LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY, INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

16th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 27–29, 2025

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

VILNIUS UNIVERSITY PRESS
Vilnius, 2025

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. (+370 5) 2109 315

https://doi.org/10.15388/DAMSS.16.2025
ISBN 978-609-07-1200-9 (digital PDF)

Improving Malware Detection by Analyzing Similarities of Multi-Category Benign Software

Juozapas Rokas Čypas, Juozas Dautartas, Olga Kurasova, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

juozapas.cypas@mif.vu.lt

In today's digital world, the importance of cybersecurity is increasing rapidly. The evolution of technology and AI enables various threat actors to evolve malware as well as the methods of evading malware detection. In this study, we aim to analyze modern malware evasion methods presently used in the wild. It's important to identify such methods so that they can be analyzed and studied by security researchers for the purpose of improving the defense infrastructure of software systems. Traditionally, machine learning based analysis of Windows Portable Executable (PE) file static features uses datasets that have either two classes (benign or malicious), or multiple malware classes (e.g., worms, Trojans, ransomware, spyware), and one benign software class. Our proposed method and dataset (DOI:10.18279/MIDAS.265677) focus on the analysis of multiple categories of benign software (office tools, security, media, etc.) and just one class of malware. These categories are used to train a classifier that can distinguish benign software based on its static features. Our concept is that for a given malware sample, it is possible to identify a benign category corresponding to the lowest expected detection rate. This approach analyzes the similarity between each malware sample and multiple categories of benign software. For each instance of malware, we identify the closest benign category and then inject the most characteristic static features of that benign cluster into the malware sample to trick the classifier and maximize the evasion rate. Preliminary results indicate that, after injecting features from the category of benign files selected based on similarity, the initial classifier demonstrates a decrease in the detection rate of such concealed malware in comparison to the original malware file.

Acknowledgments. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-116.