

X. CONCLUSION

In this work, we conduct the first study on using explainable retrieval for tracing the vulnerable patch of a security vulnerability. First, we construct a dataset for the automated tracing of a security vulnerability patch. Second, we train a multi-modal retrieval model for the commit message and diff code given the CVE description. Third, we apply LIME, a widely applied explainable machine learning algorithm to the patch tracing problem. Fourth, we propose a new explanation algorithm called TfIdf-Highlight, which leverages the TfIdf information to find the most informative words in the repository and the entire dataset.

Our evaluation shows that TfIdf-Highlight consistently outperforms LIME on the faithfulness score. That is, compared with LIME, the tokens selected by TfIdf-Highlight truly reflect the decision-making process of the trained model. By asking three annotators to blindly label the patch, we find that TfIdf has a similar labeling accuracy to LIME and a higher helpfulness score. Nevertheless, the accuracy of highlighting is lower than non-highlighting, since the highlighted tokens only reflect the token-level matching. To help with the decision-making, the highlighting needs to go beyond token matching to reflect a deeper understanding of the security domain knowledge.

REFERENCES

- [1] S. Hendery, "Update delays to nist vulnerability database alarms researchers," 2024. [Online]. Available: <https://www.scmagazine.com/news/update-delays-to-nist-vulnerability-database-alarms-researchers>
- [2] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [3] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," *arXiv preprint arXiv:1911.03429*, 2019.
- [4] W. Charoenwet, P. Thongtanunam, V.-T. Pham, and C. Treude, "An empirical study of static analysis tools for secure code review," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2024, pp. 691–703.
- [5] Wikipedia, "2017 equifax data breach," 2024. [Online]. Available: https://en.wikipedia.org/wiki/2017_Equifax_data_breach
- [6] "National vulnerability database," 2023. [Online]. Available: <https://nvd.nist.gov/>
- [7] C. Hughes, "Death knell of the nvd?" 2024. [Online]. Available: <https://www.resilientcyber.io/p/death-knell-of-the-nvd>
- [8] E. Chin, "National vulnerability database: Opaque changes and unanswered questions," 2024. [Online]. Available: <https://anchore.com/blog/national-vulnerability-database-opaque-changes-and-unanswered-questions>
- [9] GitHub, "The github advisory pull request for cve-2013-1814," 2024. [Online]. Available: <https://github.com/github/advisory-database/pull/3210>
- [10] —, "Github advisory database," 2024. [Online]. Available: <https://github.com/advisories>
- [11] J. Sun, J. Chen, Z. Xing, Q. Lu, X. Xu, and L. Zhu, "Where is it? tracing the vulnerability-relevant files from vulnerability reports," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.
- [12] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang *et al.*, "Codebert: A pre-trained model for programming and natural languages," *arXiv preprint arXiv:2002.08155*, 2020.
- [13] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.
- [14] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *proceedings of the 26th Symposium on Operating Systems Principles*, 2017, pp. 1–18.
- [15] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 364–379.
- [16] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" *arXiv preprint arXiv:2004.03685*, 2020.
- [17] GitHub, "All the reviewed cves under the maven ecosystem in github advisory," 2024. [Online]. Available: <https://github.com/advisories?query=type%3AReviewed+ecosystem%3AMaven>
- [18] Y. Chen, A. E. Santosa, A. Sharma, and D. Lo, "Automated identification of libraries from vulnerability data," in *ACM/IEEE International Conference on Software Engineering: Software Engineering in Practice*, 2020. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3377813.3381360>
- [19] S. A. Haryono, H. J. Kang, A. Sharma, A. Sharma, A. Santosa, A. M. Yi, and D. Lo, "Automated identification of libraries from vulnerability data: Can we do better?" in *IEEE/ACM International Conference on Program Comprehension*, 2022. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3524610.3527893>
- [20] G. Yang, S. Dineen, Z. Lin, and X. Liu, "Few-sample named entity recognition for security vulnerability reports by fine-tuning pre-trained language models," in *Deployable Machine Learning for Security Defense: Second International Workshop*, 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-87839-9_3
- [21] Snky, "Snyk open source vulnerability database," 2024. [Online]. Available: <https://security.snyk.io/>
- [22] NVD, "The nvd page for cve-2017-6056," 2024. [Online]. Available: <https://nvd.nist.gov/vuln/detail/CVE-2017-6056>
- [23] MITRE, "The information page of cwe 79," 2024. [Online]. Available: <https://cwe.mitre.org/data/definitions/79.html>
- [24] NVD, "The nvd page for cve-2017-8085," 2024. [Online]. Available: <https://nvd.nist.gov/vuln/detail/CVE-2017-8085>
- [25] J. Singh and A. Anand, "Exs: Explainable search using local model agnostic interpretability," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 770–773.
- [26] L. Lyu and A. Anand, "Listwise explanations for ranking models using multiple explainers," in *European Conference on Information Retrieval*. Springer, 2023, pp. 653–668.
- [27] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, and Z. Zhang, "Explainable information retrieval: A survey," *arXiv preprint arXiv:2211.02405*, 2022.
- [28] 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*, ser. MSR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 508–512. [Online]. Available: <https://doi.org/10.1145/3379597.3387501>
- [29] X. Wang, S. Wang, P. Feng, K. Sun, and S. Jajodia, "Patchdb: A large-scale security patch dataset," in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2021, pp. 149–160.
- [30] OSV, "Osv database," 2024. [Online]. Available: <https://osv.dev/>
- [31] P. P. Authority, "packaging: Core utilities for python packages," 2024. [Online]. Available: <https://pypi.org/project/packaging/>
- [32] P. S. Foundation, "difflib — helpers for computing deltas," Python Standard Library, 2024. [Online]. Available: <https://docs.python.org/3/library/difflib.html>
- [33] Y. Wu, L. Zhang, U. A. Bhatti, and M. Huang, "Interpretable machine learning for personalized medical recommendations: A lime-based approach," *Diagnostics*, vol. 13, no. 16, p. 2681, 2023.
- [34] A. Halilovic and F. Lindner, "Explaining local path plans using lime," in *International Conference on Robotics in Alpe-Adria Danube Region*. Springer, 2022, pp. 106–113.
- [35] T. Chen, L. Li, B. Shan, G. Liang, D. Li, Q. Wang, and T. Xie, "Identifying vulnerable third-party libraries from textual descriptions of vulnerabilities and libraries," *arXiv preprint arXiv:2307.08206*, 2023.
- [36] D. Guo, S. Lu, N. Duan, Y. Wang, M. Zhou, and J. Yin, "Unixcoder: Unified cross-modal pre-training for code representation," *arXiv preprint arXiv:2203.03850*, 2022.
- [37] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
- [38] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.
- [39] Y. Guo, L. Nie, H. Cheng, Z. Cheng, M. Kankanhalli, and A. Del Bimbo, "On modality bias recognition and reduction," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–22, 2023.
- [40] Z. Wang, G. Li, J. Li, Y. Xiong, and Z. Jin, "M2cvd: Multi-model collaboration for code vulnerability detection," *arXiv preprint arXiv:2406.05940*, 2024.
- [41] J. Zhou, M. Pacheco, Z. Wan, X. Xia, D. Lo, Y. Wang, and A. E. Hassan, "Finding a needle in a haystack: Automated mining of silent vulnerability fixes," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2021, pp. 705–716.
- [42] X. Zhou, T. Zhang, and D. Lo, "Large language model for vulnerability detection: Emerging results and future directions," *arXiv preprint arXiv:2401.15468*, 2024.
- [43] T. Chen, L. Li, T. Qian, Z. Wang, G. Liang, D. Li, Q. Wang, and T. Xie, "Identifying vulnerability patches by comprehending code commits with comprehensive change contexts," *arXiv preprint arXiv:2310.02530*, 2023.
- [44] Wikipedia, "tf-idf," 2024. [Online]. Available: <https://en.wikipedia.org/wiki/Tf%2Eidf>
- [45] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.
- [46] Google Cloud Platform, "Google apps script," [Online]. Available: <https://developers.google.com/apps-script/guides/sheets>
- [47] M. L. McHugh, "Lessons in biostatistics," *Biochem. Med*, vol. 19, pp. 120–126, 2009.
- [48] "The mitre cve website," 2023. [Online]. Available: <https://cve.mitre.org/>

- [49] T. Chen, L. Li, L. Zhu, Z. Li, G. Liang, D. Li, Q. Wang, and T. Xie, "Vullibgen: Identifying vulnerable third-party libraries via generative pre-trained model," *arXiv preprint arXiv:2308.04662*, 2023.
- [50] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022. [Online]. Available: <https://arxiv.org/abs/2301.00234>
- [51] A. Anwar, A. Abusnaina, S. Chen, F. Li, and D. Mohaisen, "Cleaning the nvd: Comprehensive quality assessment, improvements, and analyses," *IEEE Transactions on Dependable and Secure Computing*, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9601266>
- [52] P. Kuehn, M. Bayer, M. Wendelborn, and C. Reuter, "OVANA: An approach to analyze and improve the information quality of vulnerability databases," in *International Conference on Availability, Reliability and Security*, 2021. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3465481.3465744>
- [53] Y. Lyu, T. Le-Cong, H. J. Kang, R. Widyasari, Z. Zhao, X.-B. D. Le, M. Li, and D. Lo, "Chronos: Time-aware zero-shot identification of libraries from vulnerability reports," in *International Conference on Software Engineering*, 2023. [Online]. Available: <https://doi.org/10.1109/ICSE48619.2023.00094>
- [54] Z. Li, Y. Zhu, and M. Van Leeuwen, "A survey on explainable anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 1, pp. 1–54, 2023.
- [55] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, "Does explainable artificial intelligence improve human decision-making?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6618–6626.
- [56] C. Molnar, G. König, J. Herbringer, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General pitfalls of model-agnostic interpretation methods for machine learning models," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2020, pp. 39–68.
- [57] P. Knab, S. Marton, and C. Bartelt, "Dseg-lime—improving image explanation by hierarchical data-driven segmentation," *arXiv preprint arXiv:2403.07733*, 2024.
- [58] S. Chen, S. Bateni, S. Grandhi, X. Li, C. Liu, and W. Yang, "Denas: automated rule generation by knowledge extraction from neural networks," in *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 2020, pp. 813–825.