CYBER SECURITY COOPERATIVE RESEARCH CENTRE

SCReeD Dataset Declaration Form

Dataset name:* Malicious Package Metadata Dataset

Dataset version:* 1

Dataset URL:* https://github.com/CSCRC-SCREED/CSU-Malicious-Package-Metadata-Dataset

Creation date:* 1 July 2023

Last update:* 14 October2024

Author(s):* Sajal Halder, Michael Bewong, Arash Mahboubi, Yinhao Jiang, Md Rafiqul Islam, Md Zahid Islam, Ryan HL Ip, Muhammad Ejaz, Gowri Sankar, Muhammad Ali Babar, Oscar Blessed Deho

Author(s) affiliation(s): Charles Sturt University, Queensland University of Technology, CSIRO- Data61, University of Adelaide

Author contact(s):* sajal.halder@data61.csiro.au, mbewong@csu.edu.au, amahboubi@csu.edu.au, yjiang@csu.edu.au, Ali.babar@adelaide.edu.au, g.ramachandran@qut.edu.au, ejaz.ahmed@data61.csiro.au, hoip@csu.edu.au, zislam@csu.edu.au, mislam@csu.edu.au, odeho@csu.edu.au

Keywords:* Metadata, Malicious Package, Node Package Manager

Description/background:* The dataset consists of engineered features extracted from the metadata of software packages, designed to help identify whether a package is benign or malicious. These features capture various aspects of the package, including version details, description length, license presence, repository information, and dependency structure. Together, these features provide a rich set of metadata insights that can be used for machine learning models to predict potentially malicious packages and secure software supply chains.

Dataset funding: Click or tap here to enter text.

Attribute details:* version_major, version_minor, version_patch, version_num_segments, version_is_pre_release, version_pre_release_label, description_word_count, description_length, license_exist, license_length, repository_exist, funding_exist, exports_exist, main_exist, main_file_extension, main_path_depth, types_exist, types_file_extension, types_path_depth, engines_exist, node_version_constraint_complexity, min_node_version, scripts_exist, num_scripts, has_test_script, keywords_exist, num_keywords, rare_keyword_count,

num_dev_dependencies, homepage_exist, homepage_domain, homepage_type, _nodeVersion_exist, _nodeVersion_major, author_type, author_contact_complete, author_email_domain, contributors_exist, num_contributors, contributors_info_complete, num_dependencies, version_flexibility, maintainers_exist, num_maintainers, repository_type, repository_domain

Intended target (if specified): Click or tap here to enter text.

Format:* csv

License:* Open

Standard compliance: Click or tap here to enter text.

Type:* Numerical and Categorical

Size:* 5.2 MB

Availability: Click or tap here to enter text.

Data status: Clean data

Data provenance:* The data contained in this dataset consists of primary metadata collated from existing opensource software packages including existing benign and malicious software packages. Benign software metadata information was collated from the Node Package Manager (NPM) repository by considering popular NPM packages. while malicious software metadata was collated from historically known malicious packages archived on Github (i.e. https://dasfreak.github.io/Backstabbers-Knife-Collection). Please see published paper (https://dl.acm.org/doi/10.1145/3589334.3645543) to learn more about the assumptions made about the dataset

Source computing infrastructure:* Live system

Accompanying program(s)/script(s): Click or tap here to enter text.

Software installer or VM for replication: Click or tap here to enter text.

Generated or captured via:* Software

Category/categories:* ML training data

Published in: Click or tap here to enter text.

Open research question(s) (if any): Click or tap here to enter text.

Potential use case(s) or application area(s): Click or tap here to enter text.

Data access control:* Global access

Data retention period:*	Permanent
Data validation/checksum:*	CSU Team
GDPR compliance:*	Yes
Consent:	Click or tap here to enter text.
Ethics approval:	Choose an item.
Ethics considerations:	Click or tap here to enter text.

* *denotes mandatory field*

☒ I confirm that I have read, understood, and agreed to the submission guidelines, policies, and submission declaration.

☒ I confirm that the contributors of the dataset have no conflict of interest to declare.

☒ I agree to take public responsibility for my dataset's contents.

mbewong

Signature of Corresponding Author Date: 28/October/2024

(signed on behalf of all contributors)