Contribution ID: **478**

Type: **not specified**

Optimising Synthetic Data Generation for Cybersecurity Datasets

Tuesday, 2 December 2025 14:00 (30 minutes)

In recent years, cybercrimes have become more prevalent and impactful for all users of modern technology. Consequently, various artificial-intelligence-driven intrusion detection software have been implemented to detect and prevent such cyberattacks. Some well-known tools include Microsoft's Security Copilot and SentinelOne's Singularity. However, such AI tools are of-ten difficult to train and maintain, primarily because of the lack of available cybersecurity datasets. Furthermore, even when real cybersecurity datasets are collected, they may lack balance, reliability, and variety, making them inefficient for training AI intrusion detection tools. A trending solution to this predicament is synthetic data generation, particularly for meeting commercial cyber-security dataset requirements. However, synthetic data generation simply mimics the structure and content of real datasets and often reproduces the poor characteristics of the real datasets. Therefore, this study proposes the inclusion of Data Quality Metrics and data optimisation techniques during the synthetic data generation process to improve the quality of synthetic cybersecurity datasets. At the pinnacle of this research, an optimal process for producing synthetic datasets for cybersecurity research is proposed.

Presenting Author

Email

Student or Postdoc?

Institute

Registered for the conference?

CHPC User

Primary author: DEVRAJ, Christian K.

Presenters: DEVRAJ, Christian K.; ELOFF, Jan (University of Pretoria)

Session Classification: ISSA