World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:01, 2023

Feature Engineering Based Detection of Buffer Overflow Vulnerability in Source Code Using Deep Neural Networks

Authors : Mst Shapna Akter, Hossain Shahriar

Abstract : One of the most important challenges in the field of software code audit is the presence of vulnerabilities in software source code. Every year, more and more software flaws are found, either internally in proprietary code or revealed publicly. These flaws are highly likely exploited and lead to system compromise, data leakage, or denial of service. C and C++ open-source code are now available in order to create a largescale, machine-learning system for function-level vulnerability identification. We assembled a sizable dataset of millions of opensource functions that point to potential exploits. We developed an efficient and scalable vulnerability detection method based on deep neural network models that learn features extracted from the source codes. The source code is first converted into a minimal intermediate representation to remove the pointless components and shorten the dependency. Moreover, we keep the semantic and syntactic information using state-of-the-art word embedding algorithms such as glove and fastText. The embedded vectors are subsequently fed into deep learning networks such as LSTM, BilSTM, LSTM-Autoencoder, word2vec, BERT, and GPT-2 to classify the possible vulnerabilities. Furthermore, we proposed a neural network model which can overcome issues associated with traditional neural networks. Evaluation metrics such as f1 score, precision, recall, accuracy, and total execution time have been used to measure the performance. We made a comparative analysis between results derived from features containing a minimal text representation and semantic and syntactic information. We found that all of the deep learning models provide comparatively higher accuracy when we use semantic and syntactic information as the features but require higher execution time as the word embedding the algorithm puts on a bit of complexity to the overall system.

Keywords : cyber security, vulnerability detection, neural networks, feature extraction
Conference Title : ICMLCS 2023 : International Conference on Machine Learning in Computer Security
Conference Location : New York, United States
Conference Dates : January 30-31, 2023