World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:04, 2023

Profiling Risky Code Using Machine Learning

Authors : Zunaira Zaman, David Bohannon

Abstract : This study explores the application of machine learning (ML) for detecting security vulnerabilities in source code. The research aims to assist organizations with large application portfolios and limited security testing capabilities in prioritizing security activities. ML-based approaches offer benefits such as increased confidence scores, false positives and negatives tuning, and automated feedback. The initial approach using natural language processing techniques to extract features achieved 86% accuracy during the training phase but suffered from overfitting and performed poorly on unseen datasets during testing. To address these issues, the study proposes using the abstract syntax tree (AST) for Java and C++ codebases to capture code semantics and structure and generate path-context representations for each function. The Code2Vec model architecture is used to learn distributed representations of source code snippets for training a machine-learning classifier for vulnerability prediction. The study evaluates the performance of the proposed methodology using two datasets and compares the results with existing approaches. The Devign dataset yielded 60% accuracy in predicting vulnerable code snippets and helped resist overfitting, while the Juliet Test Suite predicted specific vulnerabilities such as OS-Command Injection, Cryptographic, and Cross-Site Scripting vulnerabilities. The Code2Vec model achieved 75% accuracy and a 98% recall rate in predicting OS-Command Injection vulnerabilities. The study concludes that even partial AST representations of source code can be useful for vulnerability prediction. The approach has the potential for automated intelligent analysis of source code, including vulnerability prediction on unseen source code. State-of-the-art models using natural language processing techniques and CNN models with ensemble modelling techniques did not generalize well on unseen data and faced overfitting issues. However, predicting vulnerabilities in source code using machine learning poses challenges such as high dimensionality and complexity of source code, imbalanced datasets, and identifying specific types of vulnerabilities. Future work will address these challenges and expand the scope of the research.

Keywords : code embeddings, neural networks, natural language processing, OS command injection, software security, code properties
Conference Title : ICMLC 2023 : International Conference on Machine Learning and Cybernetics
Conference Location : Istanbul, Türkiye
Conference Dates : April 24-25, 2023