Enhancing Explainability and Trustworthiness of Intrusion Detection Systems Using Competitive Learning

Dr. Jesse Ables Ph.D.
ables@southalabama.edu
Assistant Professor
University of South Alabama
Department of Computer Science

Introduction

- Black Box AI models dominate the space
- Used everywhere: accounting, medicine, **network security**
- High Accuracies, Low Explainability
- Whitebox algorithms can be just as accurate and are far more explainable

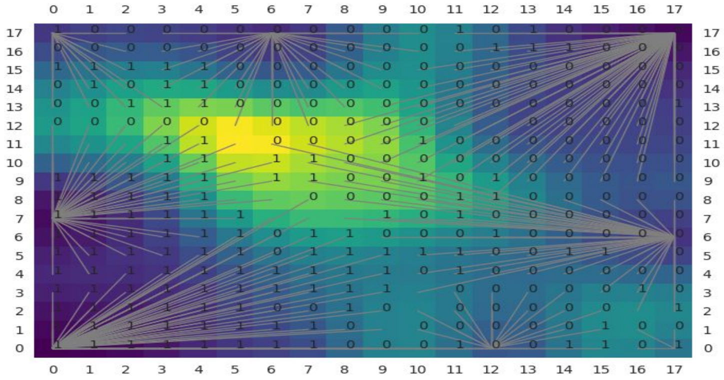
OpenAI

Google DeepMind

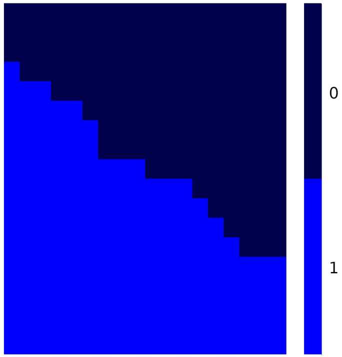
Introduction

- Competitive Learning (CL)
  - CL can be accurate
  - Easy to understand
  - Can produce visual and statistical explanations

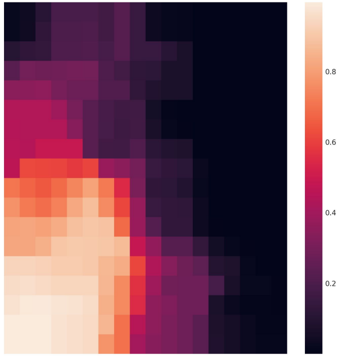
Self Organizing Map U-Matrix



Simplified Label Map



Feature Heatmap



Ables et al. “Creating an Explainable Intrusion Detection System Using Self Organizing Maps”, 2022

Overview

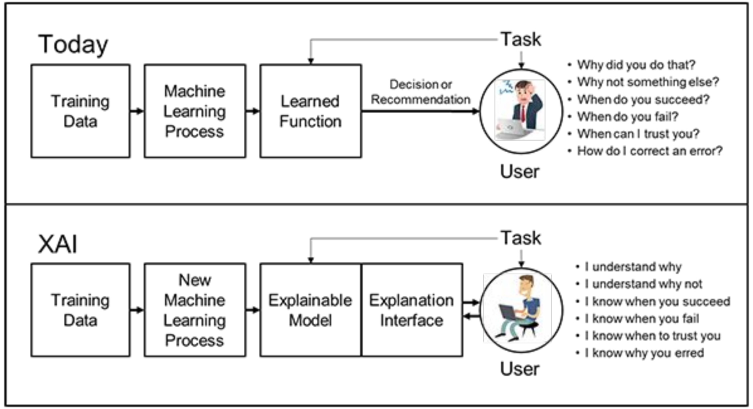
- **Explainability and Intrusion Detection**
- Competitive Learning Algorithms
  - Self Organizing Maps
  - Growing Self Organizing Maps
  - Growing Hierarchical Self Organizing Maps
- Our Proposed Architecture for X-IDS

Explainable AI

- No consensus on the definition of **Explainability**
- DARPA's Definition[1]:
    - "An AI should explain the **reasoning** for its decisions, characterize its **strengths** and **weaknesses**, and convey a sense of its **future behavior.**"
    - Solid foundation for consensus
- Goal of XAI
    - Build **Trust**
    - **Aid** user performed **Tasks**
    - Fairness, Reliability, Privacy, and Causality

Explainability and Intrusion Detection

- X-IDS **Reasoning**
  - Why is a network flow anomalous?
  - How did the model come to its conclusion?
  - What patterns did it see to create its prediction?
- X-IDS **Strengths and Weaknesses**
  - Metrics
  - Explanations on Incorrect Predictions
- X-IDS **Future Behavior**
  - Local and Global Explanations



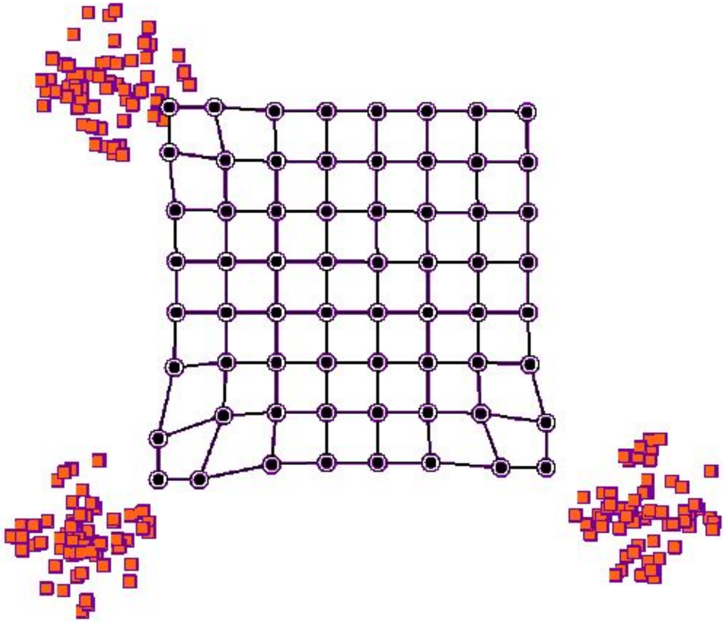
DARPA's AI vs XAI diagram [2]

Overview

- Explainability and Intrusion Detection
- **Competitive Learning Algorithms**
  - Self Organizing Maps
  - **Growing Self Organizing Maps**
  - **Growing Hierarchical Self Organizing Maps**
- Our Proposed Architecture for X-IDS

Explainable Self-Organizing Maps

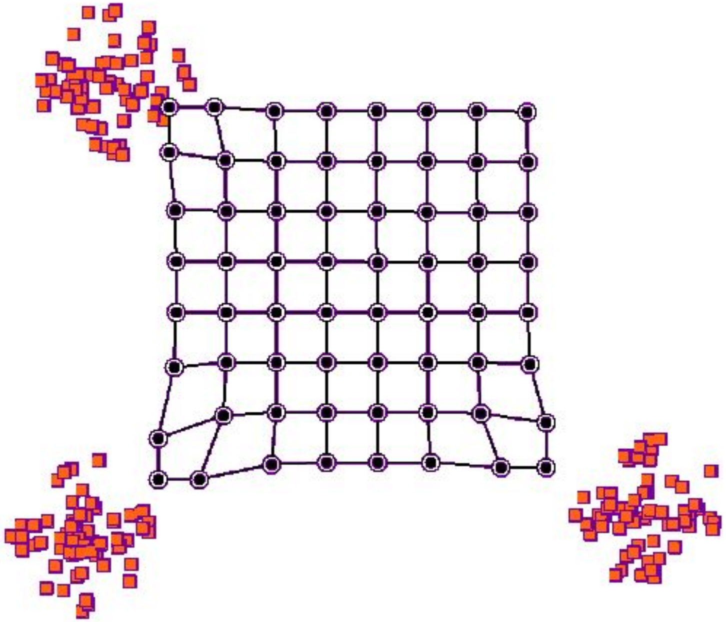
- SOM
  - Grid of nodes on a 2D plane
  - Weights are slowly adjusted towards the training data
  - Forms clusters similar to the categories in a dataset
  - Example: Red dots are training samples, black dots and lines are the model nodes
- Training Process
  - Pick a training sample
  - Find Best Matching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



SOM adjusting towards training data [3]

Explainable GSOM and GHSOM

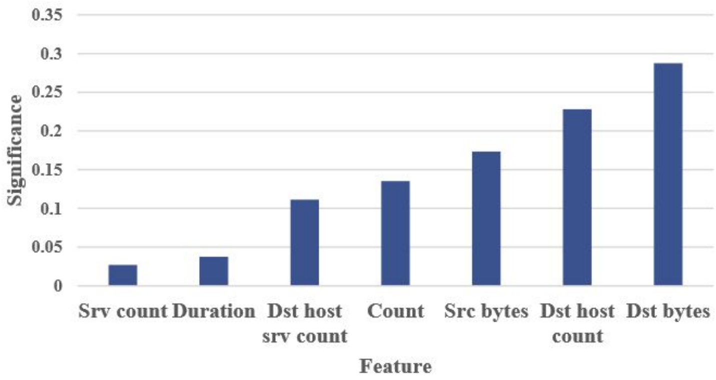
- Self Organizing Map
  - Starts with a **m** x **n** grid of nodes
  - Pick training sample
  - Find Best Matching Unit (**BMU**) using Euclidean distance
  - Adjust weights
- Growing Self Organizing Map
  - Start with **2** x **2** grid of nodes
  - **Growth Threshold** and
  - **Cumulative Error**
  - Add new **node**
- Growing Hierarchical Self Organizing Map
  - **Vertical Threshold**
  - Add new child **GSOM**



SOM adjusting towards training data [3]

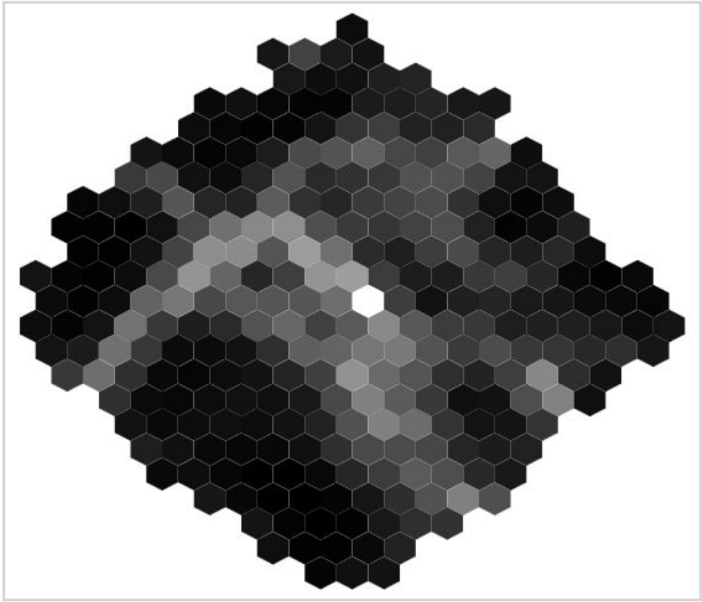
Types of Explanations

- Statistical
    - Local Feature Significance
    - Global Feature Significance
- Visual
    - Unified Distance Matrix (U-Matrix)
    - Feature Heat Map



Feature Significance Chart

U-Matrix

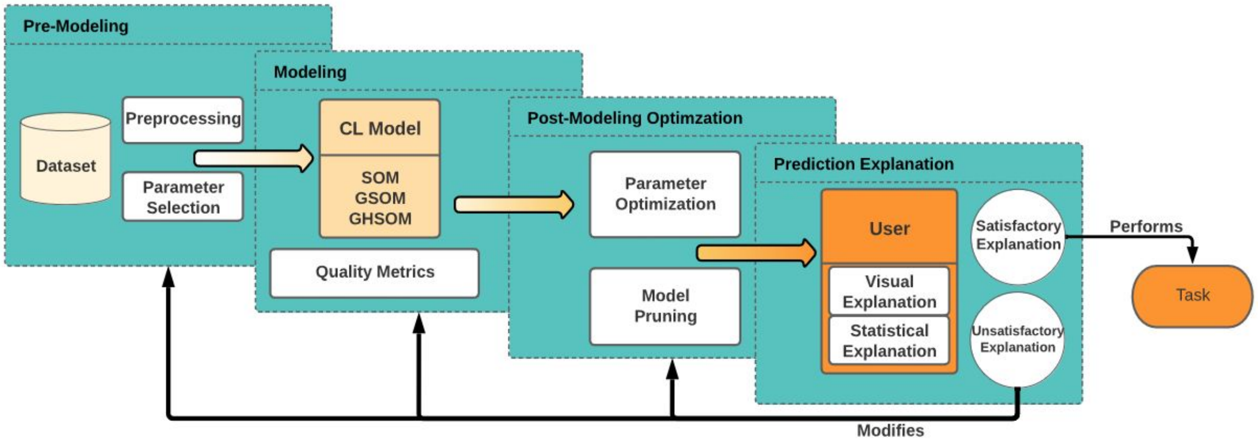


Overview

- Explainability and Intrusion Detection
- Competitive Learning Algorithms
  - Self Organizing Maps
  - Growing Self Organizing Maps
  - Growing Hierarchical Self Organizing Maps
- **Our Proposed Architecture for X-IDS**

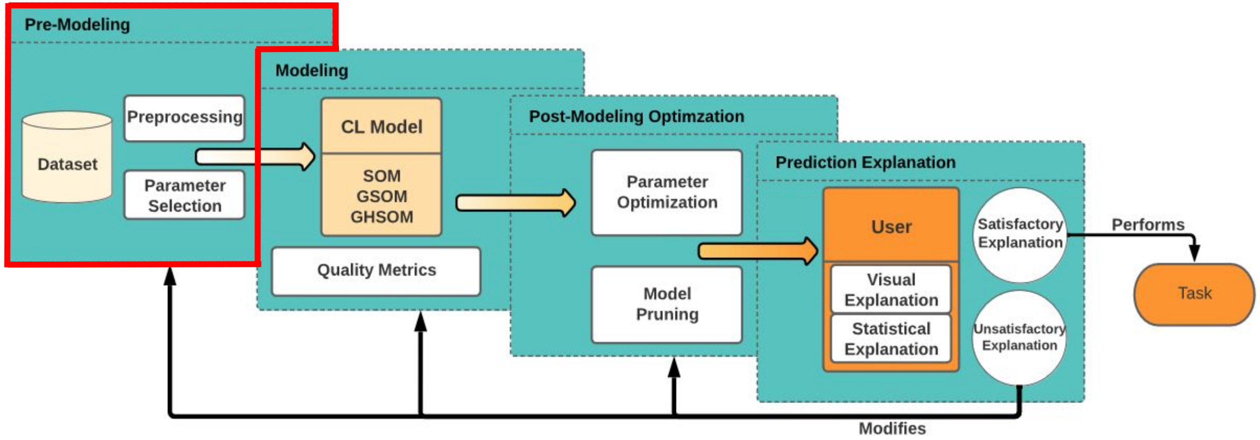
X-IDS Architecture

- Competitive Learning X-IDS
- Based on Darpa's Recommendations
- Model Optimization and Pruning
- User "in-the-loop" feedback



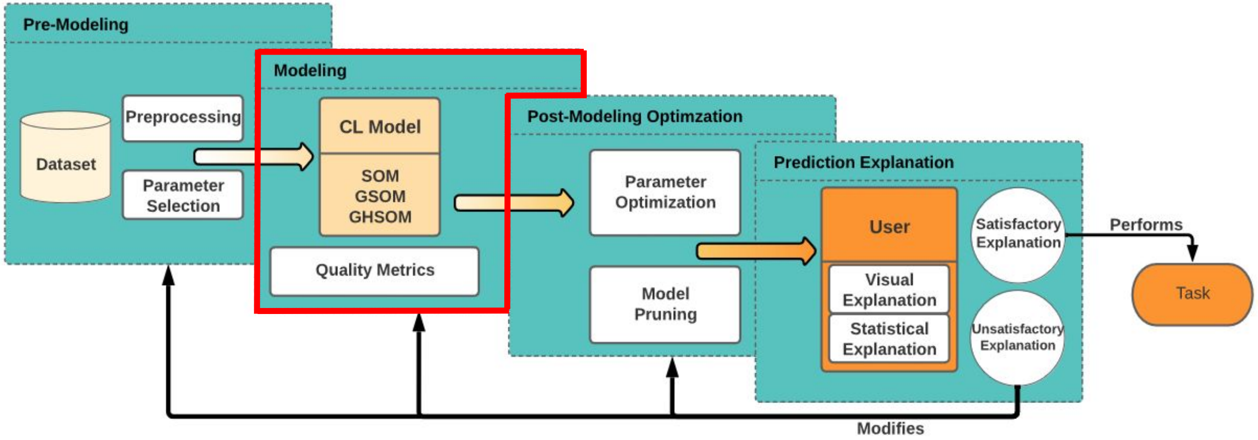
X-IDS Architecture

- Pre-Modeling Phase
  - Dataset Preprocessing
  - Model Parameter Selection



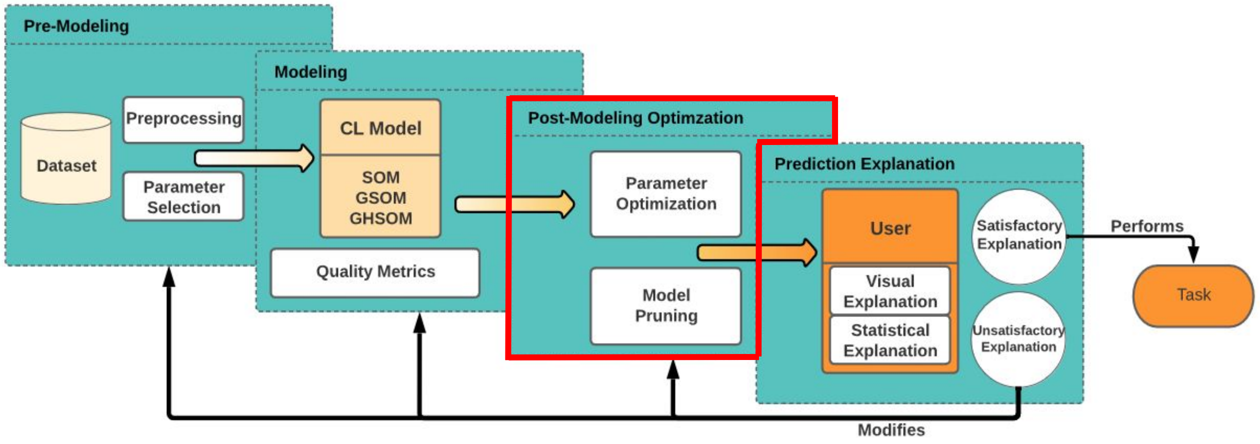
X-IDS Architecture

- Modeling Phase
  - Train Competitive Learning algorithms
  - Record quality metrics



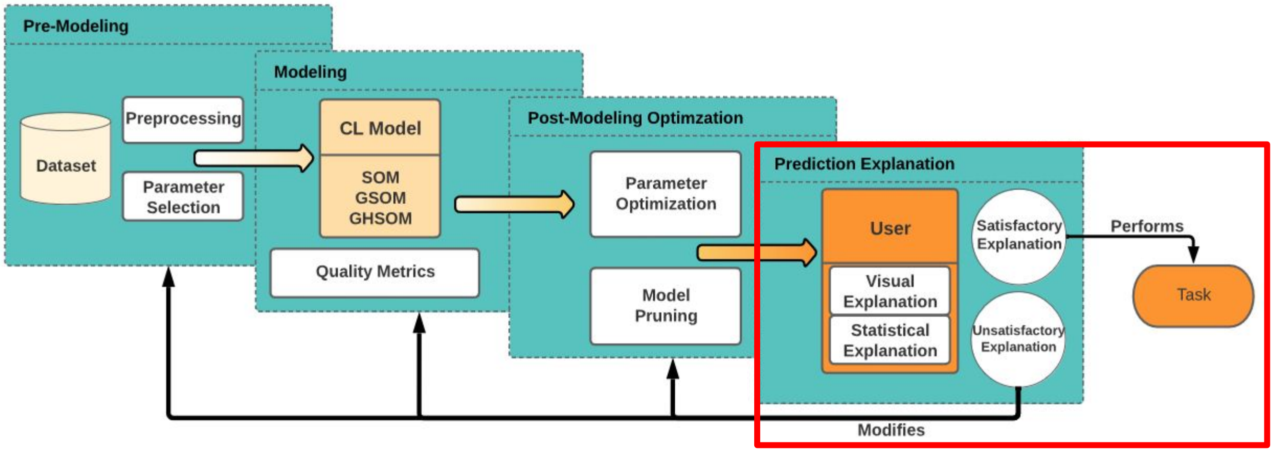
X-IDS Architecture

- Post-Modeling Optimization
  - Increase model performance
  - Parameter optimization (Bayesian Search)
  - GHSOM map pruning



X-IDS Architecture

- Prediction Explanation
  - Visual and Statistical explanations
  - Modified for GSOM and GHSOM
  - User feedback to improve X-IDS



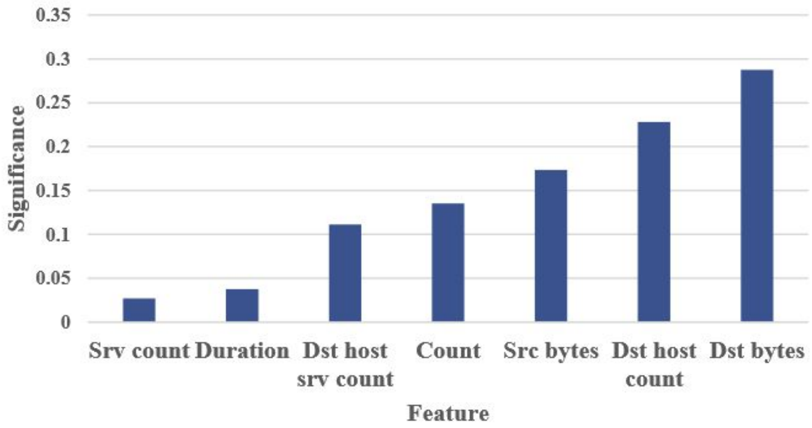
Datasets and Methodology

- NSL-KDD [4]
  - Original 1999, Updated 2009
  - 150,000 samples
  - 46.5% Contamination Rate
- CICIDS-2017 [5]
  - 2017
  - 2.8 M samples
  - 19.7% Contamination Rate
- Experiments
  - Explainability
  - Accuracy

	Parameter	NSL-KDD	CIC-IDS-2017
SOM	n	18	18
	m	18	18
	LR	.3	.3
	Epochs	1000	1000
GSOM	LR	.006	.006
	SF	.9	.9
	Epochs	100	40
GHSOM	LR	.006	.006
	SF	.3	.3
	Epochs	100	40

TABLE I: The selected parameters for each CL model.

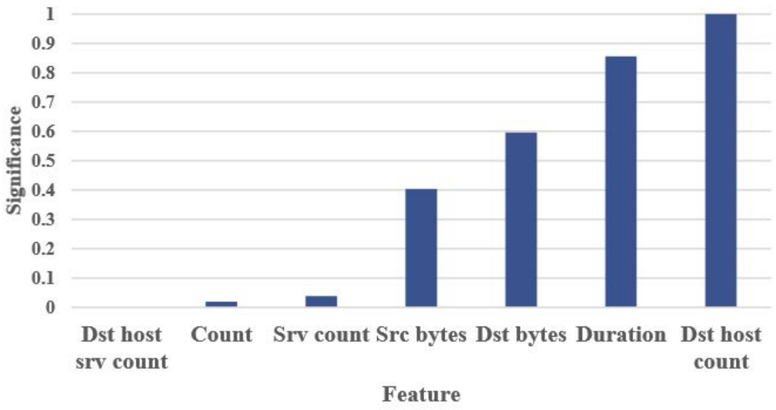
Explainability Results: NSL-KDD



Global feature significance of the trained GSOM model

- **Global Feature Significance**
  - Features - Dataset Features
  - Significance - Features with higher variability that change labels
  - If we change this feature, we are more likely to change the prediction outcome
- Greater = More Impactful!
- Most important feature is **Destination Bytes**

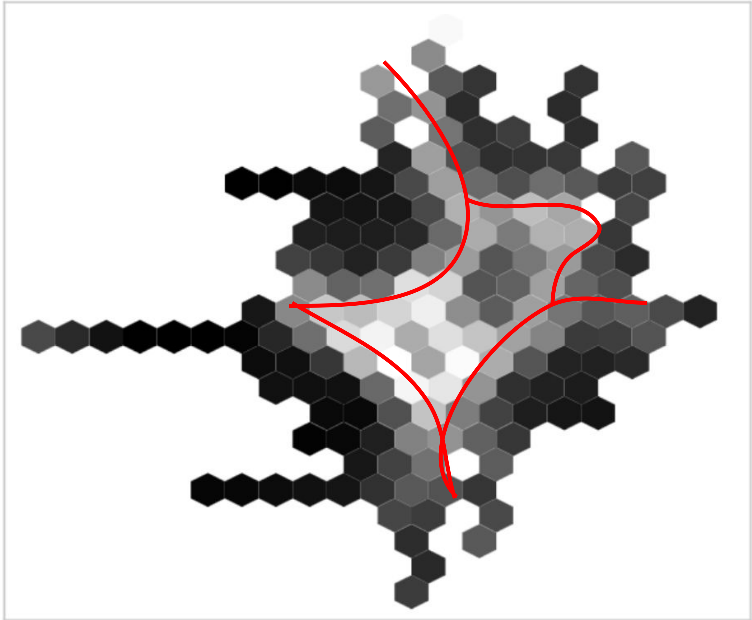
Explainability Results: NSL-KDD



Local feature significance for an Anomalous sample

- **Local Feature Significance**
  - Features - Dataset Features
  - Similarity to BMU - Euclidean distance from the **Best Matching Unit**
- Higher = More Impactful!
- **Anomalous** Sample
  - Most significant features are **Destination Bytes**, Duration, Destination Host Count

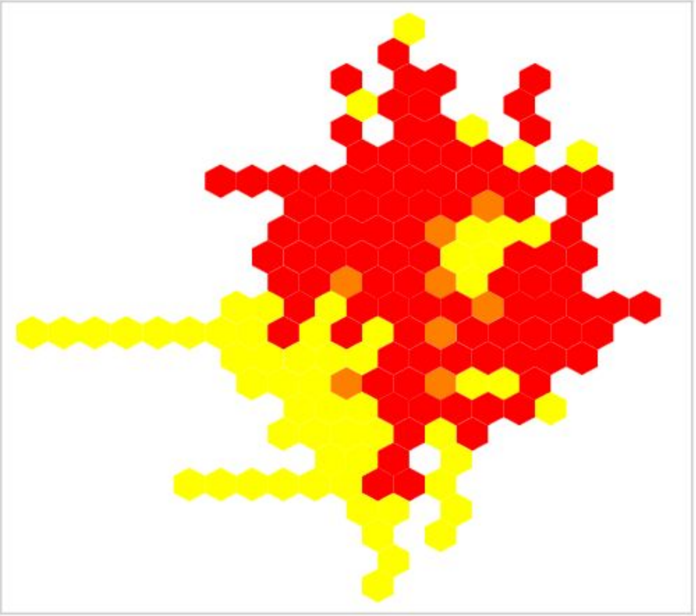
Explainability Results: NSL-KDD



U-Matrix of the trained GSOM model
Darker nodes are closer together

- **Unified Distance Matrix (U-Matrix)**
  - Displays all the nodes on the map
  - Visualizes **Clusters**
  - Darker nodes are closer to their neighbors
  - Lighter values show separation
- **4 to 5** well **separated clusters**
  - Lines indicate patterns found

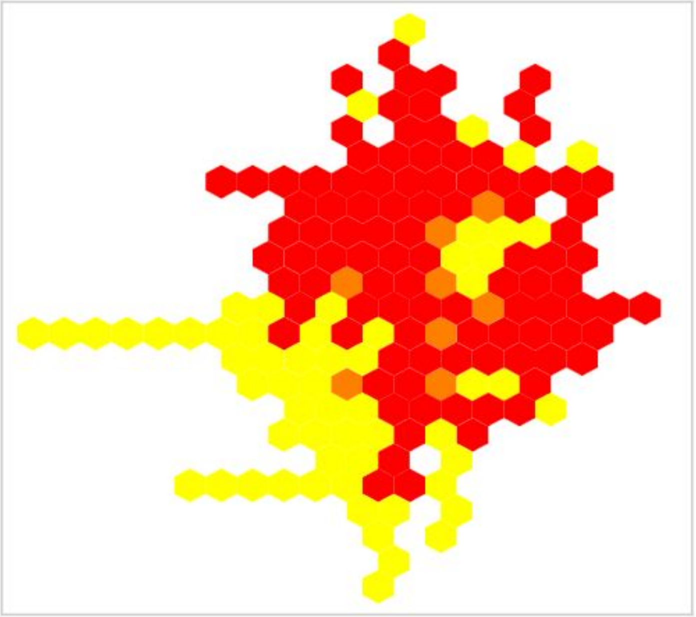
Explainability Results: NSL-KDD



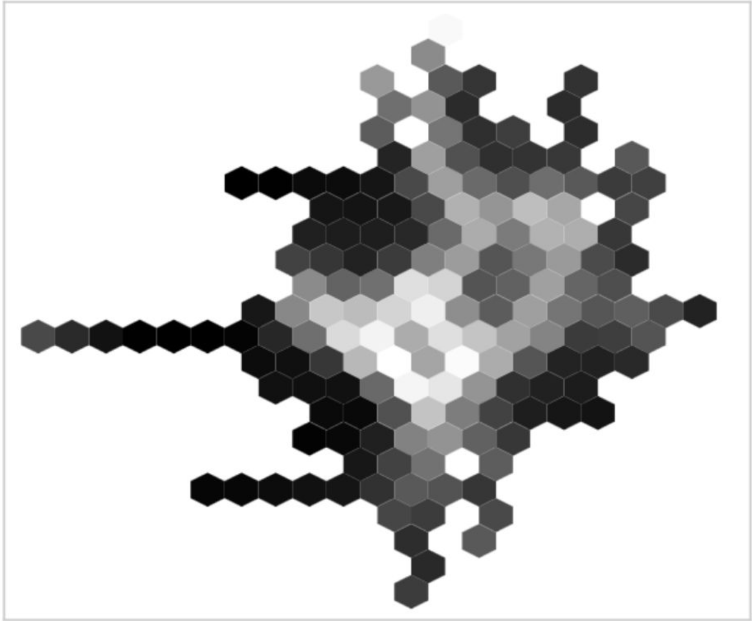
Label Map for GSOM model
Red = Benign (0), Yellow = Malicious (1)

- **Label Map**
  - Displays all nodes in the map
  - Yellow = **Malicious**
  - Red = **Benign**
  - Orange = **Probability Selection***

Explainability Results: NSL-KDD

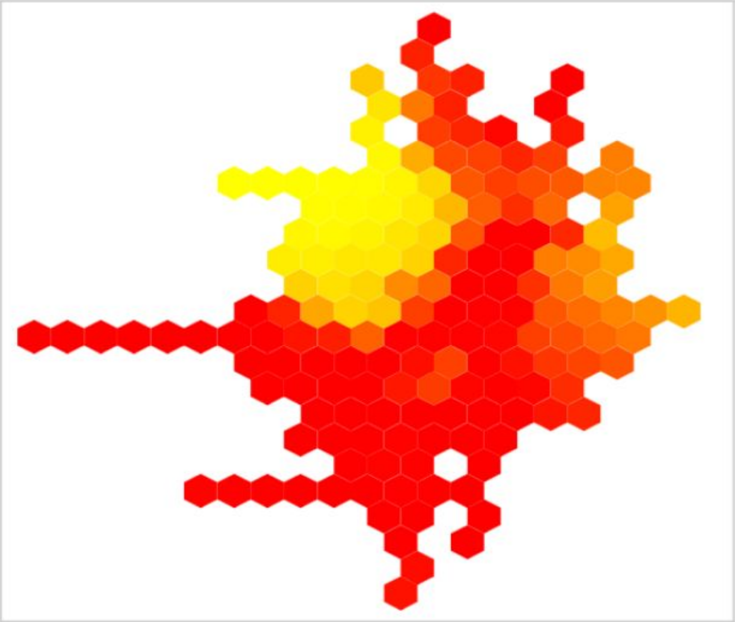


Label Map for GSOM model



U-Matrix of the trained GSOM model

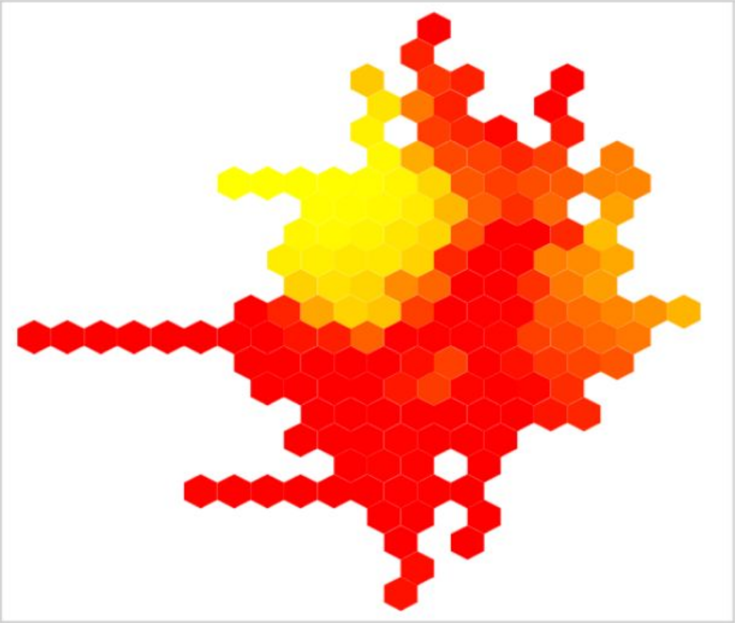
Explainability Results: NSL-KDD



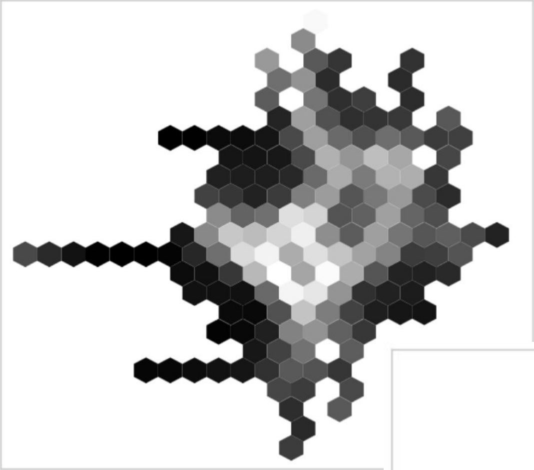
Feature Heatmap for Destination Bytes
Feature values closer to 1 are yellow

- **Feature Heatmap**
  - Displays all nodes in the map
  - All nodes contain a feature representation between [0, 1]
  - Lighter values are closer to 1
- **Example: Destination Bytes**
  - Upper-left quadrant contain **dst bytes** values closer to 1

Explainability Results: NSL-KDD



Feature Heatmap for Destination Bytes

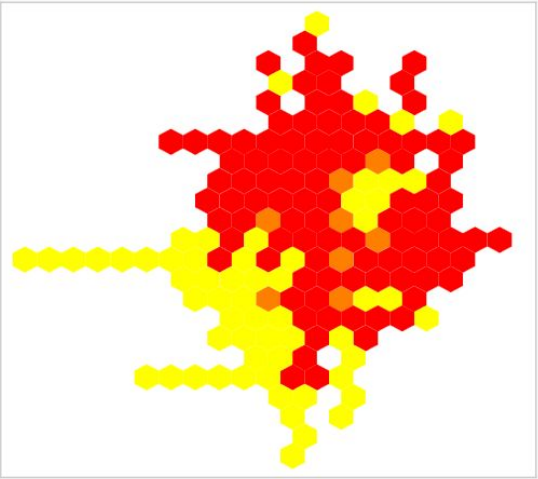


⇧

U-Matrix of the trained GSOM model

Label Map for GSOM model

⇩



Accuracy Results: NSL-KDD

NSL-KDD								
	SOM	GSOM	GHSOM	P-GHSOM	NDNN Jia et al. [6]	CNN Mohammadpour et al. [7]	BGRU+MLP Xu et al. [8]	BAT-MC Su et al. [9]
Accuracy	90.9%	96.7%	98.2%	98.0%	95.0%	99.8%	99.3%	99.2%
Precision	97.2%	96.6%	98.0%	98.0%	-	-	-	-
Recall	83.3%	96.5%	98.3%	97.8%	97.4%	-	99.3%	-
F1	89.7%	96.6%	98.1%	97.9%	91.4%	-	-	-
FPR	2.2%	3.1%	1.9%	1.8%	-	-	0.8%	-
FNR	16.6%	3.5%	1.6	2.2%	-	-	-	-
Network Size	1	1	7288	574	-	-	-	-
Training Time (s)	8	60	692	816	-	-	-	-
Prediction Time (ms)	.03	.03	.06	.04	-	-	-	-

Accuracy Results: NSL-KDD

NSL-KDD								
	SOM	**GSOM**	**GHSOM**	**P-GHSOM**	**NDNN Jia et al. [6]**	**CNN Mohammadpour et al. [7]**	**BGRU+MLP Xu et al. [8]**	**BAT-MC Su et al. [9]**
Accuracy	90.9%	96.7%	98.2%	98.0%	95.0%	99.8%	99.3%	99.2%
Precision	97.2%	96.6%	98.0%	98.0%	-	-	-	-
Recall	83.3%	96.5%	98.3%	97.8%	97.4%	-	99.3%	-
F1	89.7%	96.6%	98.1%	97.9%	91.4%	-	-	-
FPR	2.2%	3.1%	1.9%	1.8%	-	-	0.8%	-
FNR	16.6%	3.5%	1.6	2.2%	-	-	-	-
Network Size	1	1	7288	574	-	-	-	-
Training Time (s)	8	60	692	816	-	-	-	-
Prediction Time (ms)	.03	.03	.06	.04	-	-	-	-

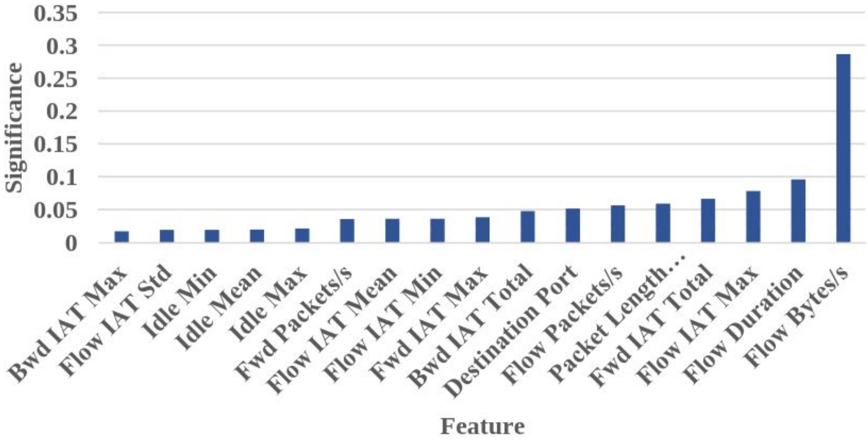
Accuracy Results: NSL-KDD

NSL-KDD								
	SOM	GSOM	GHSOM	P-GHSOM	NDNN Jia et al. [6]	CNN Mohammadpour et al. [7]	BGRU+MLP Xu et al. [8]	BAT-MC Su et al. [9]
Accuracy	90.9%	96.7%	98.2%	98.0%	95.0%	99.8%	99.3%	99.2%
Precision	97.2%	96.6%	98.0%	98.0%	-	-	-	-
Recall	83.3%	96.5%	98.3%	97.8%	97.4%	-	99.3%	-
F1	89.7%	96.6%	98.1%	97.9%	91.4%	-	-	-
FPR	2.2%	3.1%	1.9%	1.8%	-	-	0.8%	-
FNR	16.6%	3.5%	1.6	2.2%	-	-	-	-
Network Size	1	1	7288	574	-	-	-	-
Training Time (s)	8	60	692	816	-	-	-	-
Prediction Time (ms)	.03	.03	.06	.04	-	-	-	-

Accuracy Results: NSL-KDD

NSL-KDD								
	SOM	**GSOM**	**GHSOM**	**P-GHSOM**	**NDNN Jia et al. [6]**	**CNN Mohammadpour et al. [7]**	**BGRU+MLP Xu et al. [8]**	**BAT-MC Su et al. [9]**
Accuracy	90.9%	96.7%	98.2%	98.0%	95.0%	99.8%	99.3%	99.2%
Precision	97.2%	96.6%	98.0%	98.0%	-	-	-	-
Recall	83.3%	96.5%	98.3%	97.8%	97.4%	-	99.3%	-
F1	89.7%	96.6%	98.1%	97.9%	91.4%	-	-	-
FPR	2.2%	3.1%	1.9%	1.8%	-	-	0.8%	-
FNR	16.6%	3.5%	1.6	2.2%	-	-	-	-
Network Size	1	1	7288	574	-	-	-	-
Training Time (s)	8	60	692	816	-	-	-	-
Prediction Time (ms)	.03	.03	.06	.04	-	-	-	-

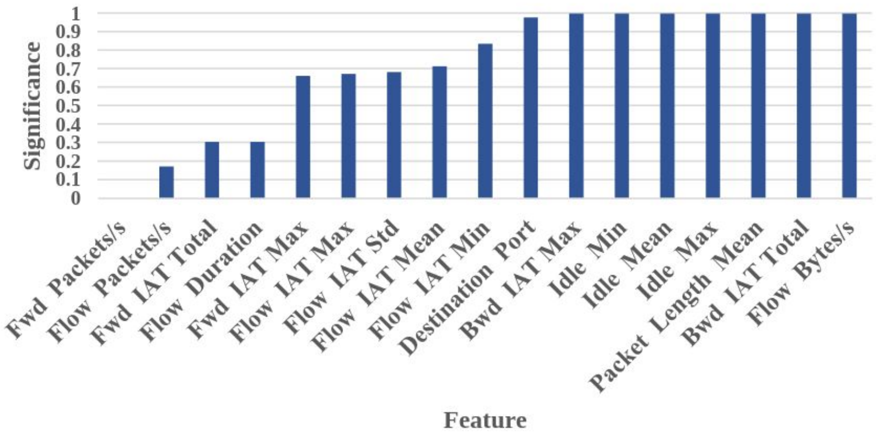
Explainability Results: CICIDS-2017



Global feature significance of the trained GSOM model
Higher values are more significant

- **Global Feature Significance**
  - Features - Dataset features
  - Significance - Features with higher variability that change labels
- Most significant feature is **Flow Bytes/s**
  - Other features have less variance

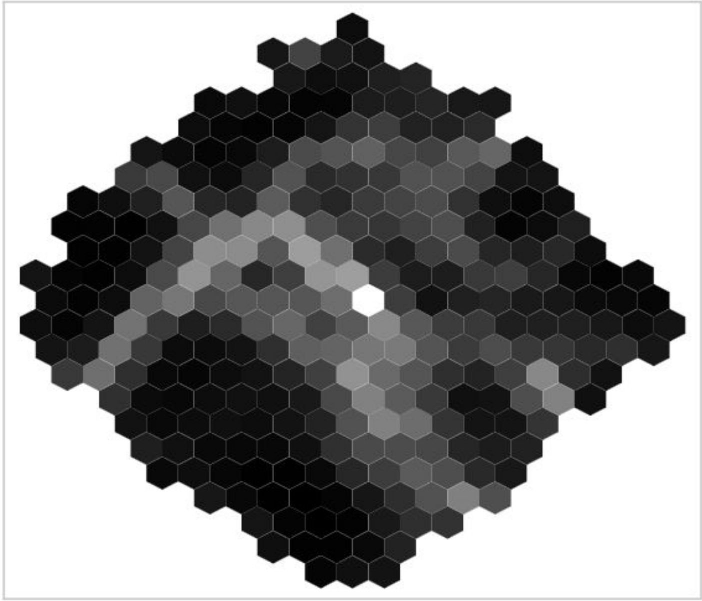
Explainability Results: CICIDS-2017



Local feature significance for an Anomalous sample
Higher values are more significant

- **Local Feature Significance**
  - Features - Dataset features
  - Significance - Impact on prediction
- **Anomalous** sample
  - Many similar features
  - 8 features close to 1
  - Why?

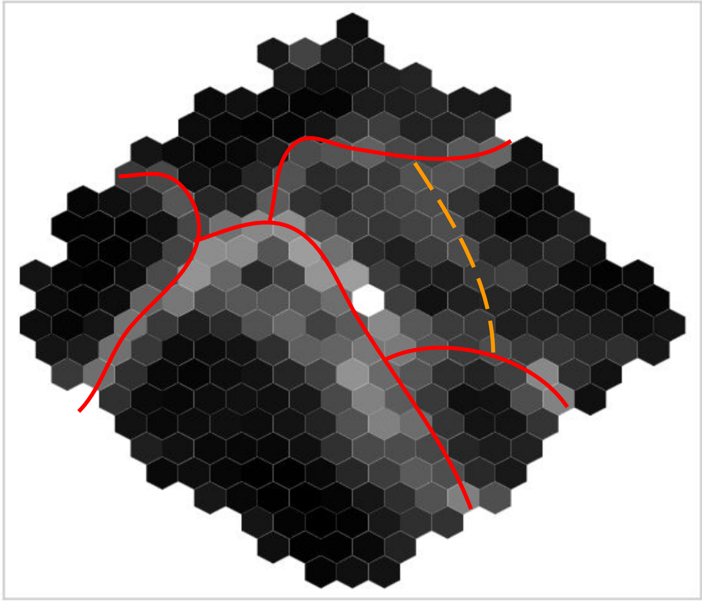
Explainability Results: CICIDS-2017



U-Matrix of the trained GSOM model
Darker nodes are closer together

- **U-Matrix**
  - Displays all the nodes in the map
  - Visualizes **clusters**
  - Darker nodes are closer to one another
  - Lighter colors show separation of clusters
- **5 to 6** well **separated clusters**
  - Uniform node spread

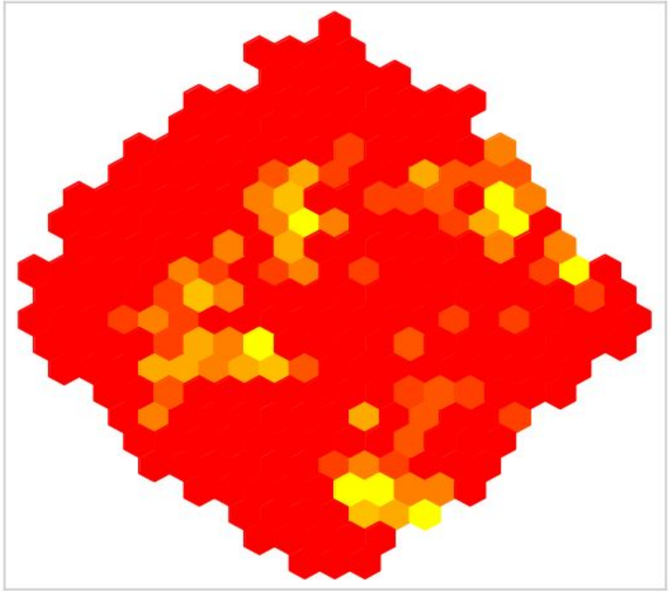
Explainability Results: CICIDS-2017



U-Matrix of the trained GSOM model
Darker nodes are closer together

- **U-Matrix**
  - Displays all the nodes in the map
  - Visualizes **clusters**
  - Darker nodes are closer to one another
  - Lighter colors show separation of clusters
- **5 to 6** well **separated clusters**
  - Uniform node spread

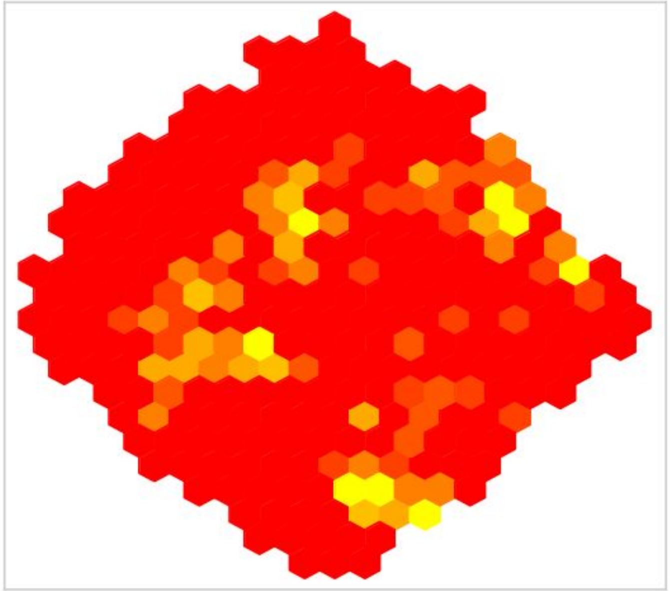
Explainability Results: CICIDS-2017



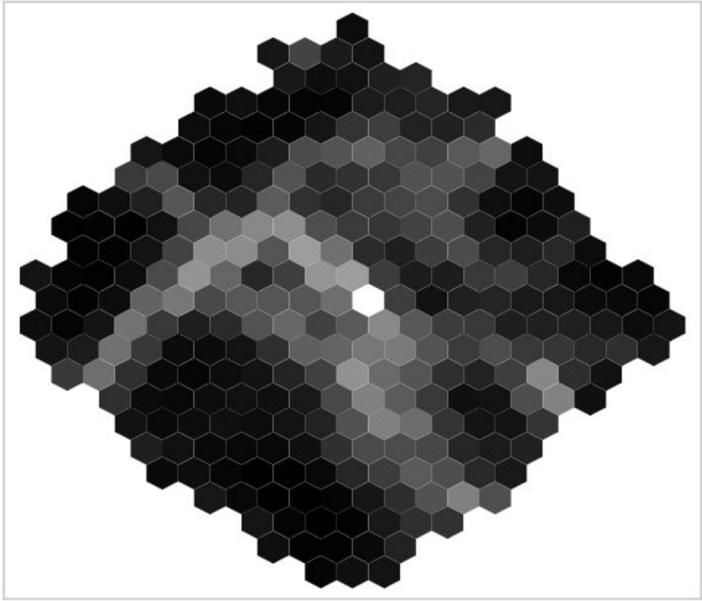
Label Map for GSOM model
Red = Benign (0), Yellow = Malicious (1)

- **Label Map**
  - Displays all nodes in the map
  - Yellow = **Malicious**
  - Red = **Benign**
  - Orange = **Probability Selection***

Explainability Results: CICIDS-2017

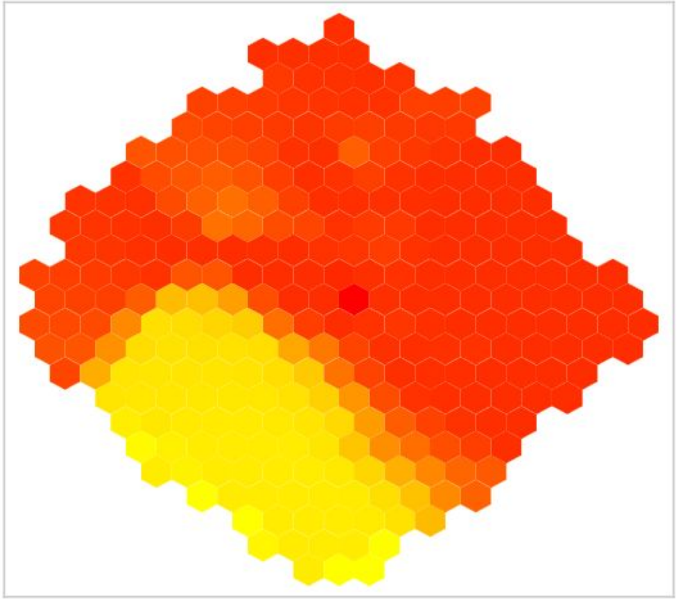


Label Map for GSOM model



U-Matrix of the trained GSOM model

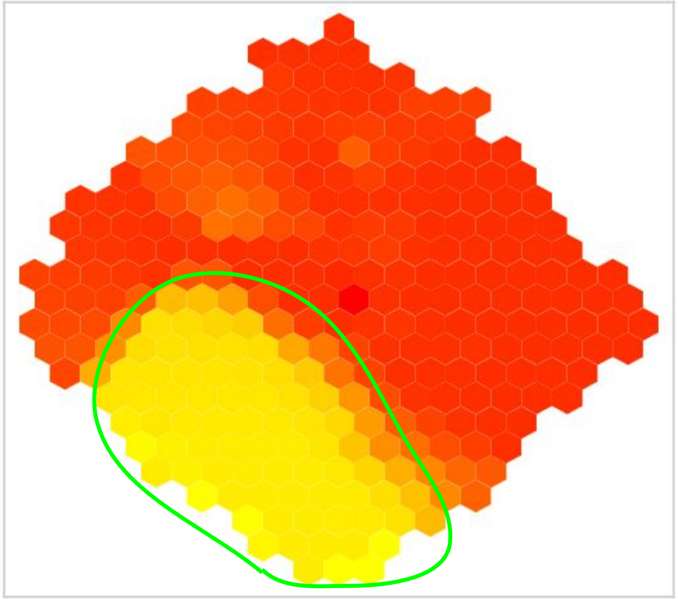
Explainability Results: CICIDS-2017



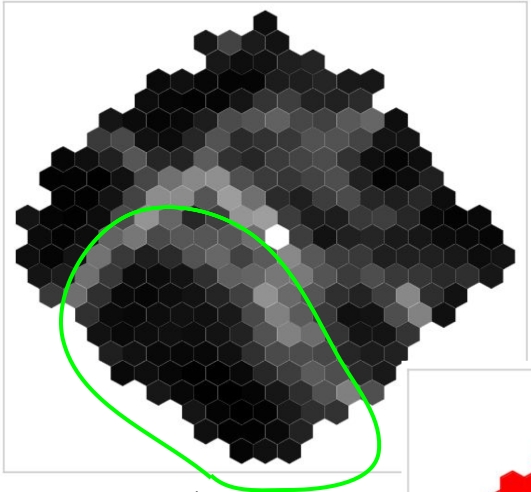
Feature Heatmap for Flow Bytes/s
Feature values closer to 1 are yellow

- **Feature Heatmap**
  - Displays all nodes in the map
  - All nodes contain a feature representation between [0, 1]
  - Lighter values are closer to 1
- **Example: Flow Bytes/s**
  - Bottom-left area contain **Flow Bytes/s** values closer to 1

Explainability Results: CICIDS-2017

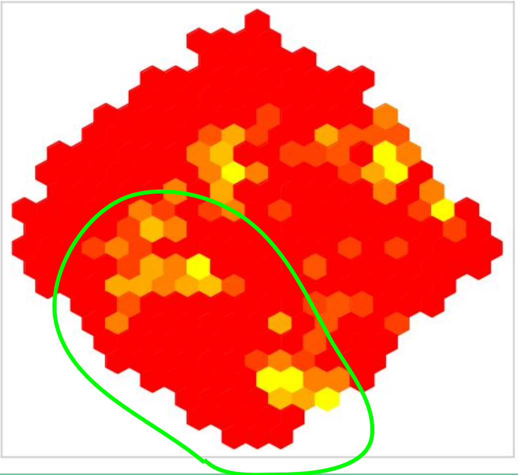


Feature Heatmap for Flow Bytes/s

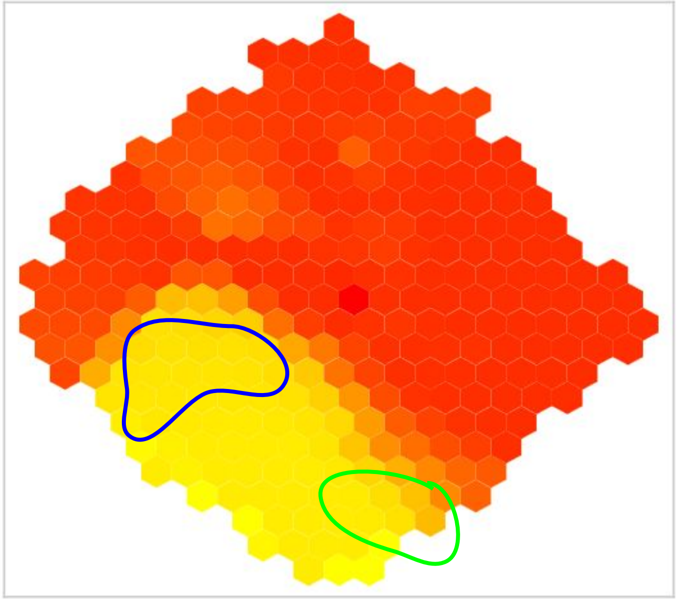


U-Matrix of the trained GSOM model

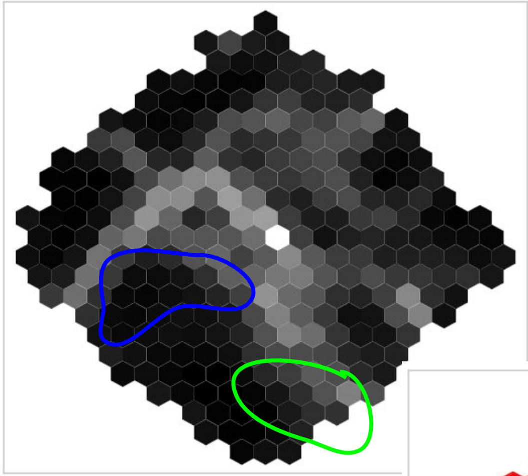
Label Map for GSOM model



Explainability Results: CICIDS-2017

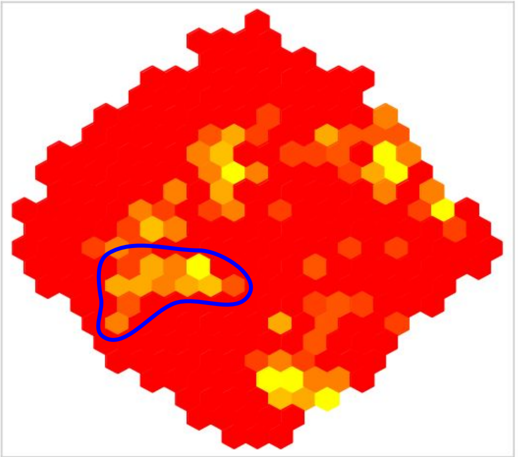


Feature Heatmap for Flow Bytes/s



U-Matrix of the trained GSOM model

Label Map for GSOM model



Accuracy Results: CICIDS-2017

CIC-IDS-2017								
	SOM	GSOM	GHSOM	P-GHSOM	SDCNN Khan et al. [10]	DNN+RE Almutlaq et al. [11]	SS-Deep-ID Abdel-Basset et al. [12]	CNN-IDS* Halbouni et al. [13]
Accuracy	79.4%	94.6%	96.7%	95.7%	99.3%	97.4%	99.6%	99.6%
Precision	83.2%	83.7%	89.1%	86.5%	99.1%	98.3%	99.5%	99.7%
Recall	42.0%	90.0%	94.5%	92.7%	99.7%	99.2%	99.2%	99.4%
F1	55.8%	86.7%	91.7%	89.5%	99.4%	98.3%	99.4%	99.7%
FPR	19.0%	4.3%	2.8%	3.5%	1.0%	-	0.7%	0.5%
FNR	23.0%	10.0%	5.5%	7.3%	1.0%	-	0.5%	-
Network Size	1	1	16894	119	-	-	-	-
Training Time (s)	260	1820	4299	11205	-	-	-	-
Prediction Time (ms)	.03	.06	1.5	.03	-	-	-	-

Accuracy Results: CICIDS-2017

CIC-IDS-2017								
	SOM	GSOM	GHSOM	P-GHSOM	SDCNN Khan et al. [10]	DNN+RE Almutlaq et al. [11]	SS-Deep-ID Abdel-Basset et al. [12]	CNN-IDS* Halbouni et al. [13]
Accuracy	79.4%	94.6%	96.7%	95.7%	99.3%	97.4%	99.6%	99.6%
Precision	83.2%	83.7%	89.1%	86.5%	99.1%	98.3%	99.5%	99.7%
Recall	42.0%	90.0%	94.5%	92.7%	99.7%	99.2%	99.2%	99.4%
F1	55.8%	86.7%	91.7%	89.5%	99.4%	98.3%	99.4%	99.7%
FPR	19.0%	4.3%	2.8%	3.5%	1.0%	-	0.7%	0.5%
FNR	23.0%	10.0%	5.5%	7.3%	1.0%	-	0.5%	-
Network Size	1	1	16894	119	-	-	-	-
Training Time (s)	260	1820	4299	11205	-	-	-	-
Prediction Time (ms)	.03	.06	1.5	.03	-	-	-	-

Accuracy Results: CICIDS-2017

CIC-IDS-2017								
	SOM	GSOM	GHSOM	P-GHSOM	SDCNN Khan et al. [10]	DNN+RE Almutlaq et al. [11]	SS-Deep-ID Abdel-Basset et al. [12]	CNN-IDS* Halbouni et al. [13]
Accuracy	79.4%	94.6%	96.7%	95.7%	99.3%	97.4%	99.6%	99.6%
Precision	83.2%	83.7%	89.1%	86.5%	99.1%	98.3%	99.5%	99.7%
Recall	42.0%	90.0%	94.5%	92.7%	99.7%	99.2%	99.2%	99.4%
F1	55.8%	86.7%	91.7%	89.5%	99.4%	98.3%	99.4%	99.7%
FPR	19.0%	4.3%	2.8%	3.5%	1.0%	-	0.7%	0.5%
FNR	23.0%	10.0%	5.5%	7.3%	1.0%	-	0.5%	-
Network Size	1	1	16894	119	-	-	-	-
Training Time (s)	260	1820	4299	11205	-	-	-	-
Prediction Time (ms)	.03	.06	1.5	.03	-	-	-	-

Accuracy Results: CICIDS-2017

CIC-IDS-2017								
	SOM	GSOM	GHSOM	P-GHSOM	SDCNN Khan et al. [10]	DNN+RE Almutlaq et al. [11]	SS-Deep-ID Abdel-Basset et al. [12]	CNN-IDS* Halbouni et al. [13]
Accuracy	79.4%	94.6%	96.7%	95.7%	99.3%	97.4%	99.6%	99.6%
Precision	83.2%	83.7%	89.1%	86.5%	99.1%	98.3%	99.5%	99.7%
Recall	42.0%	90.0%	94.5%	92.7%	99.7%	99.2%	99.2%	99.4%
F1	55.8%	86.7%	91.7%	89.5%	99.4%	98.3%	99.4%	99.7%
FPR	19.0%	4.3%	2.8%	3.5%	1.0%	-	0.7%	0.5%
FNR	23.0%	10.0%	5.5%	7.3%	1.0%	-	0.5%	-
Network Size	1	1	16894	119	-	-	-	-
Training Time (s)	260	1820	4299	11205	-	-	-	-
Prediction Time (ms)	.03	.06	1.5	.03	-	-	-	-

Conclusion

- Explainability and Intrusion Detection
- White-box Competitive Learning
- Our Proposed Architecture for X-IDS

References

[1] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2) (2019) 44–58.

[2] https://www.darpa.mil/program/explainable-artificial-intelligence

[3] Chompinha, Self Organizing Map Training Gif, https://commons.wikimedia.org/wiki/File:TrainSOM.gif

[4] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, A detailed analysis of the kdd cup 99 data set, in: 2009 IEEE symposium on computational intelligence for security and defense applications, Ieee, 2009, pp. 1–6

[5] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization., ICISSp 1 (2018) 108–116.

References

[6] Y. Jia, M. Wang, Y. Wang, Network intrusion detection algorithm based on deep neural network, IET Information Security 13 (1) (2019) 48–53.

[7] L. Mohammadpour, T. C. Ling, C. S. Liew, C. Y. Chong, A convolutional neural network for network intrusion detection system, Proceedings of the Asia-Pacific Advanced Network 46 (0) (2018) 50–55.

[8] C. Xu, J. Shen, X. Du, F. Zhang, An intrusion detection system using a deep neural network with gated recurrent units, IEEE Access 6 (2018) 48697–48707.

[9] T. Su, H. Sun, J. Zhu, S. Wang, Y. Li, Bat: Deep learning methods on network intrusion detection using nsl-kdd dataset, IEEE Access 8 (2020) 29575–29585.

[10] A. S. Khan, Z. Ahmad, J. Abdullah, F. Ahmad, A spectrogram image-based network anomaly detection system using deep convolutional neural network, IEEE Access 9 (2021) 87079–87093.

[11] S. Almutlaq, A. Derhab, M. M. Hassan, K. Kaur, Two-stage intrusion detection system in intelligent transportation systems using rule extraction methods from deep neural networks, IEEE Transactions on Intelligent Transportation Systems (2022).

[12] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, M. J. Ryan, Semi-supervised spatiotemporal deep learning for intrusions detection in iot networks, IEEE Internet of Things Journal 8 (15) (2021) 12251–12265.

[13] A. H. Halbouni, T. S. Gunawan, M. Halbouni, F. A. A. Assaig, M. R. Effendi, N. Ismail, Cnn-ids: Convolutional neural network for network intrusion detection system, in: 2022 8th International Conference on Wireless and Telematics (ICWT), IEEE, 2022, pp. 1–4.