IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 2, May 2025

Impact Factor: 7.67

Software Vulnerability Testing using Borderline SMOTE

M. Vara Lakshmi[1], Mohammed Aqeel Yousuf[2], A Sreehan Chary[3]

Assistant Professor, Department of IT[1]
B.Tech Student, Department of IT[2,3]
Mahatma Gandhi Institute of Technology, Hyderabad, India

Abstract: *Enhancing the efficiency of software vulnerability detection has become increasingly important as software systems grow and complexity. In this study, we propose a comprehensive and innovative framework that brings together four essential techniques—Source Embedding, Feature Learning, Data Resampling, and Classification—to tackle the critical challenges of identifying vulnerabilities in source code. The process begins with Source Embedding, where Abstract Syntax Trees (ASTs) generated using the Joern tool are analysed with advanced data mining techniques to capture both the structural and semantic characteristics of the code. These embedded representations form the basis for Feature Learning, which applies machine learning and deep learning algorithms to extract meaningful insights from the AST nodes, enabling the detection of both known vulnerabilities and subtle, hidden code anomalies. Recognizing the issue of data imbalance in real-world datasets, we integrate the Borderline SMOTE algorithm to generate synthetic examples near class boundaries, helping to create a more balanced and representative dataset. With this enriched dataset, the Classification stage leverages robust models trained to accurately identify potential vulnerabilities. What sets our approach apart is the seamless integration of varied techniques, which allows us to discover intricate patterns that traditional static analysis tools often miss. We validate our framework using the Verum dataset, where it delivers outstanding results across multiple performance metrics including precision, recall, F1-score, and accuracy. The findings affirm the model's capability to deliver reliable predictions while reducing false positives and negatives. This holistic methodology not only raises the bar for source code vulnerability detection but also lays a solid foundation for building more secure and resilient software. By addressing the core aspects of code representation, feature extraction, and imbalanced learning, our study contributes significantly to the development of smarter, more adaptive security tools for modern development environments.*

Keywords: Abstract Syntax Tree (AST), Source Embedding, Feature Learning, Borderline SMOTE, Classification

DOI: 10.48175/IJARSCT-26242