





Software Vulnerability Prediction Knowledge Transferring Between Programming Languages

Khadija Hanifi¹^a, Ramin F. Fouladi¹^b, Basak Gencer Unsalver²^c and Goksu Karadag²^d

¹Ericsson Security Research, Istanbul, Turkey

²Vodafone, Istanbul, Turkey

fi

Keywords: Software Security, Vulnerability Prediction, Source Code, Machine Learning, Transfer Learning.


Abstract: Developing automated and smart software vulnerability detection models has been receiving great attention from both research and development communities. One of the biggest challenges in this area is the lack of code samples for all different programming languages. In this study, we address this issue by proposing a transfer learning technique to leverage available datasets and generate a model to detect common vulnerabilities in different programming languages. We use C source code samples to train a Convolutional Neural Network (CNN) model, then, we use Java source code samples to adopt and evaluate the learned model. We use code samples from two benchmark datasets: NIST Software Assurance Reference Dataset (SARD) and Draper VDISC dataset. The results show that proposed model detects vulnerabilities in both C and Java codes with average recall of 72%. Additionally, we employ explainable AI to investigate how much each feature contributes to the knowledge transfer mechanisms between C and Java in the proposed model.


1 INTRODUCTION


Developers are concerned with the correctness of the written codes to run in a desired way and meet pre-defined design specifications. Along with that, they utilize different code analyzing techniques to ensure that the written code is robust enough and free of any weaknesses (called vulnerabilities) that could be exploited by attackers to carry out their malicious activities. Managing software vulnerabilities involves a wide range of code analyzing techniques to enhance the security and to ensure the confidentiality, integrity, and availability of the system. Source code analyzing techniques are usually classified into two main groups; static and dynamic code analysis (Palit et al., 2021). Static code analysis examines the source code without executing and running the code. It utilizes pre-defined set of rules and coding standards to analyze how much the code meets these rules and standards. On the other hand, dynamic code analysis entails running code and examining the outcome. This involves testing different possible execution paths of


the code and examining their outputs. Dynamic code analysis is able to find security issues caused by the program interaction with other system components such as SQL databases or Web services. However, in dynamic code analysis, to build a series of correct inputs for test coverage, a pre-knowledge of the program steps is needed. Although static and dynamic analysis have different approach to find the vulnerabilities, they both suffer from high false positive rate, necessitating human expertise to review the results which yields extra time, effort, and cost.

Many studies in literature have explored methods to decrease the false positive rate of both static and dynamic analysis. A common approach utilized in these studies is the use of machine learning (ML) techniques to train vulnerability detection models (Lin et al., 2020). While ML-based models show potential for improving vulnerability prediction accuracy, they also face two significant challenges. Firstly, the process of addressing various types of vulnerabilities is resource-intensive and requires both software and human expert analysis. This is further complicated by the fact that commercially used source code is subject to intellectual property rights and considered as confidential information by enterprises, making it difficult to obtain access to real labeled vulnerability

^a <https://orcid.org/0000-0001-7044-3315>

^b <https://orcid.org/0000-0003-4142-1293>

^c <https://orcid.org/0000-0001-9426-8400>

^d <https://orcid.org/0000-0002-4596-0983>

data. To mitigate this issue, synthetic datasets, such as the NIST Software Assurance Reference Dataset¹ and Draper VDISC dataset, are generated to simulate different vulnerability types for specific programming languages. However, these datasets are limited in scope, covering only published vulnerability types for a limited range of programming languages. Secondly, the structure and logic of each programming language are distinct, which results in language-specific feature extraction and varying numbers of features. This makes it challenging to develop a single model that can be used for all programming languages.

In this study, we propose a novel technique to address the challenge of limited vulnerability sample availability in some programming languages. Our approach utilizes available datasets to train a machine learning model for vulnerability detection in a related programming language and applies this knowledge to predict vulnerabilities in other programming languages with limited training samples. The main contributions of our work are:

- An automated and interpretable software vulnerability prediction model using machine learning techniques, which can learn from datasets in one programming language and apply its knowledge to predict vulnerabilities in other programming languages.
- A code representation method that converts source code into numerical vectors for machine learning analysis, combined with a syntax matching technique for matching related components between Abstract Syntax Trees (ASTs) from different programming languages.
- A new method for vulnerability prediction knowledge transfer, which involves customized processing steps for different programming languages and common final steps for vulnerability prediction.

We carried out a preliminary experiment to demonstrate our approach using the common vulnerabilities between C and Java programming languages. Our results show that we were able to transfer vulnerability prediction knowledge from C programming language to Java and that our machine learning model was able to make accurate predictions. Additionally, we leveraged explainable AI methods to better understand and verify the correctness of our model.

2 RELATED WORK

Software vulnerability prediction is dramatically rising in popularity in research community and industry,

¹<https://samate.nist.gov/SARD>

especially among cyber security professionals dealing with vulnerability management and secure software development life cycle (SSDLC). This spike of popularity is due to the rise in cyber security threats which exploit vulnerabilities for their malicious intent. Although several studies exist in the literature for software vulnerability prediction, the use of ML-based approaches are ubiquitous in the literature (Hanif et al., 2021). In general, two main approaches of metric-based and pattern-based are utilized with ML algorithms for software vulnerability prediction (Li et al., 2021). Metric-based approaches link the software engineering metrics, such as code complexity metrics (Kalouptsoglou et al., 2022; Halepmollasi et al., 2023) and developer activity metrics (Coskun et al., 2022) to software vulnerabilities and use those metrics to train ML models for vulnerabilities prediction. Although metric-based approaches are lightweight to analyze a large-scale program, they suffer from high positive rates. On the other hand, pattern-based approaches improve the efficiency and the automation of software vulnerability prediction. Authors in (Bilgin et al., 2020) proposed a method for software vulnerability prediction on the function level for C code. While preserving the structural and semantic information in the source code, the method transforms the AST of the source code into a numerical vector and then utilizes 1D CNN for software vulnerability prediction. Authors in (Duan et al., 2019), used the Control Flow Graph (CFG) and AST as the graph representation and utilized soft attention to extract high-level features for vulnerability prediction. Authors in (Zhou et al., 2019), proposed a function-level software vulnerability prediction method based on a graph representation that utilize AST, dependency, and natural code sequence information.

With respect to applying transfer learning for software vulnerability prediction, authors in (Ziems and Wu, 2021) developed various deep learning Natural Language Processing (NLP) models to predict vulnerabilities in the C/C++ source code. They used transfer learning to adapt some pre-trained models for English language such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) to be reused for software vulnerability prediction. The performance is promising despite the different structure between English language and the C/C++ programming language. Authors in (Lin et al., 2019), proposed a software vulnerability prediction method based on transfer learning and Long-short Term Memory (LSTM). They used several heterogeneous and cross-domain data sources combined to obtain a general representation of patterns for vulnerability prediction.

Although pattern-based approaches improve the prediction performance with respect to false positive rate, they are not adaptable and flexible enough to be re-used in other domains rather than the domain they are trained for. TL-based methods address this issue to some extent; however, they are mainly limited to very close domains such as the same or similar programming languages. In this study, we address the aforementioned issues by combing a pattern-based approach with transfer learning to transfer the knowledge acquired for software vulnerability prediction in C source code being used for software vulnerability prediction in Java source code.

3 PROPOSED METHOD

In this section, we discuss our proposed method to develop an ML-based programming language agnostic software vulnerability prediction model. To this end, we trained a CNN model with software vulnerability samples of C source code, and then, we updated trained model to predict vulnerabilities in Java source code. The main phases of the proposed method are discussed in the following sub-sections.

3.1 Preprocessing Source Code

Extracting discriminating features from source code and representing them as numerical vectors is considered the most critical step in training an ML-based software vulnerability prediction model. Converted numerical representation not only needs to preserve the essential information from the source code, but also needs to preserve the semantic relations between the key elements of the source code. Moreover, extracting the optimal feature set improves the vulnerability prediction performance and facilitates training a generic model.

In this phase, we explain the steps we follow to convert an input source code into an ML suitable numerical vector that contains the most important features for vulnerability prediction.

We demonstrate the steps of converting a source code into a numerical vector on the code sample provided below, as it is simple and runnable with both C and Java compilers:

```
int sum(int a,int b){
    return a+b;
```

We convert the input source code into a numerical array through the following steps:

Step-1 (Tokenization): In this step, we first make sure that all non-code related additional parts like comments, tabs, and newlines are removed. Then, we

convert the main code into a stream of tokens. Each token is represented as a sequence of characters that can be treated as a unit in the grammar of the corresponding programming language. For this, we use two lexers developed explicitly for Java and C source code. The extracted tokens from *sum* function are as follow:

```
'int': Keyword, 'sum': Identifier, '(' :
Separator, 'int': Keyword, 'a': Identifier, 'int':
Keyword, 'b': Identifier, ')': Separator, '{':
Separator, 'return': Keyword, 'a': Identifier,
'+': Operator, 'b': Identifier, ';': Separator,
'}': Separator.
```

Step-2 (AST Generation, Normalization, and Simplification): AST contains syntax and semantic information about the source code, and therefore, it is highly useful for code analysis. In this step, we extract the AST of the tokenized code using two parsers developed explicitly for both Java and C source code. Figure 1 presents the Java AST extracted from *sum* function.

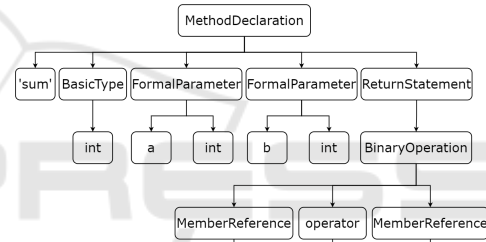


Figure 1: Java AST of "sum" function.

Because Java and C languages have distinct structures, there is no common parser for both languages, and thus generated ASTs are different for each language. Also, the AST structure is flexible and each node in the AST can have multiple attributes/children, however, with this structure, it is hard to follow the relationships between AST nodes after converting the AST to an array. Thus, we normalize and simplify C and Java ASTs to ensure that equivalent ASTs will be converted into the same numerical array by the end of the preprocessing phase. We do this by: first, normalizing the AST tokens and replacing all identifier tokens with assigned abstract representation forms. For example, all method names are replaced with "m_name", and all variable names are replaced with "v_name", etc. Then, considering the importance of structural relations between the AST nodes for vulnerability prediction, we simplify the tokens that represent features (leaves in the AST) rather than independent tokens by combining them with their parent token. In Figure 2 we represent normalized and simplified AST of the function *sum*.

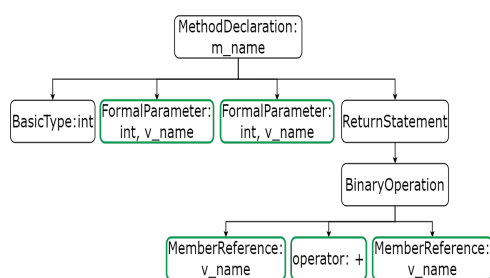


Figure 2: Normalized and simplified Java N-array-AST of "sum" function.

Step-3 (Conversion to a Complete Binary AST): In this step, we convert the AST from normal tree with unbounded number of children per node into a complete binary tree. By this conversion we not only preserve the relations between nodes but we also unify the AST structure for both C and Java. We convert the regular AST (referred as N-array-AST) into a complete binary AST (referred as CB-AST), where all leaves have the same depth, and all internal nodes have exactly two children. The conversion from the N-array-AST to the CB-AST is achieved with respect to the following rules (Bilgin et al., 2020):

- The root of the N-array-AST is the root of the CB-AST.
- The left-child of a node in the CB-AST is the left-most child of the node in the N-array-AST.
- The right-child of a node in the CB-AST is the right sibling of the node in the N-array-AST.
- When the node does not have children, then its left-child is set as NULL.
- When the node is the rightmost child of its parent, its right-child is set as NULL

The Java CB-AST for *sum* function is represented in Figure 3. Additionally, to have a complete tree, NULL children nodes are added until we reach a level when all nodes are NULL.

Step-4 (Encoding to Numerical 3-Tuples): Since majority of ML algorithms expect numerical inputs, we convert the CB-AST into a numerical vector. To this end, we leverage the encoding method presented in (Bilgin et al., 2020) to convert CB-AST nodes into n-tuples which consists of numerical values. In this study, we used 3-tuple of numbers; the first number represents the type of the token, while second and third numbers store additional information about the nodes. For example, the token "BasicType: int, 50" is encoded to (8.0, 103.0, 50.0). This representation is considered as a three-dimensional data structure where each dimension holds a value related to an associated token. Notice that the numeric values

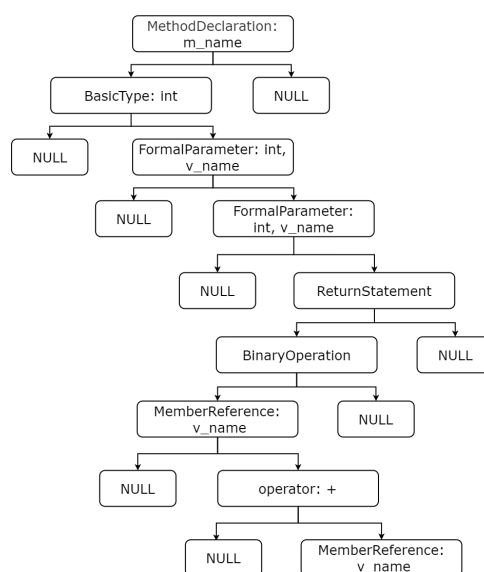


Figure 3: Java binary AST of "sum" function.

used for the encoding mechanism are chosen arbitrarily and could be changed as long as different categories take different values.

Step-5 (Vector Representation): After replacing each node in CB-AST with its 3-tuple numeric value, we transfer the CB-AST into a 1D array. To this end, we used Breadth-first search (BFS) algorithm to position the numeric tuples in their corresponding locations in a 1D array. The 1D array is filled by the CB-AST nodes starting from the root node traversing all the neighbor nodes at the present level then moving on to the nodes at the next depth level and so on. In Figure 4, we present the BFS algorithm passing through the nodes of the first three levels of the Java CB-AST.

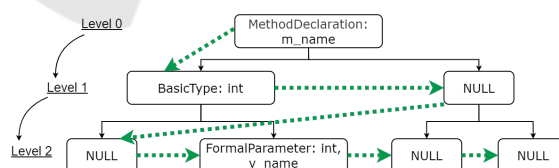


Figure 4: Reading order of the CB-AST nodes.

Proposed approach preserves the structural relations in the source code via respective positional relations between elements in the 1D array. Also, with the rich numerical encoding technique (explained in step-4), the semantic information of the AST are preserved and embedded in the 1D array. Another significant advantage of the proposed numerical array representation is that each element in the 1D array acts as a feature, and hence, it can be directly used as an input to an ML model, which facilitates conducting various automated and intelligent code analysis.

3.2 Syntax Matching and Knowledge Transferring

Analysed tokens are clustered under three main groups: tokens that exist only in C language like CompoundLiteral, tokens that exist only in Java language like ClassDeclaratio, and tokens that exist in both C and Java. Thus, we prepared a dictionary that involves these three categories. However, the tokens that exist in both C and Java are equivalent but different names are assigned due to using different lexers and parser. For example: "FuncDef", "Constant", and "IF" are used by C lexer, whereas, "MethodDeclaration", "BasicType", and "IfStatement" are used by Java lexer. In order to use one model for both Java and C codes, we need to generate the same numerical array for equivalent C and Java codes. Thus, we created a table to match the tokens used by the compiler to construct different statements and expressions in each programming language. We prepared a mapping dictionary following the steps provided in Figure 5 to encode equivalent tokens similarly. After generating the CB-AST, tokens are encoded using the prepared dictionary that contains the equivalent tokens of C and Java with their 3-tuples. In Table 1, examples of common tokens between C and Java are provided along with their numerical 3-tuples.

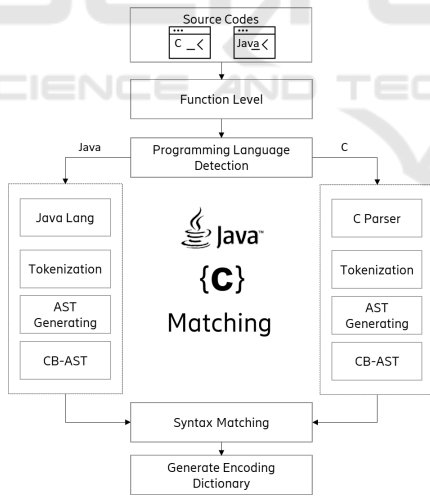


Figure 5: Steps for matching c and Java tokens.

Table 1: Examples of common C and Java tokens.

C Token	Java Token	Numerical tuples
DoWhile	DoStatement	22.0,0.0,0.0
If	IfStatement	30.0,0.0,0.0
Switch	SwitchStatement	38.0,0.0,0.0
TernaryOp	TernaryExpression	39.0,0.0,0.0
While	WhileStatement	44.0,0.0,0.0

3.3 Generic Model Generation

In this study, we aim to overcome the problem of lack of sufficient training data samples to generalize an ML-model to predict a software vulnerability in a source code of a programming language (target). To this end, we train an ML-model for the same vulnerability type using a large number of data samples for another programming language and then transfer the knowledge for vulnerability prediction in the target language, by applying transfer learning method. The main approach we followed in this study (as shown in Figure 6) is based on two main steps: First, we trained and improved a model with a big dataset that includes vulnerability examples written in C. Then, we updated the model with the limited Java examples to predict the vulnerabilities existed in Java codes using the knowledge gained from C codes.

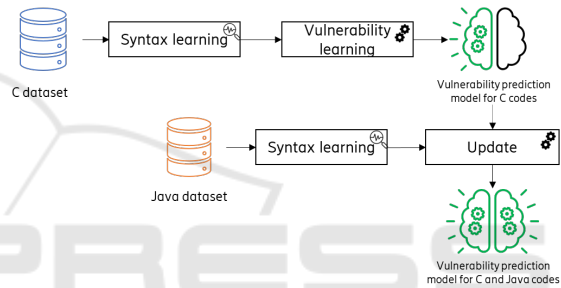


Figure 6: General approach used for vulnerability prediction knowledge transferring.

To automatically detect the important features of vulnerabilities without the need for human intervention, we used CNN architecture. Another advantage of using CNN is to extract the relations and correlations between features by applying convolution between features.

For knowledge transferring, we developed an approach that is different from the common transfer learning approaches (where the input data in the two datasets are similar but the output labels are different). As in our case, we aim to get the same output labels (vulnerable and non-vulnerable) for Java codes leveraging the features and weights learned from C dataset. However, the input Java code instructions have structural and syntax differences and they need to be reflected on their equivalent C code instructions. Thus, as shown in Figure 7, we customized the first steps of our approach (including tokenization, AST generation, and encoding to numerical tuples) to process Java input codes and convert them into numerical arrays peer to the ones extracted from C codes. Then, we use the trained CNN for the feature extraction and classification steps. One of the advantages of this ap-

proach other than knowledge transferring, is that the trained feature extraction and classification layers can be used as they are, and they also can be improved and updated with new data samples (C or Java).

Moreover, to decrease the impact of structural differences between the two programming languages (C and Java), we built the system to analyse source code on function level.

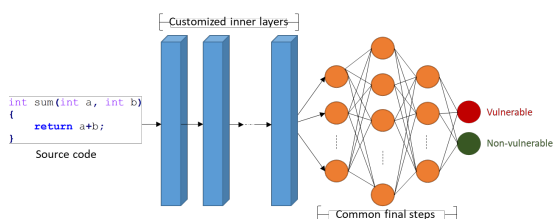


Figure 7: Proposed model for vulnerability prediction knowledge transferring.

4 DATASET

For this study, we used data samples from two datasets; Draper VDISC Dataset (Russell et al., 2018) and Software Assurance Reference Dataset (SARD) (Black, 2018). While the former consists of the vulnerability samples for C source code, the latter includes vulnerabilities for Java source code. A brief summary related to each dataset is provided in the following subsections.

4.1 Draper VDISC Dataset

VDISC dataset involves function-level source code of C and C++ programming languages, labeled by static analysis for potential vulnerabilities. The data is provided in HDF5 files, and raw source code are stored as UTF-8 strings. Five binary 'vulnerability' labels are provided for each function, corresponding to the four most common CWEs as: CWE-120 (3.7% of functions), CWE-119 (1.9% of functions), CWE-469 (0.95% of functions), CWE-476(0.21% of functions), and CWE-other (2.7%of functions).

4.2 SARD Dataset

SARD is a growing collection of test programs with documented weaknesses. Test cases vary from small synthetic programs to large applications. The source of the test cases in SARD dataset are: Wild code sampling (code from known bugs available in industry and open source software), artificial code constructing (codes produced by researchers to cover a wide range of weaknesses), academic code (Code collected from

computer science and programming courses), and automatically generated codes. SARD dataset is provided in terms of test cases that provide the code files of each CWE type. In this paper we used Java source code test cases from Juliet Test Suite.

4.3 Dataset Preparation

The evaluation of the model was carried out by analyzing NULL Pointer Dereference vulnerabilities (CWE-476), as it is a prevalent CWE that provides suitable data for both C and Java programming languages for the purposes of this study. For model training and validation, we obtained C data samples from the VDISC dataset, resulting in a prepared dataset of 120,609 non-vulnerable C functions and 1160 vulnerable C functions. However, the ratio of samples in the vulnerable class to the non-vulnerable class was observed to be significantly imbalanced, with more than 100 non-vulnerable samples for every 1 vulnerable sample, which impacts the model's ability to gain knowledge of vulnerable samples. To address this issue, we down-sampled the dataset to a ratio of 1:4. Additionally, we acquired Java samples from the SARD dataset, and after performing preprocessing steps, we generated a labeled dataset with 616 vulnerable Java functions and 328 non-vulnerable Java functions. Due to the limited number of Java samples in either class, we resorted to using transfer learning to apply the ML model trained on C codes to detect vulnerabilities in both C and Java codes. Finally, the labels were converted to a binary classification format, with 0 indicating non-vulnerable and 1 indicating vulnerable code sample.

5 RESULTS AND DISCUSSION

In this section, we present and discuss the performance result of the proposed method. First, we present the experimental results of the proposed model. Then, we discuss the validity, limitations, and implications of the new work.

5.1 Experimental Results

We used a binary classifier based on a 1D CNN that consists of two convolutional layers, a maxpooling layer, dropout layers, a fully connected layer, and a softmax layer. The CNN architecture is chosen due to its ability to automatically detect the important features without the need for human intervention Also, by applying convolution between features, it extracts the relations and correlations between features. A

dropout layer is applied to avoid over-fitting to the training data.

Since we are training a binary classifier, it is better to train it with better balanced dataset. Thus, we used the down-sampled C dataset for the initial training of the CNN model. On the other hand, unbalanced test set represents the expected real world scenario where there are less functions that could induce vulnerabilities. Therefore, we use the full test set to measure the performance of the model. The sizes and the number of vulnerable and non-vulnerable samples used for training and test steps are shown in Table 2. As could be noted, Java related samples are few and not enough to train a model.

Table 2: Number of vulnerable and non-vulnerable samples used in train and test sets.

Data	Language	Down-sampled	Vulnerable	Non-vulnerable
Train	C	Yes	1160	4640
Train	Java	No	305	167
Test	C	No	140	15,062
Test	Java	No	311	161

To assess the performance of proposed model in vulnerabilities predicting, we analyze accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The confusion table on both C and Java test sets is presented in Table 3. And the evaluating metrics both C and Java source code are provided in Table 4. As could be observed, the developed model is able to detect vulnerabilities in both C and Java source code with accuracy of 99% and 91% respectively. Hence, we transferred the knowledge the model gained from C vulnerable samples to be used to also detect vulnerabilities in Java source code. Moreover, the additional Java samples used to update the model helped to improve the model, as Precision, Recall, F1-score, and AUC percentages have been increased.

Table 3: Confusion tables for C and Java examples (No = non-vulnerable, and Yes = vulnerable).

		Predictions	
		No	Yes
C test set			
Actuals	No	14977	85
	Yes	64	76
Java test set			
Actuals	No	128	33
	Yes	10	301

Table 4: Vulnerability prediction results for both C and Java samples.

Performance	C	Java	Average
Accuracy	0.99	0.91	0.95
Precision	0.47	0.90	0.68
Recall	0.54	0.97	0.75
F1-Score	0.51	0.93	0.72
AUC	0.77	0.88	0.82

5.2 Model Explanation

We utilize Lime (Local Interpretable Model-agnostic Explanations) method to explain and interpret the predictions of our proposed model. We aim to examine the effect of our code representation method on the classification model for each programming language. For this aim, we apply Lime on both C and Java test sets separately, and analyze the features used to differentiate vulnerable source code for each language. Analysing results of Lime are presented in Figure 8. We can notice that the most effective features (encoded values) used for differentiating vulnerable from non-vulnerable source code are the same for both programming languages. For example, the value 1277, 1280, 1219, 641, and 317 have similar effect on Java and C source code. This proves that our method have successfully converted the source code from both programming languages (C and Java) into unique, equivalent, simple, ML friendly numeric vector. In addition, it shows that proposed model is generic and can be used to detect trained vulnerabilities in other programming languages once same code representation method is applied.

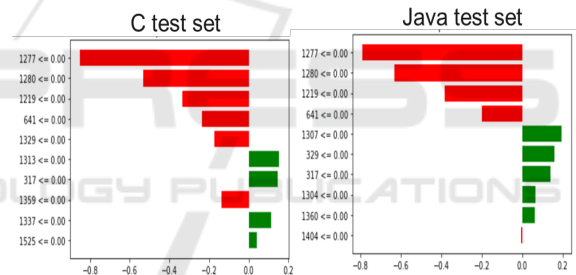


Figure 8: Local explanation of vulnerable examples detection in C and Java test sets.

5.3 Threats to Validity

Data Insufficiency: The limited availability of data for both C and Java source code may impact the representativeness of the results. To address this, data from two different datasets were used. However, this could lead to potential biases in the results if the data sets are not fully representative of real-world applications.

Preprocessing: The preprocessing steps involved matching the syntax between C and Java tokens, which may have introduced errors or biases due to the structural differences between the two languages.

Model Training: To address the lack of labeled source code from real projects, synthetic codes from the SARD and VDISC datasets were used for training the model. This could affect the validity of the

results, as the model may not generalize well to real-world applications.

6 CONCLUSION AND FUTURE WORK

In this study, we focus on the problem of having vulnerability samples for some programming languages but not others. To overcome this problem, we design a method that extract vulnerability prediction knowledge from available data samples and then use it to predict vulnerabilities in another programming language. We also, add flexibility to update the model once new samples are provided. Specifically, in this study, we built a model that is able to detect vulnerabilities in both Java and C source code. We trained a CNN-based model with C source code from VDISC dataset. Then, we modified the model to detect the learned vulnerabilities in Java source code. We extracted Java sample codes from SARD dataset. By the end of our experiments, we were able to show that despite the many differences between programming languages, we were able to train one model to detect vulnerabilities in more than one programming language. This study could be further extended to detect vulnerabilities in other commonly used programming languages such as Python and Javascript. The study could be also improved by training the model on other common vulnerability types from different programming languages.

ACKNOWLEDGEMENTS

This work was funded by The Scientific and Technological Research Council of Turkey, under 1515 Frontier R&D Laboratories Support Program with project no: 5169902.

REFERENCES

- Bilgin, Z., Ersoy, M. A., Soykan, E. U., Tomur, E., Çomak, P., and Karaçay, L. (2020). Vulnerability prediction from source code using machine learning. *IEEE Access*, 8:150672–150684.
- Black, P. E. (2018). A software assurance reference dataset: Thousands of programs with known bugs. *Journal of research of the National Institute of Standards and Technology*, 123:1.
- Coskun, T., Halepmollasi, R., Hanifi, K., Fouladi, R. F., De Cnudde, P. C., and Tosun, A. (2022). Profiling developers to predict vulnerable code changes. In *Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering*, pages 32–41.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Duan, X., Wu, J., Ji, S., Rui, Z., Luo, T., Yang, M., and Wu, Y. (2019). Vulsniper: Focus your attention to shoot fine-grained vulnerabilities. In *IJCAI*, pages 4665–4671.
- Halepmollasi, R., Hanifi, K., Fouladi, R. F., and Tosun, A. (2023). A comparison of source code representation methods to predict vulnerability inducing code changes. In *ENASE*, page In Press.
- Hanif, H., Nasir, M. H. N. M., Ab Razak, M. F., Firdaus, A., and Anuar, N. B. (2021). The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches. *Journal of Network and Computer Applications*, 179:103009.
- Kalouptsoglou, I., Siavvas, M., Kehagias, D., Chatzigeorgiou, A., and Ampatzoglou, A. (2022). Examining the capacity of text mining and software metrics in vulnerability prediction. *Entropy*, 24(5):651.
- Li, X., Wang, L., Xin, Y., Yang, Y., Tang, Q., and Chen, Y. (2021). Automated software vulnerability detection based on hybrid neural network. *Applied Sciences*, 11(7):3201.
- Lin, G., Wen, S., Han, Q.-L., Zhang, J., and Xiang, Y. (2020). Software vulnerability detection using deep neural networks: a survey. *Proceedings of the IEEE*, 108(10):1825–1848.
- Lin, G., Zhang, J., Luo, W., Pan, L., De Vel, O., Montague, P., and Xiang, Y. (2019). Software vulnerability discovery via learning multi-domain knowledge bases. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2469–2485.
- Palit, T., Moon, J. F., Monroe, F., and Polychronakis, M. (2021). Dynpta: Combining static and dynamic analysis for practical selective data protection. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1919–1937. IEEE.
- Russell, R., Kim, L., Hamilton, L., Lazovich, T., Harer, J., Ozdemir, O., Ellingwood, P., and McConley, M. (2018). Automated vulnerability detection in source code using deep representation learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 757–762. IEEE.
- Zhou, Y., Liu, S., Siow, J., Du, X., and Liu, Y. (2019). Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. *Advances in neural information processing systems*, 32.
- Ziems, N. and Wu, S. (2021). Security vulnerability detection using deep learning natural language processing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE.