Reproduction Plan

Research Question: The paper we're reproducing focuses on anomaly detection in network traffic using a dataset from the CESNET3 network. The research question revolves around comparing the performance of anomaly detection methods, such as Isolation Forest, in identifying network anomalies.

Dataset: The dataset, CESNET-TimeSeries24, contains time series data of network traffic (from over 275,000 IP addresses). It includes metrics like number of flows, packets, bytes, and various ratios. Given the large size of the dataset (over 275,000 time series entries and petabytes of data), computational limitations, especially RAM, are a concern. To mitigate this, a smaller, representative subset of the data was used in the analysis.

Tools and Methodology:
Isolation Forest was used for anomaly detection.

Metrics: Includes the number of flows, packets, bytes, as well as ratios like UDP/TCP, directionality of traffic, and flow durations.

Preprocessing: Due to the dataset's size, a subset of the data was selected, which means the results differed from the paper's findings.

Challenges:

Computational Constraints: The dataset's large size required using a subset to avoid memory overload.

Data Preprocessing: We're working with pre-aggregated time series data (10-minute, 1-hour, and 1-day windows), and we need to account for any filtering, normalization, or gaps in time series that may affect model performance.

Parameter Selection: Tuning parameters for the Isolation Forest algorithm (e.g., number of estimators, contamination level) could lead to slight differences in results.