Anomaly Detection in Network Traffic using Multivariate State Machines

Vasileios Serentellos (TU Delft)
Sicco Verwer (TUD) Christian Hammerschmidt (TUD)

Goal: Profile and Detect Hosts by Behavior

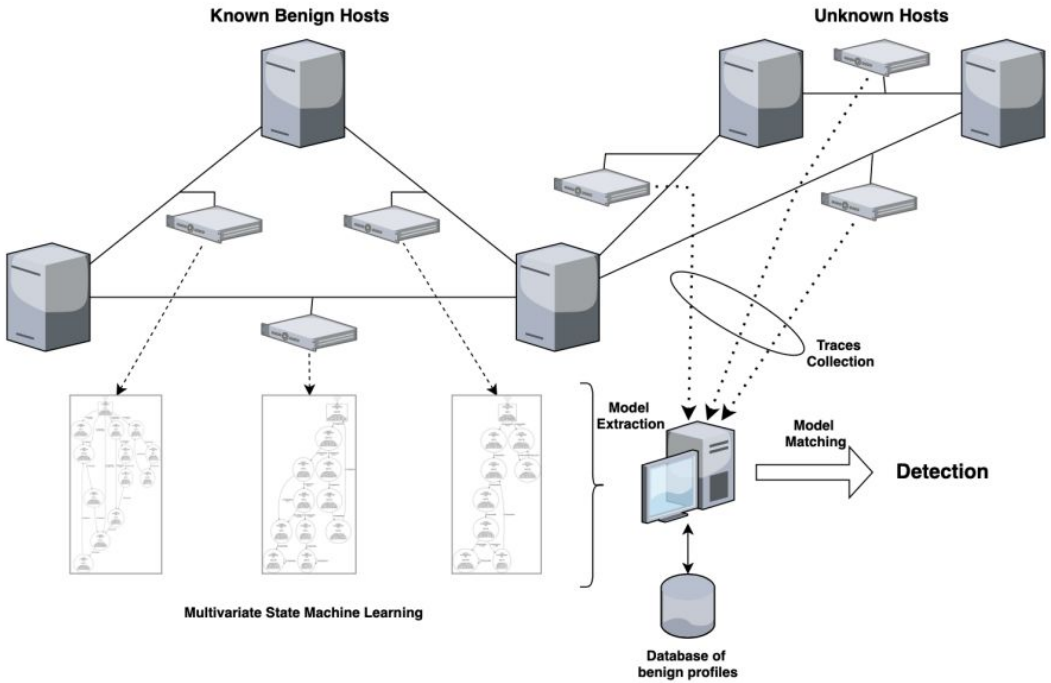

Figure 1.1: High level illustration of the detection process on a virtual network topology

Goals: Profile and Detect Hosts by behavior

Build profiles of known parts of the network and match behavior with traffic from unknown/outside sources.

Constraints:

- Only use NetFlow data, no host/network-level information, no threat intelligence

Approach:

- Cluster flows using a stateful sequence model, identify outliers in each cluster

Contribution

- A effective combination of state-machine learning for sequence clustering with classical outlier detection algorithms
- An application for robust traffic classification for malware detection including a comparison with BotFP

Multivariate Finite State Machines

Sequential model used in our work:

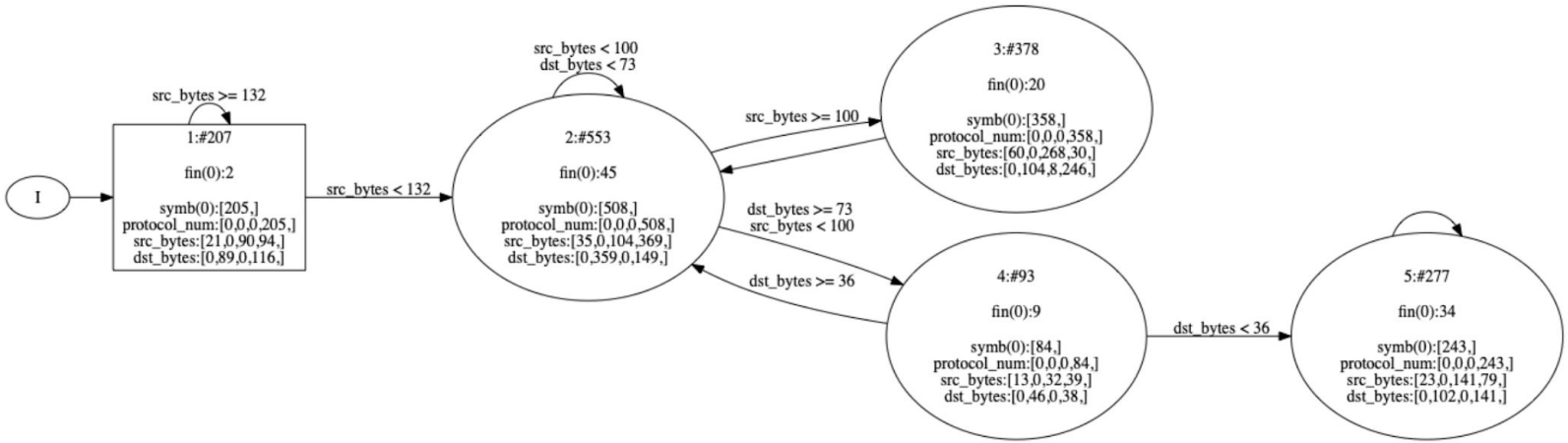

Figure 5.3: Example of a multivariate model extracted from NetFlows with 5 states and 3 attributes in each state

Pipeline: Overview

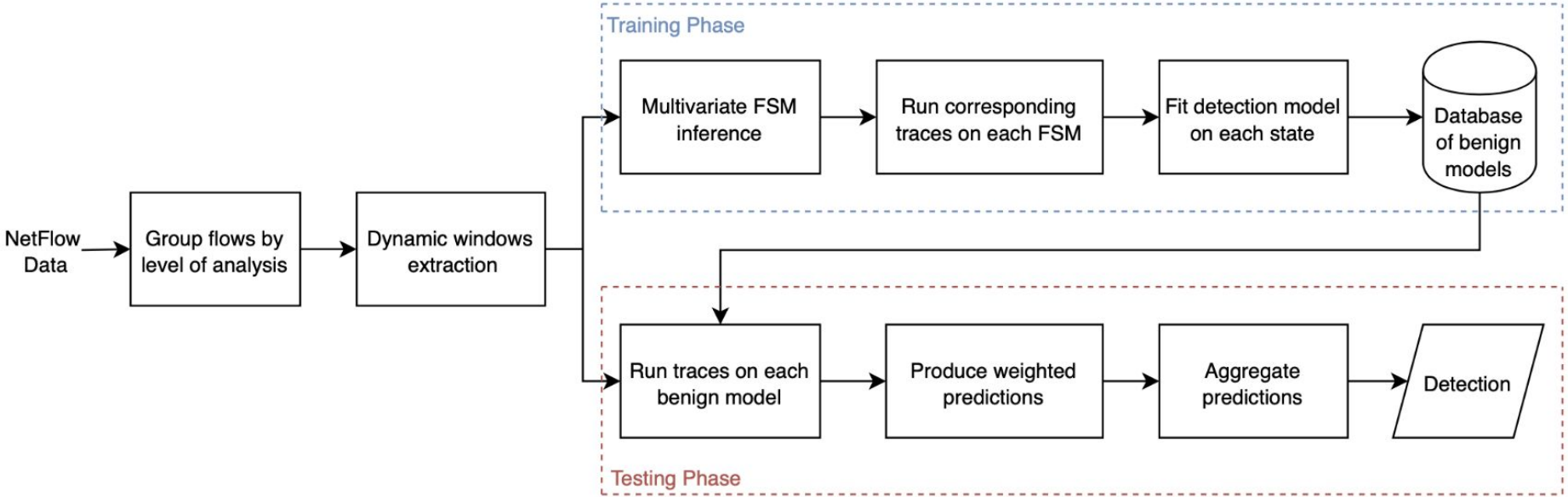

Figure 6.1: High level flowchart of the detection pipeline

Pipeline: Training

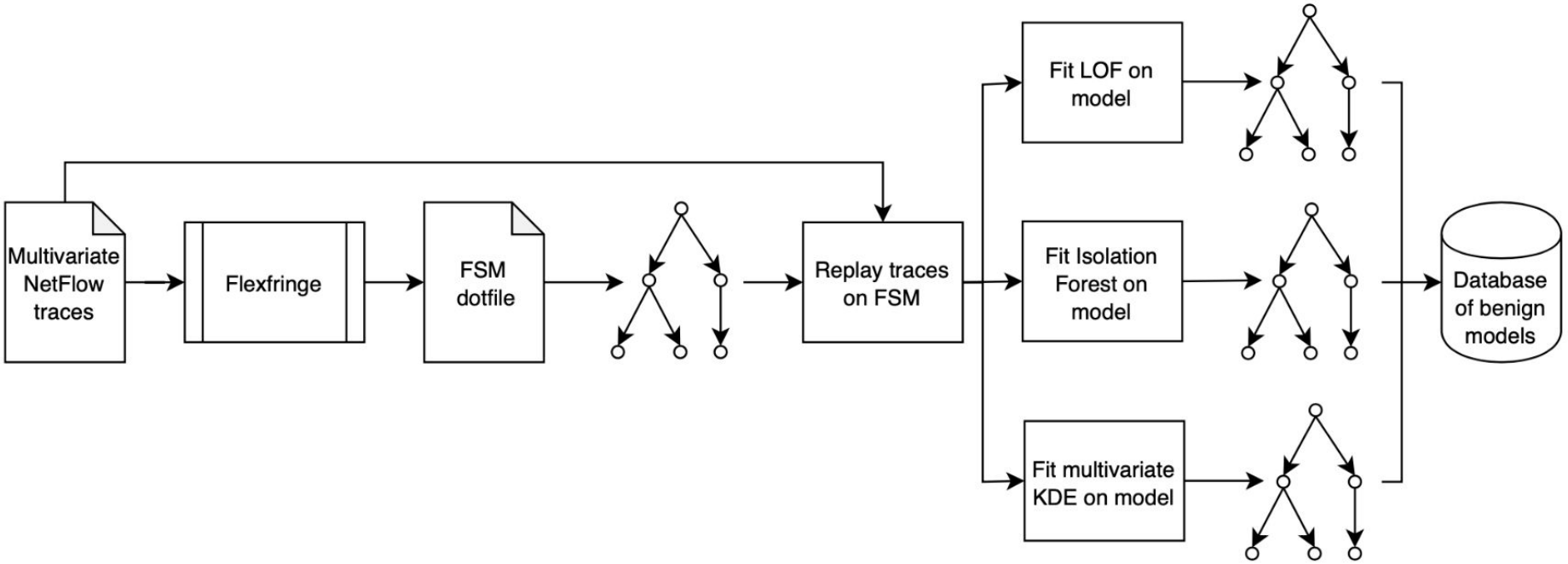

Figure 6.2: High level flowchart of the training pipeline

Pipeline: Testing

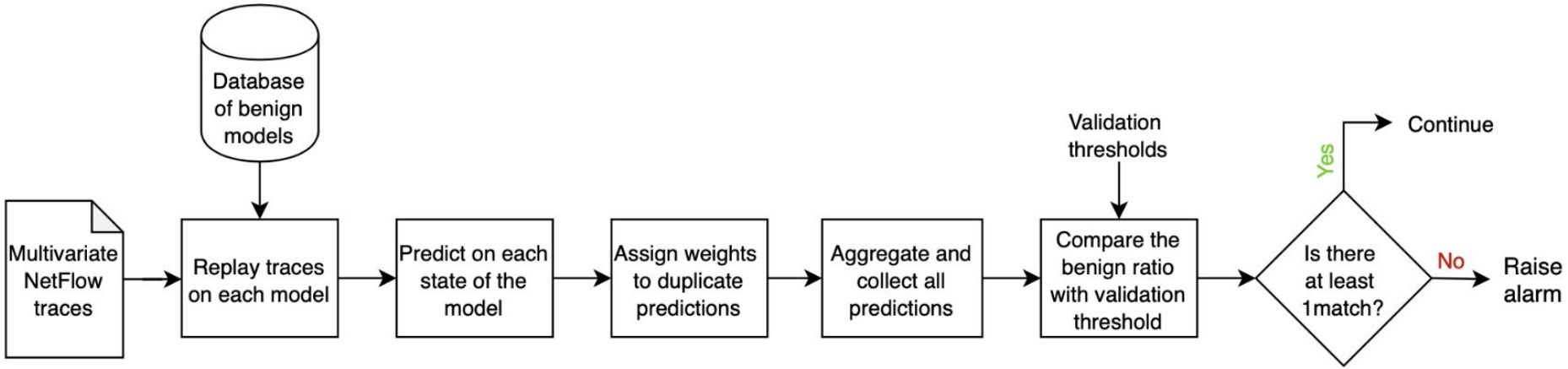

Figure 6.3: High level flowchart of the testing pipeline

Experiments

- Experiments on three well-known open datasets
- Evaluation of different features and outlier detection methods
- Comparison with state-of-the-art classifier (BotFP)

Dataset	Training set	Test set
CTU-13	Benign flows from Scenario 3	Remaining Scenarios + Malicious flows from Scenario 3
UNSW-NB15	Benign flows from Scenario 1	Remaining Scenarios + Malicious flows from Scenario 1
CICIDS2017	Monday	Rest Days

Table 7.1: Training and Test sets split for each dataset

Results: CTU13 Dataset

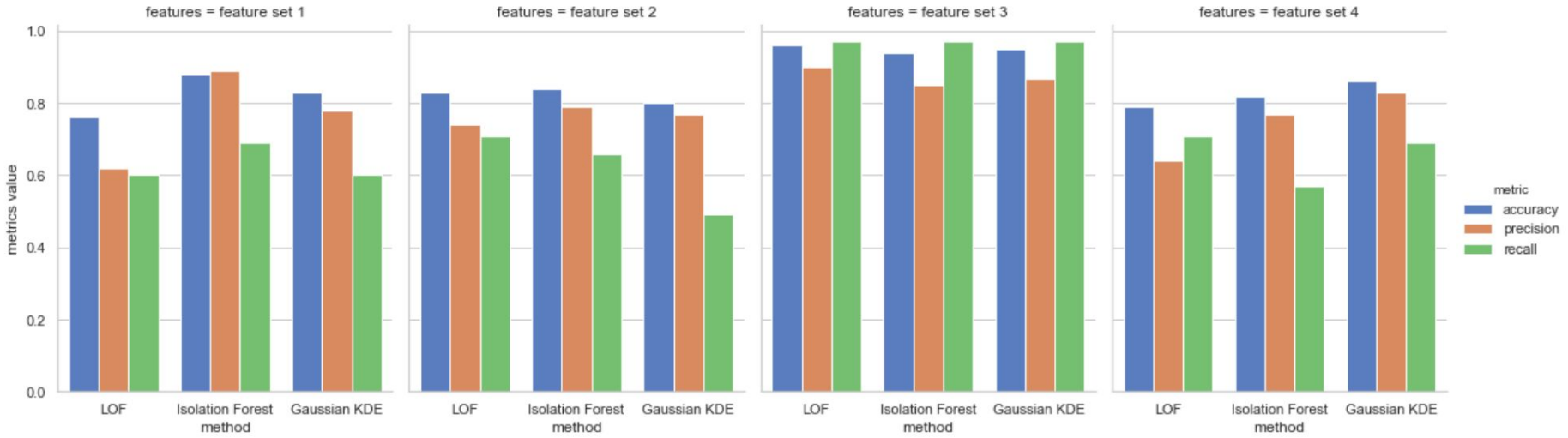

Figure 7.2: Aggregated evaluation metrics per detection method and feature set for the CTU-13 dataset with all hosts included

Results: CUT13 (Comparison with BotFP/SotA)

Scenario	Best Multivariate				BotFP			
	TP	TN	FP	FN	TP	TN	FP	FN
1	1	166	0	0	1	163	3	0
2	1	131	0	0	1	131	0	0
6	1	111	0	0	1	111	0	0
8	1	167	3	0	1	165	5	0
9	10	133	1	0	10	133	1	0

Confusion results

Method	1	2	6	8	9
Best Multivariate	1	1	1	0.98	0.99
BotFP	0.98	1	1	0.97	0.99

Accuracy results

Table 7.11: Comparative results between the designed system and BotFP on the CTU-13 dataset

Limitations

- Misclassified hosts had mixtures of benign/malicious traffic with a minority of traffic being malicious
- Approach requires enough sequential information to be effective