Data sets

Data source: The data we are using are originally labelled by CVE and NVD which are two authorized Publicly available vulnerability data repository. And then collected by the group represented by Danial Lin. This data sets are widely used in this team and cited many times by other group and papers.

Data features:
Open source: Everyone has the access to the data sets and do the validation to their own models
From real-world projects: the performance of the models trained by this data sets would be more representative when dealing with the real world projects.

Data structure: Containing two different domains: multimedia (FFmpeg) and image (LibPNG, LibTIFF) applications which is related to our project topic: Cross domain vulnerability detection

FFmpeg:	Libpng:	Libtiff:
Non-Vuln: 5563	Non-Vuln: 577	Non-Vuln: 731
Vuln: 191	Vuln: 44	Vuln: 94

Datasets cited many time across the vulnerability detection field:
Datasets:
FFmpeg: https://ffmpeg.org/
10.1109/TDSC.2021.3051525
10.1109/tdsc.2019.2954088
10.1007/978-3-030-47426-3_54
10.1145/3133956.3138840
…
LibPNG:
10.1109/tdsc.2019.2954088
10.1007/978-3-030-47426-3_54
10.1145/3133956.3138840
…

LibTIFF:
10.1109/TII.2019.2942800
10.1109/tdsc.2019.2954088
10.1007/978-3-030-47426-3_54
10.1145/3133956.3138840
…