**GEORGE WASHINGTON UNIVERSITY**

**ADA UNIVERSITY**

**COMPUTER SCIENCE AND DATA ANALYTICS**

**GUIDED RESEARCH I**

**Report 2**

**Asiman Mammadzada**

# Title: Network Intrusion Detection System using Machine Learning

**Instructors: Prof. Dr. Stephen Kaisler, Assoc. Prof Dr. Jamaladdin Hasanov**

**Problem Description**

The problem aims to research possible application of machine learning algorithms in Network Intrusion Detection System (NIDS). NIDS is considerably significant to maintain the security and integrity of computer networks. Currently, the majority of NIDS is implemented on rule-based methods and signatures to identify network intrusions. Considering the evolving cyber-threats and complexity of network traffic, these rules created to detect intrusions become obsolete and less effective against sophisticated attacks. The goal of the project is to develop NIDS for specific sets of cyber-threats using Machine Learning algorithms for effective and accurate detection. The system is expected to be able to process real-time network traffic data based on learning from the historically labeled data. The ML algorithms should be trained on diverse dataset to make sure the robust identification of known network intrusions.

**Strategy Definition:**

In this project, both qualitative and quantitative approach will be utilized during the research.

The variety of ML multi-classification models will be targeted to be trained on the dataset. Having applied the hyper-parameter tuning to attain the highest accuracy, the model will be final tested on testing dataset. Qualitative approach will help to understand the contexts, nuances of network data which in fact cannot be adequately captured by quantitative approach solely. Examination of network packet payloads and their semantic content will be one aspect of qualitative research in ML application in NIDS.

In addition, quantitative research plays an important role in advancing the field of ML for NIDS. Based on the using aggregated numerical dataset, statistical analysis and modelling relying on mathematical formulas, the quantitative approach will provide an objective to deeply understand NIDS and its combination with ML. Quantitative methods in ML-NIDS research sometimes entail gathering and analyzing vast amounts of network traffic information. Packet headers, flow logs, connection logs, and other network-related data are examples of this data. Utilizing quantitative methods, researchers can find statistical trends, patterns, and anomalies in the data, providing insightful information on network intrusions.

**Dataset Selection:**

This implementation will be carried out using KDD Cup dataset. The KDD Cup dataset is open-source dataset which actually has 42 features on variety of intrusions. It has around 494K records of intrusions.

**Dataset Preparation:**

The datasets are labeled and structured dataset. But, considering the extensive number of features, some feature engineering techniques will be applied to get the most predictive factors among them. Label Encoder technique will be applied in feature engineering and using the RandomForest the best predictors will be selected and trained in different models. One hot encoding will not be utilized in this case as when it is applied, the number of the columns is getting extensively big. Thus, label encoding is thought to be appropriate for the specific dataset.

**Flow Diagram:**