**Title:** Synthetic Data Generation with Machine Learning for Network Intrusion Detection Systems

**Authors:** Marvin Newlin, Mark Reith, Mark DeYoung

**Abstract:** Machine learning is becoming an integral part of cybersecurity today, particularly in the area of network anomaly detection. However, machine learning techniques require large volumes of data to be effective. Although there are some datasets available for training Network Intrusion Detection Systems (NIDS), many of them are outdated or do not contain enough useful information for training/classification of NIDS. Therefore, generating synthetic data that is realistic is imperative for training effective intrusion detection systems. Currently, the most common methods for generating synthetic data are simulation or emulation through a software package like OPNET, and then machine learning is used to analyze the dataset for correctness. This paper argues for an approach to utilize machine learning to develop models in order to generate the datasets themselves for NIDS, which is an approach that is not commonly used. In this paper, we discuss some of the well-known available datasets, the features that make up a good dataset, the reasons for utilizing generative modeling to synthesize network data, and layout a basic approach to developing generative models for synthetic data by leveraging machine learning.

**Key Words:** Synthetic Data Generation, Network Intrusion Detection System, Machine Learning