Contribution ID: **32** Type: **not specified**

The influence of dataset bias on the efficiency of the Intrusion Detection Systems

Thursday 18 April 2024 09:30 (12 minutes)

Internet intrusion detection systems (IDS) use machine learning models, which one needs to train using public datasets. The training process requires a training set, which is a majority part of such a dataset, while validation is performed on its second part - the validation set. Finally, to evaluate the quality of the output model, one utilizes the test set, which is the third part. The measure (accuracy, precision, recall, or F1) obtained on the latter determines the quality of the model. In our paper, we investigate how the model prepared in the above classic way performs on other data, i.e., to what extent the model is biased to the public dataset used in the IDS model preparation. Our investigation uses cross-validation of models based on four internet traffic datasets: UNSW-NB15, BoT-IoT, ToN-IoT, and CIC-CSE-IDS2018. The results obtained show that quality measures of a model trained on one public dataset are only partially repeatable on others. It confirms the necessity of careful selection of data used in the machine learning models in IDS that guarantee high data diversity.

Author: IWANOWSKI, Marcin

Co-author: PELC, Franciszek (Warsaw University of Technology)

Presenter: PELC, Franciszek (Warsaw University of Technology)

Session Classification: Session A (Presentation)