# Enhancing Cybersecurity with Machine Learning: Beaconing Detection in PCAP Data


Graduate

**Anastasiia Graftceva**


Definition of Task: The objective of this study is to develop a model capable of detecting beaconing activity within network traffic capture (PCAP) files. Although not designed for real-time detection, this model is intended for integration into existing Network Detection and Response (NDR) or Security Orchestration, Automation and Response (SOAR) systems for analyzing post-capture files. It combines machine learning (ML) algorithms and neural networks (NN) to address the challenge of detecting beaconing, which involves repetitive, periodic communication often used by malicious actors to maintain connections with compromised systems. The model includes a Histogram Gradient Booster Classifier (HGBC) for binary classification of PCAP data features and a Long Short-Term Memory (LSTM) neural network for analysis of temporal dependencies in the communication patterns. This combination aims to achieve higher detection accuracy and demonstrates resilience against ever evolving tactics of cyber threats.

Approach / Technology: The approach applies advanced ML algorithms and NN, along the collection of network traffic data from various sources for model training and evaluation. Malicious data, primarily from botnet traffic involving DNS and HTTP protocols, highlights repetitive beaconing patterns. For benign activity, data from IoT-enabled environments, both open-sourced and private, was used to develop a dataset of regular and semi-automated non-malicious communication. Initially, raw PCAP data is processed to extract unique packet flows and structured into a format where relevant metrics indicative of beaconing are computed. The HGBC then analyzes this data to identify potential malicious patterns, and these results are processed by the LSTM to examine temporal relationships. During testing, the model was challenged with both benign and malicious files from a simulated cyber attack environment, ensuring it was tested against realistic scenarios. This approach led to the development of a highly accurate beaconing detection framework, achieving a 99.37% accuracy rate on a modestly sized training set, and demonstrating advanced cybersecurity threat detection capabilities.

Conclusion: Integrating this sequenced ML and NN model with existing cybersecurity systems such as SIEM, IDS, NDR, SOAR and firewalls can significantly increase their ability to identify and mitigate network threats. The model's modularity allows for application across different network environments. Future enhancements should include expanding the dataset with more diverse data collected over extended periods, to capture a comprehensive view of network activity. Continuous data recording at different times will also enrich the dataset, reflecting a broader range of network behaviors and potential threats. Additionally, the developed framework shows promise for detecting various network traffic patterns, which could extend its use beyond beaconing detection. Adapting the model to recognize different network anomalies could enable NDR systems to identify a wider array of malicious activities, such as unusual data exfiltration and command and control communications. Integrating this model could improve incident response times, reduce undetected breaches, and enhance overall network security. Future research should focus on dataset augmentation and extending the framework's capabilities to ensure even greater protection and resilience in network security.

**HGBC-LSTM results on unseen data**
Own presentment

file2.csv

```
Statistics:
Total sequences: 362
Malicious sequences: 48
Percentage of malicious sequences: 13.26%

Sequence 249: Benign
Sequence 250: Malicious
Sequence 251: Malicious
Sequence 252: Benign
Sequence 253: Benign
Sequence 254: Benign
Sequence 255: Benign
Sequence 256: Benign
Sequence 257: Benign
Sequence 258: Malicious
```

file3.csv

```
Statistics:
Total sequences: 47
Malicious sequences: 0
Percentage of malicious sequences: 0.00%

Sequence 5: Benign
Sequence 6: Benign
Sequence 7: Benign
Sequence 8: Benign
Sequence 9: Benign
Sequence 10: Benign
Sequence 11: Benign
Sequence 12: Benign
Sequence 13: Benign
Sequence 14: Benign
```

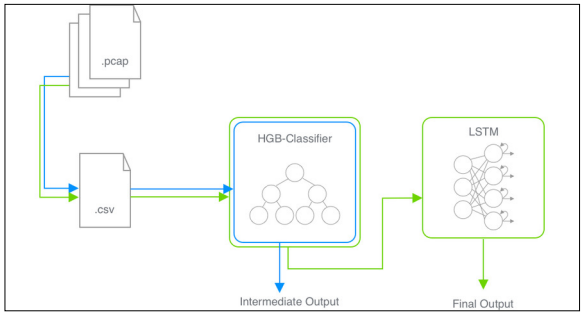
**HGBC-LSTM accuracy report**
Own presentment

```
Training Summary:
Total epochs: 10

Epoch-wise Loss and Accuracy:
Epoch     Loss              Accuracy
1         0.5479            0.7475
2         0.2196            0.9939
3         0.0879            0.9939
4         0.0517            0.9935
5         0.0392            0.9933
6         0.0343            0.9939
7         0.0315            0.9935
8         0.0311            0.9935
9         0.0298            0.9935
10        0.0299            0.9937

Final Metrics:
Loss: 0.0299
Accuracy: 0.9937
```

**HGBC-LSTM Pipeline**
Own presentment




Advisor
**Nikolaus Heners**

Co-Examiner
**Ludovico Bessi, Zürich, ZH**

Subject Area
**Security**