

**Learning Journey Project Proposal**

|  |  |
| --- | --- |
| Name of the Organization | Deep Learning for Network Traffic Monitoring |
| Project location (city or timezone) | Online/Michigan |

**A. Project Description**

**Problem definition**

[50-100 word description of the problem which the candidates need to solve]

|  |
| --- |
| The objective of this project is to build a machine learning model capable of analyzing networks, a model capable of categorizing network traffic, classifying the network to detect or predict cyber incidents.  As society becomes more interactively connected to the internet, better and more efficient capacity of our network traffic becomes mandatory. I hope to explore all the development done to date using Machine Learning models and contrast them with Deep Learning techniques.  Being able to detect network traffic is a key skill to secure new IoT devices. This can be done by identifying any anomaly through anomaly detection, malware detection, but also any kind of cyber deception, for example in IoTs used in Supply Chain.  **1. Data acquisition**  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 a fine-grained analysis of network flows from the port level as decisions are made at the port level versus the flow level. This can lead to rapid identification of potential intrusions. We also address the limitation of the presence of tail classes. Our dataset ensures that there are enough samples for ML classifiers to achieve high F-Measure scores, uniquely.  Introduction: <https://www.tapadhirdas.com/unr-idd-dataset>  **2. Model development**  Model usage source knowledge  <https://www.sciencedirect.com/science/article/pii/S0140366421000426>  Data acquisition:  The UNR-IDD Dataset is an official dataset for network-traffic analysis.  It consists of Flow Simulation (IPerf - TCP/UDP) data streams, the data collection uses a custom application to collect and log the available statistics. It not only maps the flow between ports (Flow Entry/Flow Table Statistics) but also uses statistics in time-interval definition (OpenFlow OF switches) - Port Statistics/Delta Port Statistics.  Five label are used in the dataset *TCP-SYN Flood, Port scan, Flow Table Overflow, Blackhole and Traffic Diversion*, binary & multi-calss classification. |
| Filter Model Development:  On the model development part of this project, the goal is to develop a model capable of classify which type of network traffic happend;  Two approaches could be in place:  1) Create a streaming solution where the data input is being consumed on real-time, whenever data drift happens retrain the model;  OR/AND  2) Timeseries analysis (each 5 second) classification may be a possible batch solution; |



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

|  |
| --- |
| Tech Constraints:   * Streaming vs. Batch load * Real time timeseries analysis OR small batch load classification through binning a timeseries (every five seconds)   Key Challenges – Dealing With:   * Map properly network traffic; * Compare distintic network routes; * Class imbalance (possibly, since data can change really fast)   Key Questions   * How to properly classify network traffic between different IoT devices? Can a model be generalized in this scenario? * Is the fictious scenario built by UNR a "scalable" scenario? * How can we understand and balance the bias-variance tradeoff for this model / use case? * How can we understand and balance the tradeoff between precision and recall for this model / use case and justify why one or the other should be emphasized for any model fine-tuning? |



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

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

|  |
| --- |
| - Data partition to simulate streaming input vs. output in S3 + Streaming in Amazon Kinesis, data streaming solution in AWS mapping a fictious network-traffic;  - A solution capable of perceive data drift to retrain the model;  - An alert (email) whenever malicious network traffic is detected; |

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

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

Technology stack:

* Python
* Pandas
* Keras or TensorFlow or PyTorch
* Amazon Kinesis Data Streams

Machine Learning:

* LSTM
* Recurrent neural networks
* DL for traffic classification
* <https://www.sciencedirect.com/science/article/pii/S0140366421000426>

**Expected learning outcomes**

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

|  |
| --- |
| 1. Network traffic classification using Deep Learning models;  2. AWS Streaming/batch data pipeline (maybe?);  3. Network traffic data understanding; |

|  |  |
| --- | --- |
| Desired Team Size (if any): | 1- 3 |