NSL-KDD Intrusion Detection

Applying Machine Learning to Cybersecurity



November 21, 2025

By Naoko Kamo and Dr. Andrew Pownuk

Agenda

- Challenge Overview
- What is NSL-KDD Dataset?
- Machine Learning Approach
- Submission Requirements
- Learning Outcomes

NSL-KDD Challenge Introduction

Challenge Overview

The Data Analytics Challenge focuses on NSL-KDD Intrusion Detection, where students build machine learning models to identify malicious network access and enhance cybersecurity defenses.

Relevance to Cybersecurity

This challenge provides valuable IDS* exposure to detecting cyber threats, improving network security, and understanding real-world intrusion scenarios using open-source datasets.

**IDS (Intrusion Detection System): A network security tool that monitors network traffic and devices for malicious activity, suspicious behavior, or policy violations.*

Learning Goals

Participants will practice machine learning techniques to detect harmful network access, gaining skills in data analysis, model training, and threat identification relevant to cybersecurity careers.



Dataset

Understanding NSL-KDD Dataset

Open-Source Benchmark

NSL-KDD is a publicly available dataset designed for evaluating intrusion detection systems, providing researchers with realistic network traffic data.

Training and Testing Sets

The dataset is divided into training (125,973 records) and testing (22,544 records) subsets to enable effective model building and evaluation.

41 Distinct Features

It includes 41 attributes capturing connection details such as protocol type, service, source and destination IPs, and traffic characteristics.

Diverse Attack Categories

Attacks are categorized mainly as Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), and Probing, supporting comprehensive intrusion analysis.

Introduction

Timeline of the NSL-KDD Intrusion Detection Challenge

1998

DARPA Dataset Creation

The original dataset was created by DARPA at MIT Lincoln Labs to simulate network traffic with intrusions for research.

1999

KDD Cup 99 Dataset

Derived from the DARPA dataset, the KDD Cup 99 became a widely used benchmark for intrusion detection system development.

2009

NSL-KDD Dataset Release

An improved version of KDD Cup 99 was released to address issues of redundancy and imbalance in the data.

2017

Rise of Zero-Day Attacks

Cybersecurity threats evolved with more sophisticated and unknown attacks, increasing the need for advanced detection methods.

2019

Advanced ML Techniques

Machine learning and deep learning methods gained popularity for enhancing intrusion detection accuracy and efficiency.

2023

Challenge Announcement

The NSL-KDD Intrusion Detection Challenge was introduced to engage students in practical cybersecurity and machine learning applications.

2024

Use of Google Colab

Sample base code provided on Google Colab enables participants to experiment with model training and improvements.

2025

Future Learning and Innovation

Participants will gain valuable skills in machine learning, cybersecurity evaluation, and visualization through this challenge.

Techniques

Machine Learning in Intrusion Detection

01

Supervised Learning Techniques

Supervised methods use labeled data to train classifiers like decision trees, support vector machines (SVM), and neural networks to distinguish between normal and malicious network activities accurately.

02

Anomaly Detection Advantage

Anomaly-based IDS models normal behavior and flags deviations as potential intrusions, enabling detection of unknown or zero-day attacks that signature-based methods may miss.

03

Model Training and Improvement

Participants can leverage provided sample base code on Google Colab to train models, modify algorithms, and experiment with feature selection to improve detection accuracy and reduce false alarms.

04

Practical Machine Learning Applications

Machine learning empowers IDS to adapt to evolving cyber threats by learning complex patterns from network data, thus enhancing the system’s robustness and real-time detection capabilities.

Threat Landscape

Challenges and Opportunities in IDS

01

Zero-Day Attacks

These attacks exploit unknown vulnerabilities, making them invisible to signature-based IDS. Detecting zero-day threats requires adaptive and learning-based anomaly detection methods.

02

Evasion Techniques

Attackers use fragmentation, flooding, obfuscation, and encryption to mask malicious traffic, challenging IDS effectiveness and requiring robust detection strategies.

03

Fragmentation & Flooding

Fragmentation splits malicious packets to evade detection, while flooding overwhelms IDS with high traffic volumes, creating blind spots for attackers to hide.

04

Anomaly-Based IDS Importance

Unlike signature-based systems, anomaly-based IDS profiles normal behavior and flags deviations, enabling detection of novel and sophisticated attacks.

Sample base code in Google Colab

🚀 NSL-KDD Intrusion Detection — Google Colab Notebook

This notebook demonstrates a **simple cybersecurity ML pipeline** on the **NSL-KDD** public dataset (a cleaned successor to KDD'99).
We build a classifier that predicts whether a network connection **is malicious or normal**.

What you'll see:

- Download + load **NSL-KDD** (`KDDTrain+.txt`, `KDDTest+.txt`)
- Preprocess categorical & numeric features (one-hot + scaling)
- Train a model (Logistic Regression; Random Forest optional)
- Evaluate: accuracy, precision, recall, F1, ROC-AUC
- Plot: confusion matrix and ROC curve

> Dataset source (public mirror): `https://github.com/defcom17/NSL_KDD`
> Task: **Binary classification** (attack vs normal). You can easily extend to **multiclass** by using the specific attack labels.

Setup

```
1
2 #@title Setup
3 import os, io, sys, math, zipfile, urllib.request
4 import numpy as np
5 import pandas as pd
6 import matplotlib.pyplot as plt
7
```

<u>15 sections implement the following flow:</u>

1. Download + load NSL-KDD
(KDDTrain+.txt, KDDTest+.txt)
2. Preprocess categorical & numeric features
(one-hot + scaling)
3. Train a model
(Logistic Regression; Random Forest optional)
4. Evaluate:
accuracy, precision, recall, F1, ROC-AUC
5. Plot:
confusion matrix and ROC curve

Where to start in Google Colab?

Build preprocessing + model pipeline

```
#@title Build preprocessing + model pipeline

preprocess = ColumnTransformer(
    transformers=[
        ("cat", OneHotEncoder(handle_unknown="ignore"), cat_cols),
        ("num", StandardScaler(with_mean=False), num_cols),
    ]
)

# Choose model here:
use_random_forest = False  #@param {type:"boolean"}

if use_random_forest:
    model = RandomForestClassifier(n_estimators=200, random_state=42, n_jobs=-1)
else:
    model = LogisticRegression(max_iter=200, n_jobs=None, solver="lbfgs")

pipe = Pipeline([("prep", preprocess), ("clf", model)])

pipe

```

Two models are set as default:

-Random Forest

-Logistic Regression

Try to switch between two by editing the code of #Choose model here



Requirements

Submission Requirements

- Submit an improved and optimized version of the base code provided.
- Include a detailed explanation of what changes were made and the rationale behind them.
- Present a visual dashboard or demonstration showcasing how the detection accuracy or rate has improved.
- Demonstrate data insights and model evaluation metrics clearly.
- Ensure submissions highlight the practical impact of the improvements on intrusion detection.

Summary

Conclusion and Learning Outcomes

- Gain hands-on experience applying machine learning to cybersecurity challenges.
- Work with a real-world intrusion detection dataset, NSL-KDD, for practical analysis.
- Develop skills in building and improving machine learning models for detecting malicious network activity.
- Understand different types of intrusion detection systems and their strengths and limitations.
- Create visual dashboards or demonstrations to effectively communicate detection improvements.

Resources

Links

- **Challenge Website:** https://www.online-computing.org/datathon-2025-fall/
- **Google Colab Notebook:** https://colab.research.google.com/github/apownukepcc/fall-2025-datathon/blob/main/NSL_KDD_Intrusion_Detection_Colab_v2.ipynb
- **NSL-KDD Dataset Source:** Open-source dataset widely used for intrusion detection research and benchmarking.https://github.com/defcom17/NSL_KDD
- **Machine Learning Demonstrations:** Practical examples showcasing core ML concepts and applications.https://www.online-computing.org/datathon-2024-fall/examples.html
- **Introductory Machine Learning Course:** MATLAB Machine Learning Onramp for foundational knowledge in ML techniques. https://matlabacademy.mathworks.com/en/details/machine-learning-onramp/machinelearning
- **Hackathon:** https://www.hacktheborder.org/