CERIAS
The Center for Education and Research in Information Assurance and Security

# E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection

Osvaldo Arreche[1], Tanish Guntur[2], Jack Roberts[3] and Mustafa Abdallah[1]
[1]Indiana University-Purdue University Indianapolis, Indiana, USA
oarreche@iu.edu tguntur@iu.edu robertjaw@iu.edu mabdall@iu.edu

## Introduction

• The research focuses on the increasing need for **artificial intelligence (AI)** techniques in **intrusion detection systems (IDS)** due to the rising number of network intrusions.

• The utilization of **explainable AI (XAI)** methods is crucial for **enhancing the transparency and interpretability** of AI models in real-world **IDS**.

• The study introduces an **end-to-end framework** for **evaluating black-box XAI methods** for network intrusion detection, aiming to provide insights into the strengths and limitations of these techniques.

## Motivation

• The **lack of interpretability** in current AI-based **IDS** poses **challenges** for security analysts, requiring them to sift through large volumes of data to identify abnormal behaviors.

• Achieving **high accuracy** in results and providing **comprehensible explanations** for AI algorithms are **essential** in **network intrusion** detection to build **trust** and understanding among security **analysts**.

• **XAI frameworks** play a vital role in assisting analysts by **elucidating AI decisions**, facilitating **efficient** investigations, and ultimately **saving time** for human security analysts in **network security tasks.**

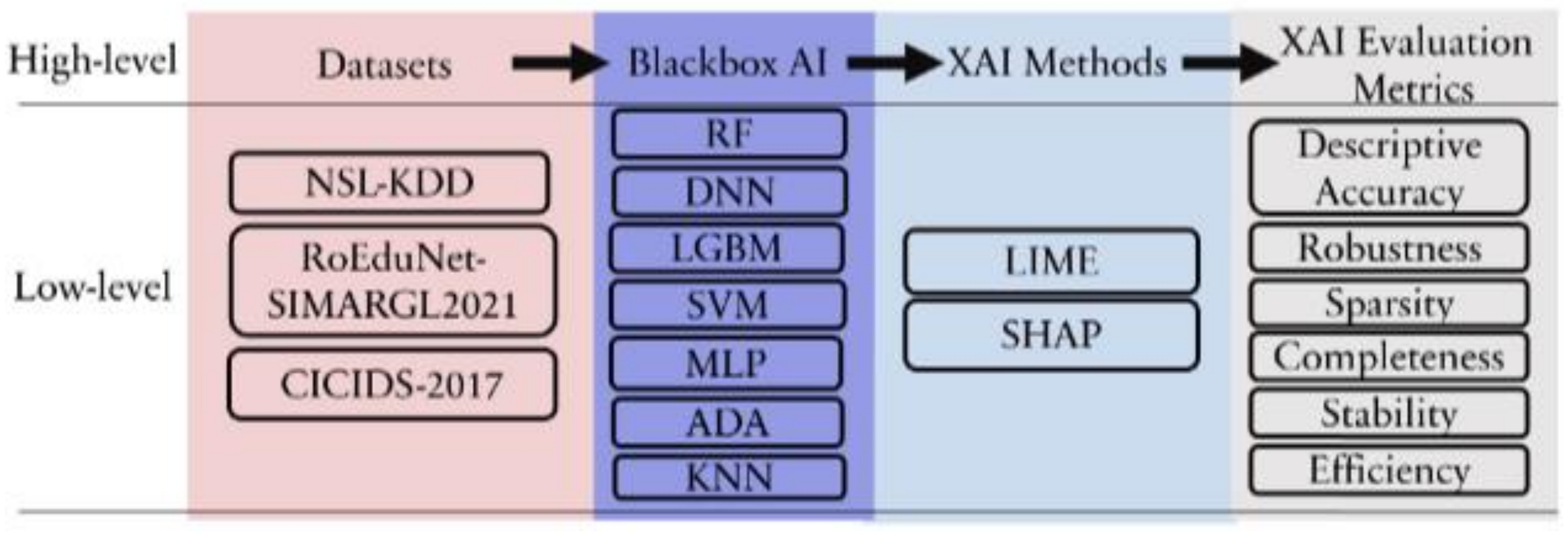
## Our Contribution

• The research proposes an **end-to-end framework for evaluating XAI techniques for network** intrusion detection tasks, assessing both **global and local explanations**.

• **Six** different **evaluation metrics** are analyzed for two popular black-box XAI techniques, **SHAP** and **LIME**, under **three** network intrusion **datasets** and **seven AI models**.

• The **source codes** of the evaluation framework are **released to the community**, serving as a baseline for XAI evaluation in network intrusion detection and encouraging further development with new datasets and models.

## Framework



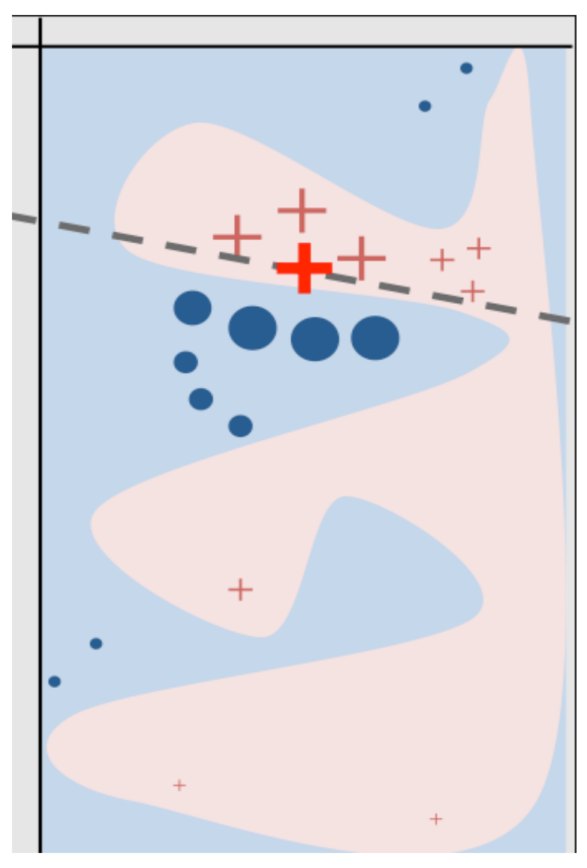
## XAI - Evaluation

### Datasets

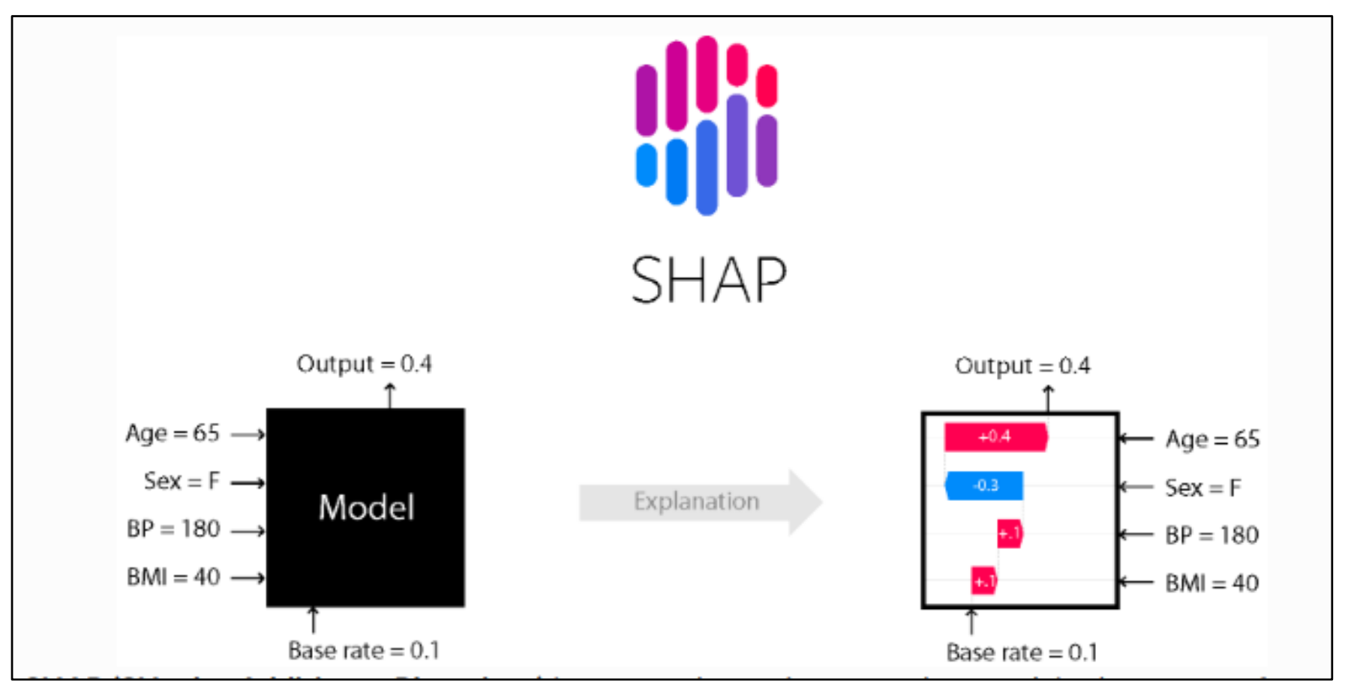
| Dataset | Number of Labels | Number of Features | Number of Samples |
|---|---|---|---|
| CICIDS-2017 | 7 | 78 | 2,775,364 |
| RoEduNet-SIMARGL2021 | 3 | 29 | 31,433,875 |
| NSL-KDD | 5 | 41 | 148517 |

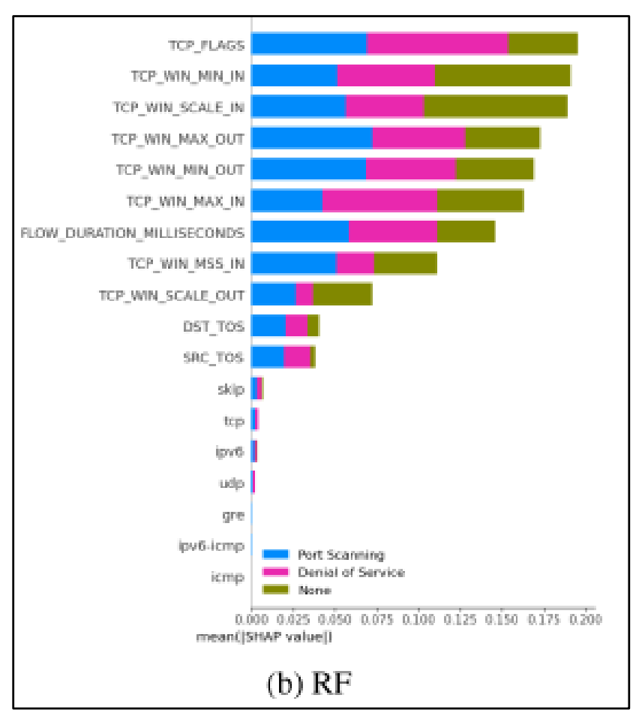
### XAI Methods

LIME



marcotcr/lime: Lime: Explaining the predictions of any machine learning classifier (github.com)





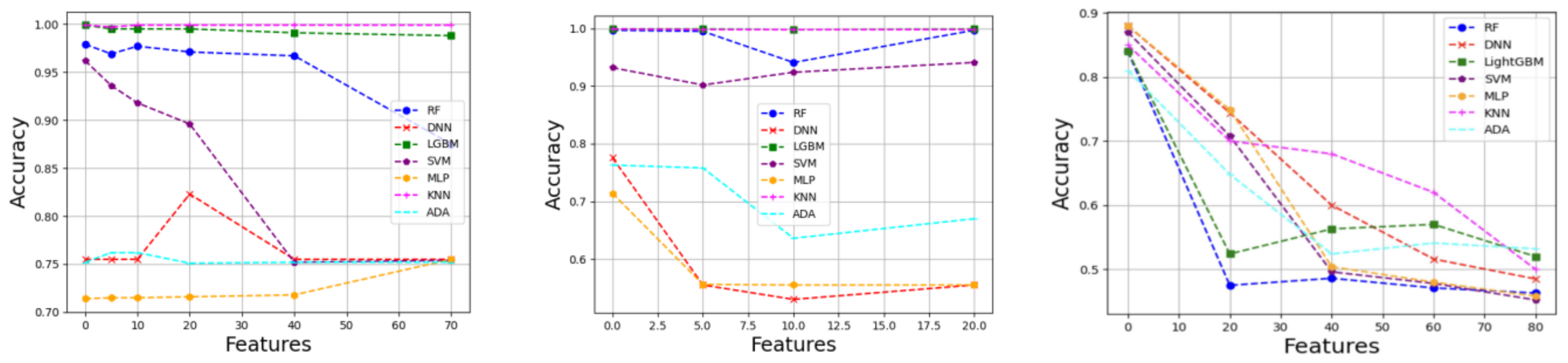
(b) RF

Welcome to the SHAP documentation — SHAP latest documentation

SHAP

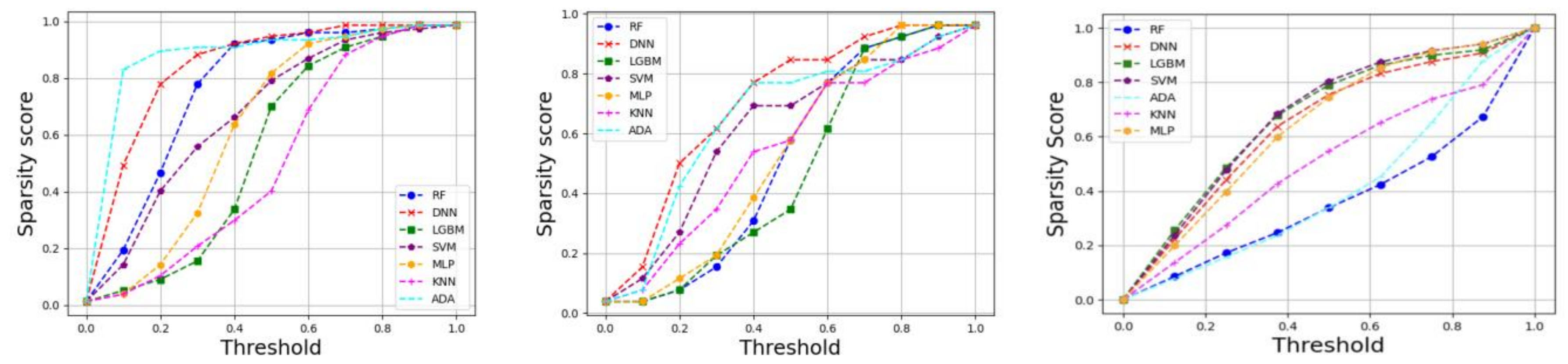
## Results

### Descriptive Accuracy



(a) CICIDS-2017 (b) RoEduNet-SIMARGL (c) NSL-KDD

### Sparsity



(a) CICIDS-2017 (b) RoEduNet-SIMARGL (c) NSL-KDD

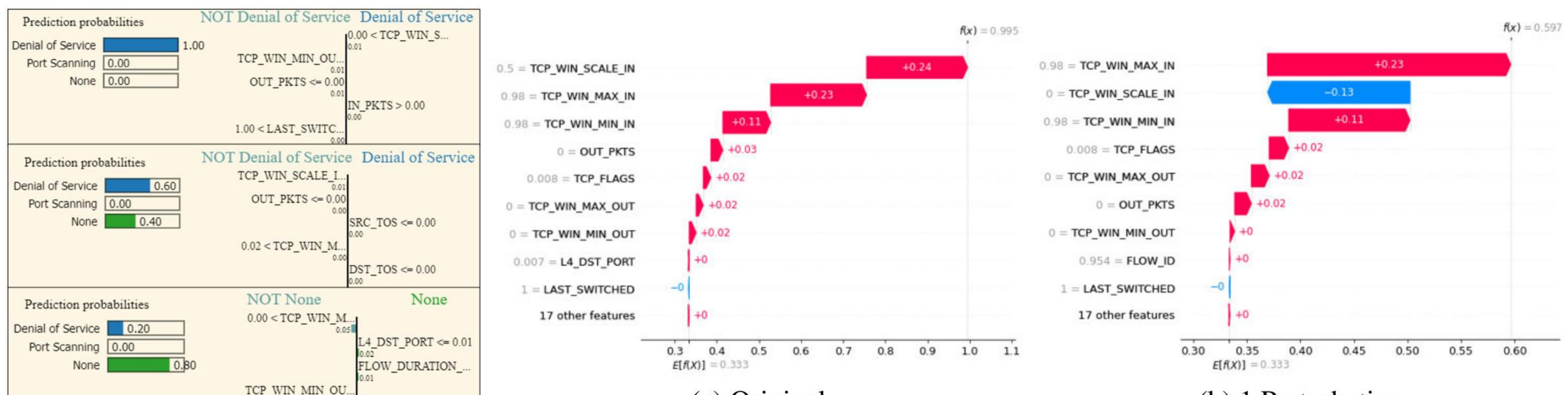
### Stability

| Dataset (XAI Method) | RF | DNN | LGBM | ADA | MLP | SVM | KNN |
|---|---|---|---|---|---|---|---|
| CICIDS-2017 (LIME) | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 0.99 |
| CICIDS-2017 (SHAP) | 0.35 | 0.50 | 0.65 | 1.00 | 0.28 | 1.00 | 1.00 |
| RoEduNet-SIMARGL2021 (LIME) | 1.00 | 0.20 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| RoEduNet-SIMARGL2021 (SHAP) | 1.00 | 0.40 | 0.60 | 1.00 | 1.00 | 1.00 | 1.00 |
| NSL-KDD (LIME) | 1.00 | 0.95 | 0.90 | 0.90 | 0.95 | 0.65 | 0.80 |
| NSL-KDD (SHAP) | 1.00 | 0.95 | 1.00 | 0.95 | 1.00 | 1.00 | 0.85 |

### Efficiency

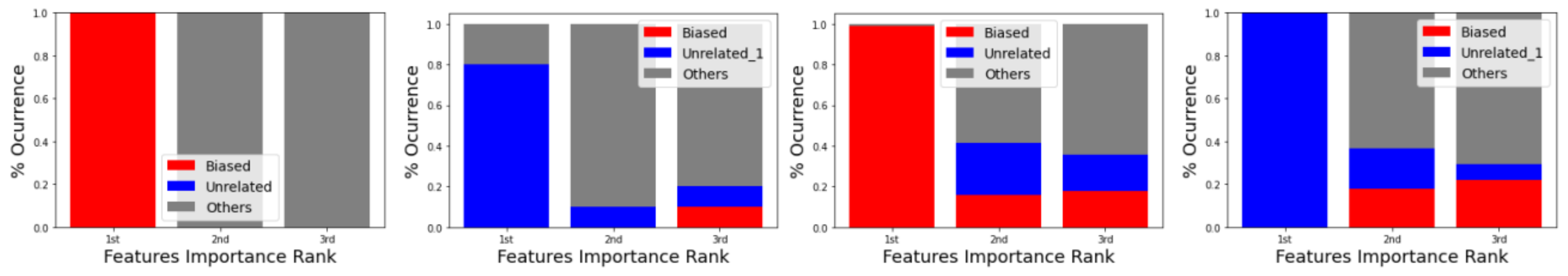
| SHAP Efficiency (Samples x Hours) | RF | DNN | LGBM | MLP | ADA | SVM | KNN |
|---|---|---|---|---|---|---|---|
| 100 | 0.001 | 0.001 | 0.001 | 0.014 | 0.129 | 0.041 | 1.104 |
| 500 | 0.002 | 0.004 | 0.001 | 0.925 | 3.557 | 0.841 | 28.33 |
| 2500 | 0.011 | 0.029 | 0.002 | 21.18 | 99.00 | 20.87 | 706.0 |
| 10000 | 0.046 | 0.306 | 0.006 | 385.0 | 1512 | 343.0 | 1654 |
| **LIME Efficiency (Samples x Hours)** | **RF** | **DNN** | **LGBM** | **MLP** | **ADA** | **SVM** | **KNN** |
| 1000 (RoEduNet-SIMARGL2021 ) | 1.401 | 1.452 | 1.423 | 1.466 | 1.416 | 1.483 | 0.333 |
| 1000 (CICIDS-2017 ) | 3.812 | 3.766 | 3.750 | 3.855 | 3.956 | 3.816 | 15.81 |
| 1000 (NSL-KDD) | 2.287 | 2.110 | 2.337 | 2.389 | 2.478 | 3.112 | 12.87 |
| **SHAP Local Efficiency (Minutes)** | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **LIME Local Efficiency (Minutes)** | 0.351 | 0.311 | 0.294 | 0.284 | 0.294 | 0.306 | 0.357 |

### Completeness



(a) Original (b) 1 Perturbation

### Robustness



(a) SHAP - Biased (b) SHAP - Advers. (c) LIME - Biased (d) LIME - Advers.

## Key Takeaways

• Creation of a **novel end-to-end framework for evaluating XAI techniques for network intrusion detection** tasks, assessing both global and local explanations.

• **XAI techniques still need improvement in the six metrics** to be applied in a production environment, but this work is **an important step in this direction**.

| Metric | Descriptive Accuracy | Sparsity | Stability | Completeness | Robustness | Efficiency |
|---|---|---|---|---|---|---|
| SHAP | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| LIME | | | ✓ | | ✓ | ✓ |

## Acknowledgement

This work was supported by AnalytixIN, Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER); By the 1st Year Research Immersion Program (1RIP) Grants from the Office of the Vice Chancellor for Research at Indiana University–Purdue University Indianapolis; and in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.