WatchDogX: AI Cyber Threat Detection and Response System

WatchDogX is focused on utilizing artificial intelligence to enhance threat detection and response to cyber threats.

Submitted by - Mridul Goyal(210C2030141)

Project Overview

Goal

Detect and respond to cyber threats with advanced AI techniques.

Scope

Backend model training and detection pipeline development.

Problem Statement

Address the growing sophistication of network-based attacks.

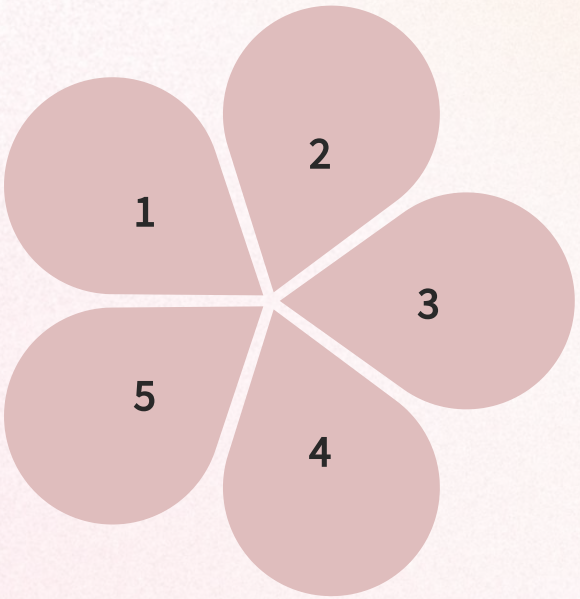
Literature Review

Outdated Signature-based IDS

Traditional signature-based Intrusion Detection Systems (IDS) are no longer effective against modern cyber threats, necessitating new approaches.

Identified Gap

There is a need for a real-time, interpretable, and alert-capable system that can quickly inform users of potential threats.



Adaptability of ML/AI

Machine learning and AI techniques provide the adaptability needed to respond to evolving cyber threats in real-time.

CICIDS2017 Dataset

The CICIDS2017 dataset is widely utilized in IDS research and serves as a benchmark for evaluating detection methods (Sharafaldin et al., 2018).

Importance of Feature Selection

Effective feature selection can significantly improve the speed and accuracy of IDS, enhancing overall system performance.

Overview of CICIDS2017 Dataset

Dataset Name

The CICIDS2017 dataset is crucial for training and evaluation.

Types of Attacks

It features 14 types of cyber attacks along with normal traffic data.

Data Size

Approximately 12GB when compressed, offering extensive data for training.

Source of Data

This dataset is provided by the UNB Canadian Institute for Cybersecurity.

Data Preprocessing Steps

1

Data Merging

Merging and filtering multiple CSV files.

2

Irrelevant Columns Removal

3

Handling Missing Values

Employing strategies to manage missing values.

4

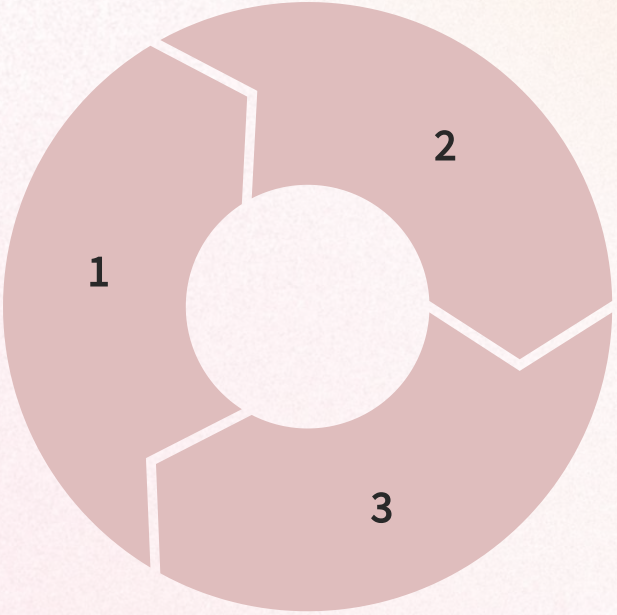
Attack Pair Creation

Generating Attack_vs_BENIGN.csv for comparisons.

Undersampling and Feature Selection

Class Imbalance Fix

The RandomUnderSampler technique was applied to address the class imbalance in the dataset, enhancing model training effectiveness.



Feature Selection Process

Top 3-4 features were selected for each attack type based on correlation metrics and information gain, optimizing the model's predictive capabilities.

Example Features

Key selected features include Flow Duration, Packet Length Standard Deviation, and Bwd Packet Length Mean, which are crucial for distinguishing between attack types.

Machine Learning Models and Their Performance

This presentation focuses on the training and evaluation of various ML models, highlighting the superior performance of the MLP model in cyber threat detection while ensuring reliability throughout the process.

Trained Models

Three machine learning models were trained, including MLP, Naive Bayes, and QDA.

Best Accuracy

MLP demonstrated the best accuracy in detecting cyber threats.

Reliability Evaluations

The models underwent repeated evaluations to ensure reliability.

Exporting MLP

The trained MLP model was exported using joblib for deployment.

Performance Metrics Analysis

1 **MLP Accuracy**

The MLP model achieved an impressive accuracy range of 95–99%, indicating its strong performance in threat detection.

2 **Comparative Speed**

Naive Bayes and QDA provided lower accuracy but offered faster processing times, beneficial for real-time applications.

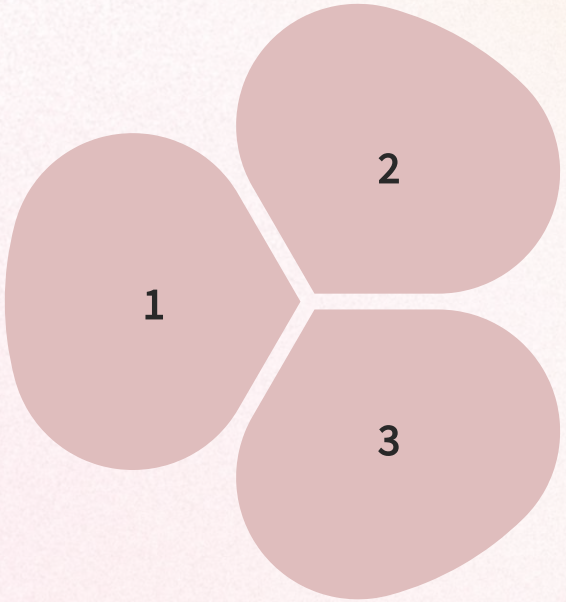
3 **Early Detection**

Models showed promising results in early detection capabilities for DDoS and DoS attacks, critical for mitigating damage.

Architecture Overview (So Far)

Backend Pipeline

The current architecture consists of a structured backend pipeline involving dataset ingestion, preprocessing, feature selection, and ML model training.



Outputs

The pipeline produces a trained .pkl model along with a scaler, providing threat predictions along with confidence scores.

Current Interface

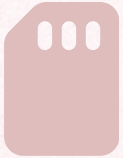
At this stage, the system accepts CSV input and provides outputs via the terminal, highlighting areas for future user interface development.

Key Features and User Flow

This document outlines the key features of the system, detailing user flow, supported data, and future enhancements for a practical cybersecurity tool.

User Interaction Flow

Allows users to upload CSV files for prediction, logging results back into CSV.



Supported Data

Accepts filtered logs, detecting known attack types for cybersecurity.

Future Additions

Enhancements planned include a real-time sniffer, alerts, geo-location logging, and a web dashboard.

Thank You