# 

# Student Project Network Analysis

|  |  |
| --- | --- |
| Project Title | Network Behavior Classification for Network Intrusion Detection |
| Industry Sponsorship (if Any) | n/a |

## 

## Project Description

**Problem definition**

*[50-100 word description of the problem which you will solve]*

|  |
| --- |
| With the expanded applications of modern-day networking (ie. internet of things, smart homes), network infrastructures are at risk from cyber-attacks and intrusions, which can compromise their availability, confidentiality, and integrity. And these threats are difficult to detect unaided, as they display indistinguishable network traffic patterns from normal traffic.  A Machine Learning based Network Intrusion Detection system can provide rapid identification of potential intrusions, letting network administrators take corrective action to cancel these threats. |

**Key Research Questions/ Technological constraints that the Project will Answer**

|  |
| --- |
| What is a threat to network flow? How to monitor it?  What are the key indicators for classifying a network flow as an intrusion?  The ability to use a multi-classification analysis provides a robust identification of the network flow giving the security team a better understanding of the type of port flow. |

**Final deliverables at the end of the project**

*[Please list the desired technical deliverables from the project team in as much detail as possible]*

|  |
| --- |
| Multi-classification analysis of network behavior with the ability to explore port statistics and delta port statistics to provide insight into the network’s behavior.  Our proposed dataset is based on a custom application that collects and logs the available statistics captured periodically (once every 5 seconds) from OpenFlow (OF) switches.  The statistics are collected by means of *OFPPortStatsRequest* and *OFPPortStatsReply* messages between controller and switches. The delta port statistics are computed on the controller side by taking the difference between the last two collected data instances.  The analysis of this dataset will provide a better understanding of the network architecture and its fine-grained network flows at the port level. This information can help the rapid identification of potential intrusions. |

**Key activities/ technologies the project team may be expected to undertake/ work with**

*[E.g. What kind of technology stack will you work with, the datasets you may need to work on, what kind of analysis you may be expected to undertake, etc.]*

|  |
| --- |
| Python, Pandas, Matplotlib; Keras, TensorFlow or PyTorch; Deep Learning for traffic classification;  Source for research:  <https://www.sciencedirect.com/science/article/pii/S0140366421000426>  <https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm> |

**Expected learning outcomes**

*[What do you expect to learn from the project? Please mention the technical skills you will imbibe over the project.]*

|  |
| --- |
| We expect to gain a deeper understanding of network traffic classification based on network traffic data and Deep Learning models.  AWS Streaming/batch data pipeline, |

|  |  |
| --- | --- |
| Team Size: | 2 |
| Member names: | Renato Barroco & Joaquin Gianantonio |

## System Design

From the System design perspective, outline the following:

***Data***

University of Nevada - Reno Intrusion Detection Dataset (UNR-IDD) that provides a wide range of samples and scenarios.

***https://www.tapadhirdas.com/unr-idd-dataset***

The main difference between UNR-IDD and existing datasets is that UNR-IDD consists primarily of network port statistics. These refer to the observed port metrics recorded in switch/router ports within a networking environment. The dataset also includes delta port statistics which indicates the change in magnitude of observed port statistics within a time interval. Compared to datasets that primarily use flow-level statistics, these port statistics can provide an analysis of network flows from the port level as decisions are made at the port level versus the flow level. The dataset also addresses the limitation of the presence of tail classes. According to UNR-IDD, the dataset ensures that there are enough samples for ML classifiers to achieve high F-Measure scores.

***Process (Models, iterations)***

Neural networks for Multi-Label Classification

<https://machinelearningmastery.com/multi-label-classification-with-deep-learning/>

***Outcome (output and recommendations)***

Multi-label classification for network analysis

What are the system design considerations for your deployable ML model? Describe the iterations, delivery formats, and limitations you may face and some solutions to overcome the limitations

The model should be deployed to be hit from an API or some sort of streaming process as events are generated.

What sort of infrastructure will be required for training? If a model requires a lot of resources, where is the best place to train?

We expect that the model will be trained using collab with GPU as a runtime;

## Ethical Considerations

Are there any ethical considerations for your project? Consider the data source, the intended outcome, and the eventual use cases.

Our dataset is publicly available to be used. It is offered by the University of Nevada Reno. It was developed as a test scenario for network analytics.

Did you modify anything about your plan based on these considerations?

We decided to use a publicly available dataset that serves the purpose of research in the field of network analytics.

Can you anticipate any issues that might arise during the process?

Whenever we use real-life datasets for network analytics privacy considerations can be an issue; in our project, we are using a dataset provided in an experimental scenario.