Creating an Explainable Intrusion Detection System Using Self Organizing Maps

Jesse Ables
jha92@msstate.edu
Mississippi State University
Department of Computer Science

Introduction

- Rise in Cyber Attacks
- Attacks are changing everyday
- We have accurate IDS
- There can be a more effective IDS

Intrusions and Intrusion Detection Systems

- Intrusion - An action that obtains unauthorized access to a network or system
- Violates the CIA principles:
  - C - Confidentiality
  - I - Integrity
  - A - Availability
- IDS consist of tools, methods, and resources used to protect networks or systems
- Detect **anomalous** network packets
  - Signatures
  - AI Anomaly Detection

Current IDS

- Black Box
  - Most State-of-the-Art approaches
  - Opaque decision process
  - Focus on accuracy over explainability
  - ANN, Deep Learning, SVM
- White Box
  - Less Complex
  - Decision Process can be analyzed
  - Decision Trees, SOM, Regression-based Approaches

Explainable AI and Explainable-IDS

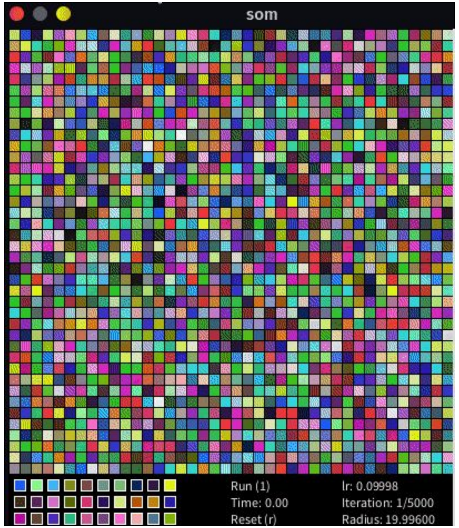
- Darpa’s Definition:
  - “An AI should explain the reasoning for its decisions, characterize its strengths and weaknesses, and convey a sense of its future behavior.”
- X-IDS should answer questions such as:
  - Why is a network flow anomalous?
  - How did the model come to its conclusion?
  - What patterns did it see to create its prediction?
- Goal of XAI and X-IDS
  - Build **Trust**
  - Fairness, Reliability, Privacy, and Causality
  - Aid user performed tasks

Why Explainability in IDS

- Explainability can help professionals perform tasks
  - Doctors making medical decisions []
  - Accountants making credit score decisions []
  - **Security Analysts** making **network security** decisions []
- Benefits to the stakeholders of an IDS
  - Security Analysts
    - Quicker and more accurate action on security decisions
  - IDS Developers
    - Fortify the model by finding flaws in its logic
  - Investors
    - Performative metrics to aid in overall business desicions

Explainable Self-Organizing Maps

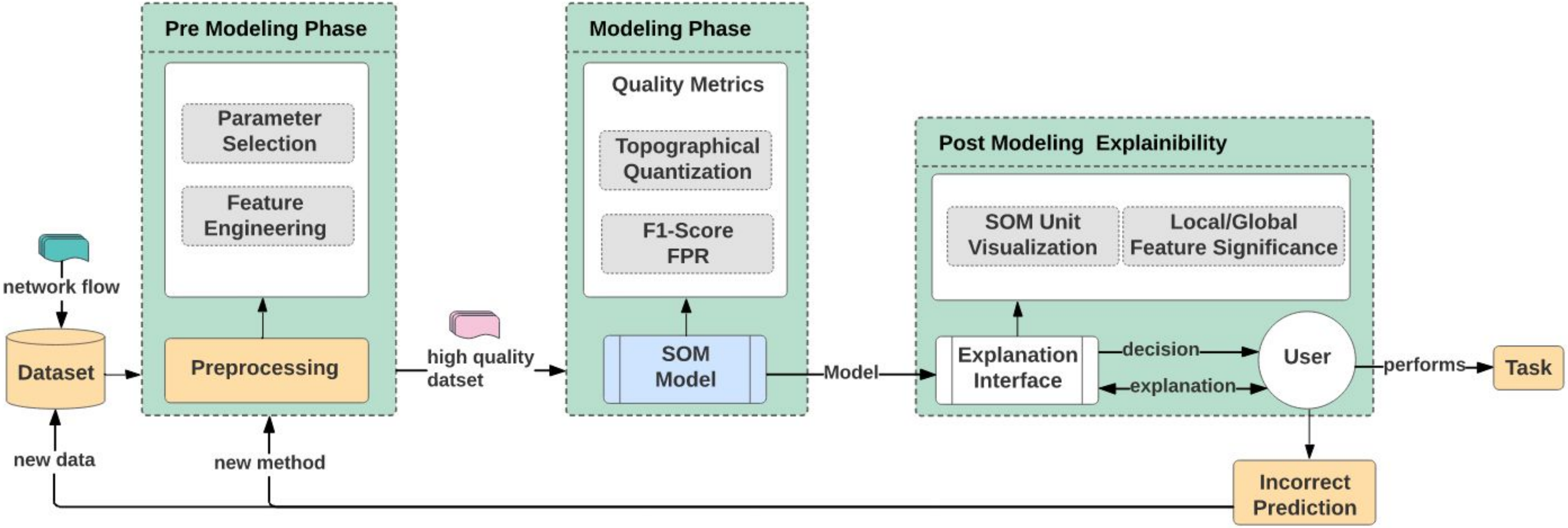
- SOM
  - Grid of nodes on a 2D plane
  - Weights are slowly adjusted towards the training data
  - Forms clusters similar to the categories in a dataset
  - Toy Example: RGB colors using Processing 4
- Training Process
  - Pick a training sample
  - Find Best Matching Unit (BMU) using Euclidean Distance
  - Adjust weights of BMU and its neighbors
  - Closer neighbors are adjusted more
  - Repeat!
- Testing/Predicting
  - Use training data to set a label on each node
  - Pick a testing sample
  - Find BMU
  - Predict!



Explainable Self Organizing Maps

- Visual
  - U-Matrix, Starburst U-Matrix [] (maybe add U-matrix from processing SOM)
  - Feature Heat Map
  - Label Cluster Map
- Statistical
  - Local Feature Significance
  - Global Feature Significance

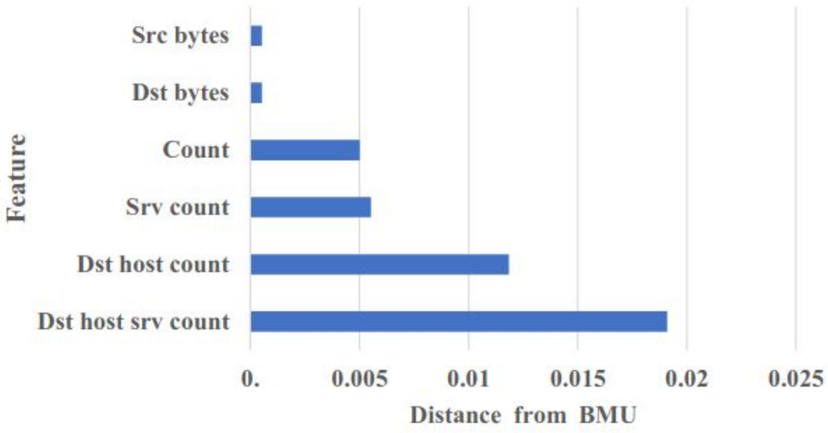
Proposed SOM X-IDS Architecture



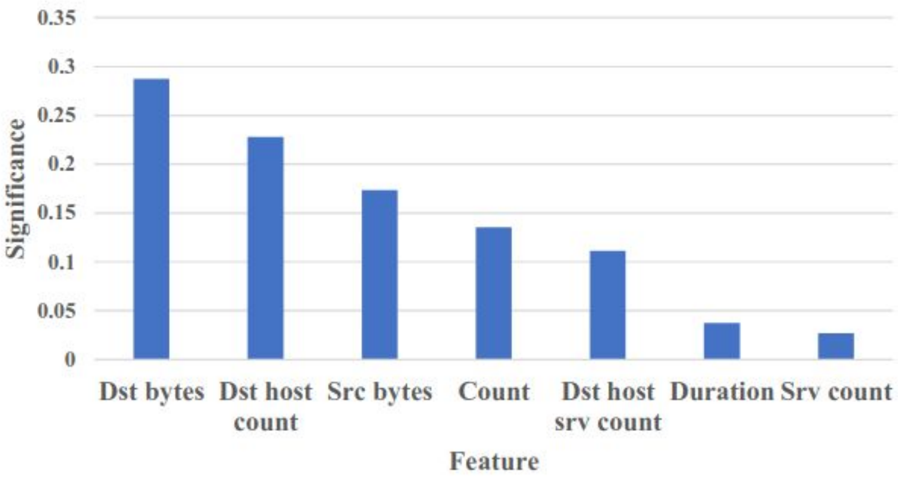
Datasets and Methodology

- NSL-KDD []
  - Year
  - Samples
  - Ratio
- CICIDS-2017 []
  - Year
  - Samples
  - Ratio
- SOM Testing Parameters
  - 18 x 18 units
  - 1000 Training Iterations
  - 70/30 Training-Testing Split

Explainability Results: NSL-KDD

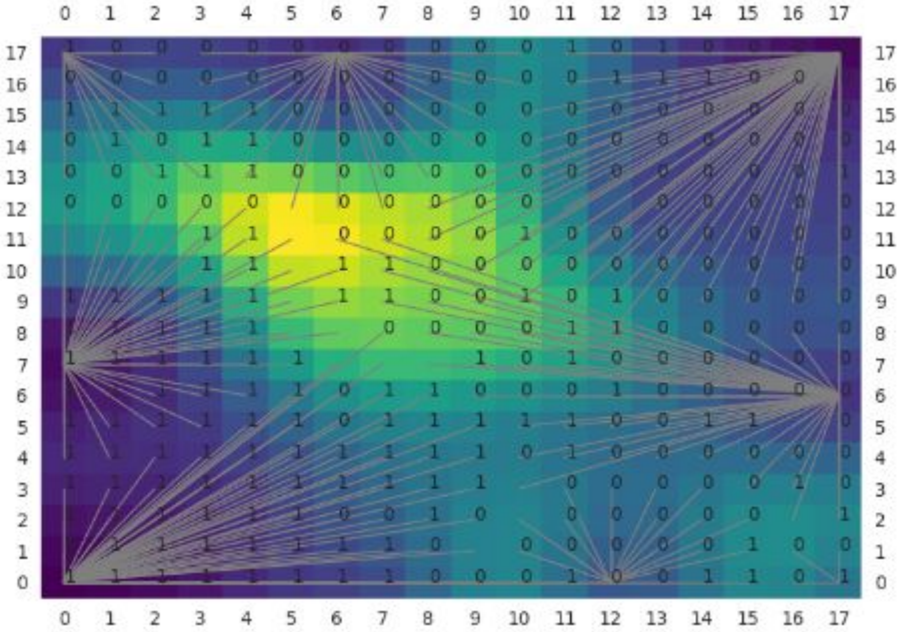


Local Prediction Explanation

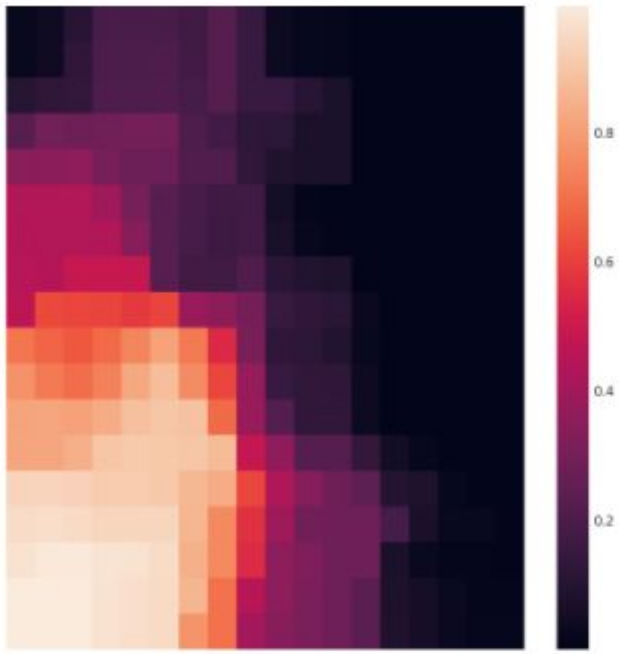


Global Feature Explanation

Explainability Results: NSL-KDD

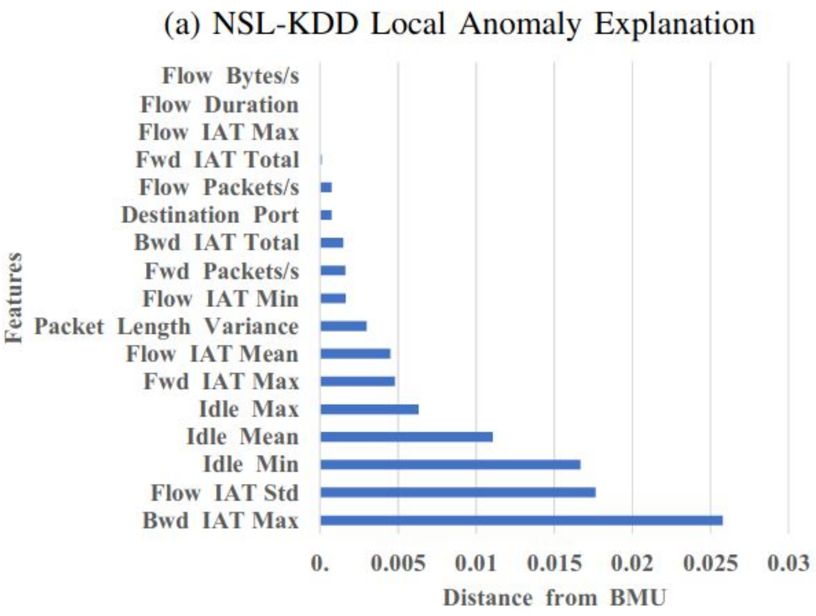


NSL-KDD Starburst U-Matrix

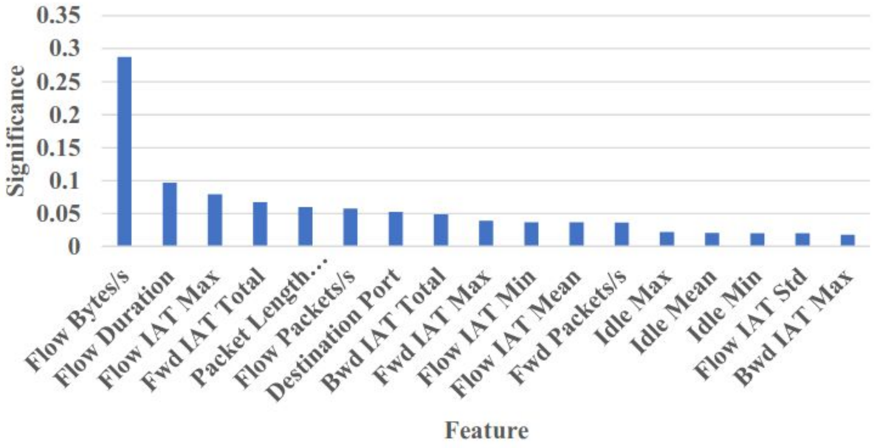


“DST Bytes” Feature Heatmap

Explainability Results: CICIDS-2017

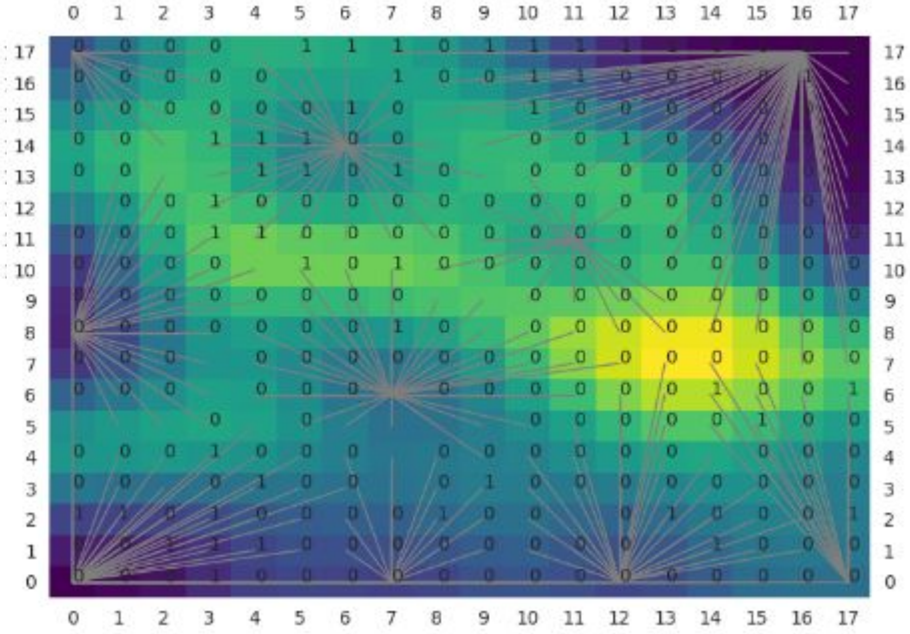


Local Prediction Explanation

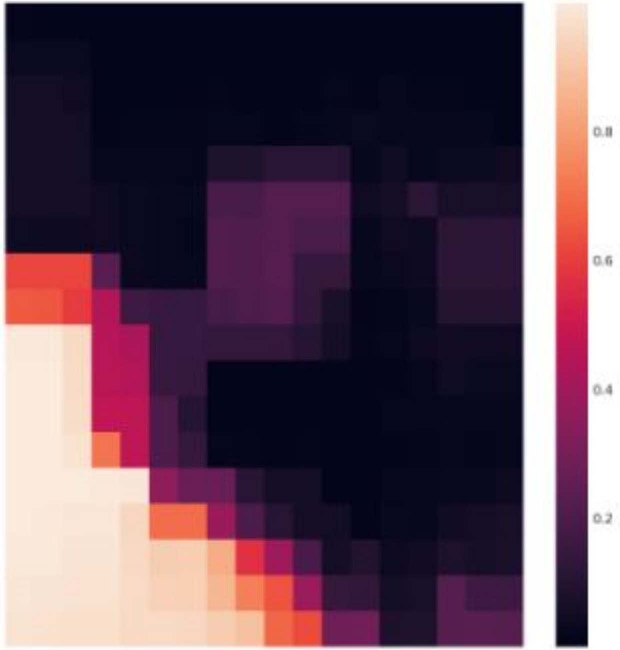


Global Feature Explanation

Explainability Results: CICIDS-2017



CICIDS-2017 Starburst U-Matrix



“Flow Bytes per Second” Feature Heatmap

Results

Trained SOM Results

Dataset	F1	Precision	Recall	FPR	FNR
NSL-KDD	91.0%	91.0%	91.4%	9.4%	8.0%
CIC-IDS-2017	80.0%	77.4%	81.8%	22.5%	4.5%

SOM Compared to Black-box algorithms

Dataset	SOM	Random Forest	DBN
NSL-KDD	91.0%	99.67%	97.5%
CIC-IDS-2017	80.0%	97.1%	94.0%