Explainable AI-based Intrusion Detection in the Internet of Things

M. Siganos et al.
K3Y LTD
msiganos@k3y.bg

Authors: Marios Siganos, Panagiotis Radoglou-Grammatikis, Igor Kotsiuba, Evangelos Markakis, Ioannis Moscholios, Sotirios Goudos, Panagiotis Sarigiannidis

Acknowledgments

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070214 (TRUSTEE).

Disclaimer: Funded by the European Union.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission.

Neither the European Union nor the European Commission can be held responsible for them.

TRUSTEE

Introduction

- cyberthreats have grown in sophistication and scope
- Intrusion Detection Systems (IDS) are important for the detection of potential cyberattacks and anomalies in a timely manner
- IDS can be classified into two main categories:
  - signature/specification-based detection - pre-defined patterns
  - anomaly-based detection - statistical analysis and Artificial Intelligence (AI)
- AI-powered IDS have already demonstrated their efficiency
  - but they suffer from false alarms and explainability issues
- development of an AI-powered IDS for the IoT, including explainable AI (XAI) functions

Related Work
Cybersecurity mechanisms with XAI

Zebin et al. (2022)	Patil et al. (2022)	Barnard et al. (2022)	Mane and Rao (2021)	Wang et al. (2020)
• XAI solution for the detection of DNS over HTTPS (DoH) attacks • balanced and stacked Random Forest classifier • CIRA-CIC-DoHBrw-2020 dataset • SHAP	• XAI for intrusion detection • voting classifier that utilises an ensemble of several models • CICIDS2017 dataset • LIME	• A framework for network intrusion detection using XAI • Gradient Boosting (XGboost) • NSL-KDD dataset • SHAP	• XAI for the creation of a network intrusion detection system • fully connected network with three hidden layers • NSL-KDD dataset • SHAP, LIME, CEM	• A framework that uses ML and XAI for IDS • a one-vs-all and a multiclass classifier based on fully connected networks • NSL-KDD dataset • SHAP

Related Work
Cybersecurity mechanisms with XAI

- provide useful solutions and methodologies
- none of them considers the unique characteristics of Internet of Things and Industrial Internet of Things network environments of Critical Infrastructures, such as the smart electrical grid

Contributions

- Implementation of an AI-powered IDS for the IoT
  - utilized CIC-IoT-Dataset-2022 and IEC 69870-5-104 Intrusion Detection datasets
  - applied various Machine Learning (ML) / Deep Learning (DL)
- Investigating and development of explainability functions
  - provided an explainability mechanism (SHAP)

Proposed Intrusion Detection System

1. Network Traffic Data Capturing Module
2. Network Flow Generation Module
3. Data Pre-processing Module
4. Detection Module
5. Explainability Module
6. Notification Module

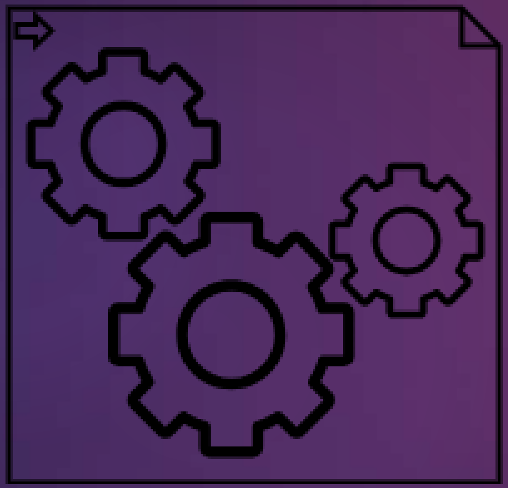
Network Traffic Data Capturing Module

- captures the network traffic data (i.e., pcap files)
- utilizes a Switch Port Analyzer (SPAN) (i.e., port mirroring) and tcpdump

Network Flow Generation Module

- generates flow statistics
  - TCP/IP network flow statistics
  - IEC 60870-5- 104 payload flow statistics
- reduces the volume of data
- provides a more meaningful representation of the network traffic data

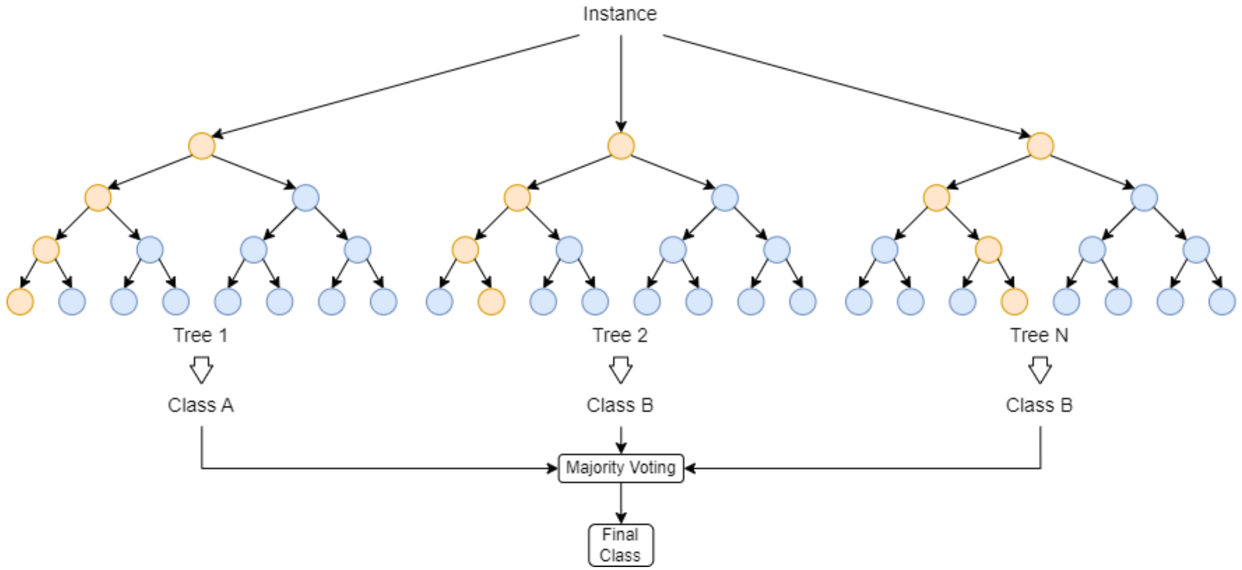
Data Pre-processing Module



- cleans the data and removes noise
  - handles missing values – rows with missing values are removed
  - handles label – categorical values are encoded with numerical ones
- performs feature scaling
  - scales data to the range [0, 1] or standardises features
- reduces feature dimensionality and performs feature selection and feature extraction
  - removes features with only one unique value, low variance (0.1) or Pearson correlation (0.9)
  - performs recursive feature elimination and sequential feature selection (forward and backward)

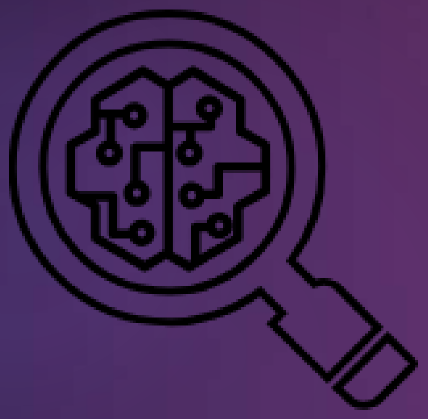
Detection Module

- discriminate potential attacks using pre-trained ML/DL models
  - IoT
  - IEC 60870-5-104 IIoT
- Random Forests is the best-performing model
  - Tree-based model
  - Ensemble method – bootstrap aggregating / bagging



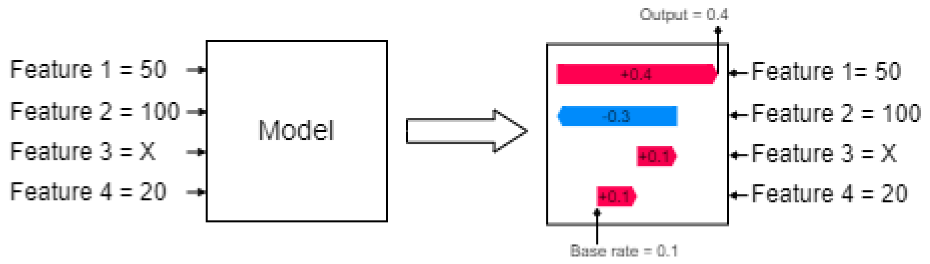
ARES conference

Explainability Module



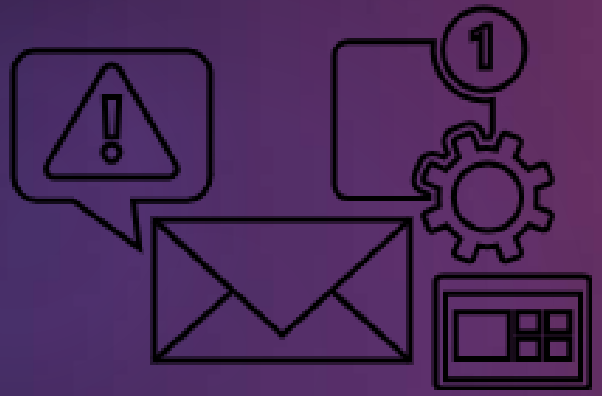
- consistent and reliable explanations
- model-agnostic post-hoc XAI techniques
- SHAP method (feature importances)
- local explanations – individual predictions
- global explanations – overview of the entire dataset
- visualizations through a dashboard

$$g\,(z') = \phi_0 + \sum_{\{i=1\}}^{M} \phi_i z_i'$$



$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right]$$

Notification Module



- alerts the security administrator
  - e-mail
  - Short Message/Messaging Service (SMS)
  - push notifications
  - dashboard that displays the intrusion details and explanation

Performance Evaluation

AI Models

- Naive Bayes
- SVM Linear
- SVM RBF
- Decision Trees
- Random Forest
- XGBoost
- Adaboost
- Logistic Regression
- Quadradic Discriminant Analysis
- DNN

Evaluation Metrics

- Accuracy $= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
- True Positive Rate (TPR) $= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
- False Positive Rate (FPR) $= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{FN}}$
- F1 Score $= \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$

Datasets

IEC 60870-5-104

Parser: CICFlowMeter

Timeframes: 15, 30, 60, 90, 120, 180

Columns: 84

IEC 60870-5-104

Parser: Custom

Timeframes: 15, 30, 60, 90, 120, 180

Columns: 112

CIC-IoT-Dataset-2022

Parser: CICFlowMeter

Timeframes: NA

Columns: 84

CIC-IoT-Dataset-2022

Parser: NFStream

Timeframes: NA

Columns: 40

Evaluation results
IEC 60 870-5-104 - CICFlow

AI Models	Accuracy	TPR	FPR	F1-Score
Naïve Bayes	0.4196	0.4196	0.512	03554
SVM Linear	0.4944	0.4944	0.0453	0.4727
SVM RBF	0.4940	0.4940	0.0448	0.4538
Decision Trees	0.6007	0.6009	0.0363	0.5994
Random Forest	**0.6632**	**0.6634**	**0.0306**	**0.6601**
XGBoost	0.6358	0.6360	0.0330	0.6324
Adaboost	0.3532	0.3532	0.0574	0.3014
Logistic Regression	0.4841	0.4841	0.0463	0.4628
Quadratic Discriminant Analysis	0.5572	0.5572	0.0395	0.5236
DNN	0.5811	0.5811	0.0381	0.5586

Evaluation results

IEC 60 870-5-104 - Custom

AI Models	Accuracy	TPR	FPR	F1-Score
Naïve Bayes	0.5582	0.5582	0.0402	0.4749
SVM Linear	0.6514	0.6514	0.0317	0.6384
SVM RBF	0.5942	0.5942	0.0369	0.5588
Decision Trees	0.8333	0.8333	0.0152	0.8281
Random Forest	**0.8521**	**0.8521**	**0.0134**	**0.8473**
XGBoost	0.8348	0.8348	0.0150	0.8280
Adaboost	0.2826	0.2826	0.0652	0.2121
Logistic Regression	0.6223	0.6223	0.0343	0.6053
Quadratic Discriminant Analysis	0.6233	0.6233	0.0342	0.5594
DNN	0.6958	0.6958	0.0277	0.6851

Evaluation results
CIC IoT dataset 2022 - CICFlow

AI Models	Accuracy	TPR	FPR	F1-Score
Naïve Bayes	0.7428	0.7427	0.1287	0.7409
SVM Linear	0.9312	0.9311	0.0344	0.9314
SVM RBF	0.9583	0.9583	0.0209	0.9585
Decision Trees	0.9985	0.9985	0.0007	0.9985
Random Forest	0.9983	0.9983	0.0008	0.9983
XGBoost	**0.9992**	**0.9992**	**0.0004**	**0.9992**
Adaboost	0.9583	0.9583	0.0208	0.9582
Logistic Regression	0.9308	0.9308	0.0346	0.9311
Quadratic Discriminant Analysis	0.9363	0.9363	0.0319	0.9364
DNN	0.9888	0.9888	0.0056	0.9888

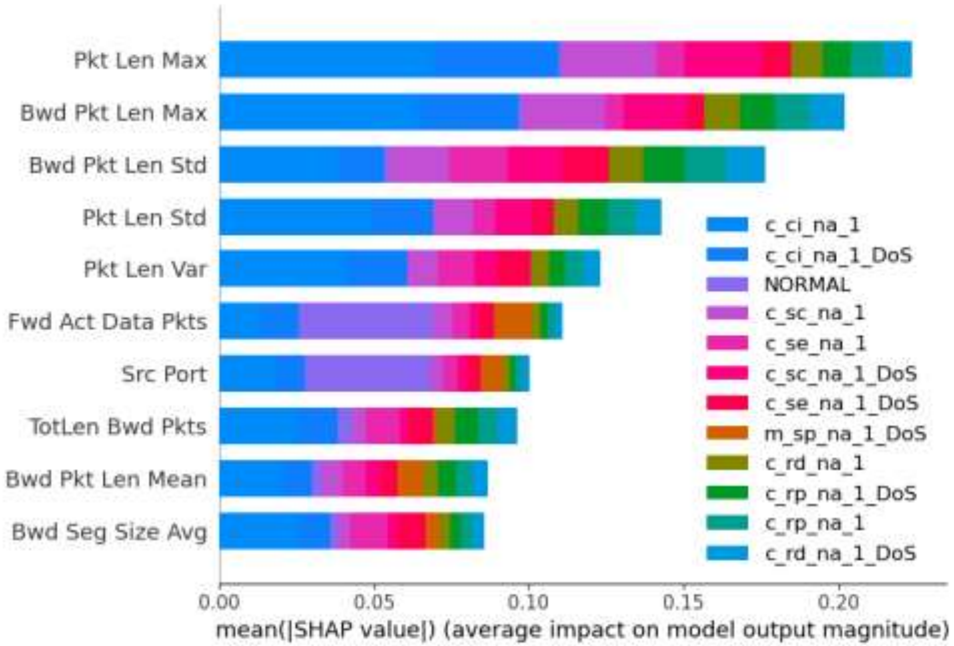
Evaluation results

CIC IoT dataset 2022 - NFStream

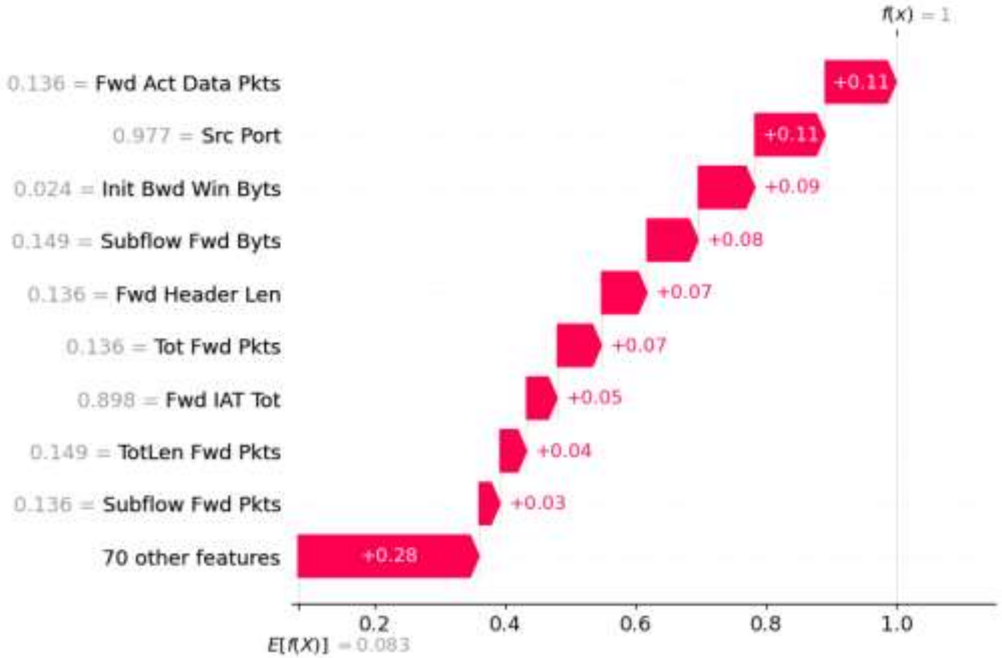
AI Models	Accuracy	TPR	FPR	F1-Score
Naïve Bayes	0.9700	0.9700	0.0150	0.9701
SVM Linear	0.9581	0.9581	0.0209	0.9583
SVM RBF	0.9879	0.9879	0.0060	0.9879
Decision Trees	0.9988	0.9988	0.0006	0.9988
Random Forest	**0.9999**	**0.9999**	**0.0000**	**0.9999**
XGBoost	0.9998	0.9998	0.0001	0.9998
Adaboost	0.9106	0.9106	0.0447	0.9112
Logistic Regression	0.9620	0.9620	0.0190	0.9621
Quadratic Discriminant Analysis	0.5530	0.5530	0.2235	0.5051
DNN	0.9985	0.9985	0.0007	0.9985

Explainability results
IEC 60 870-5-104 - CICFlow

SHAP Summary Plot

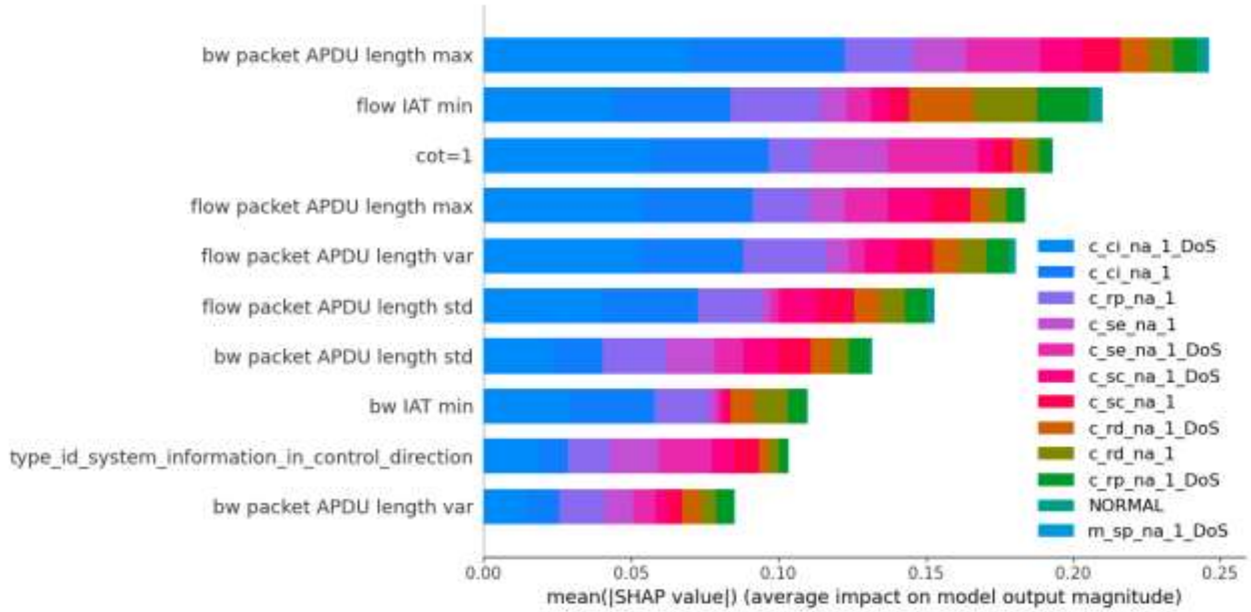


SHAP Waterfall Plot

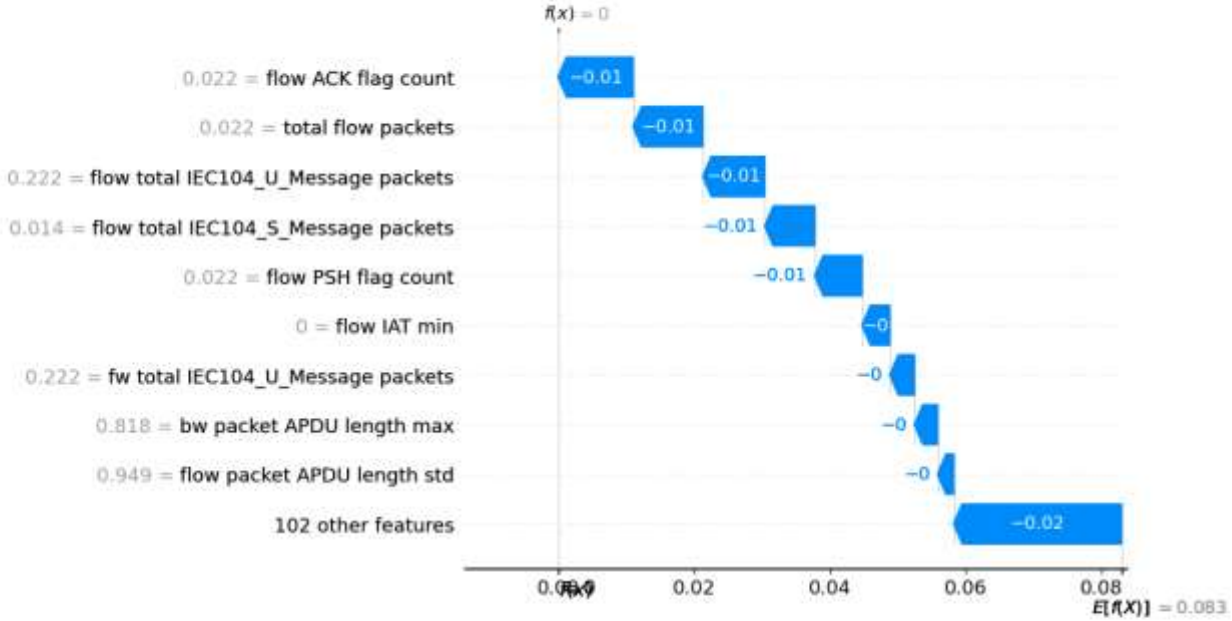


Explainability results
IEC 60 870-5-104 - Custom

SHAP Summary Plot

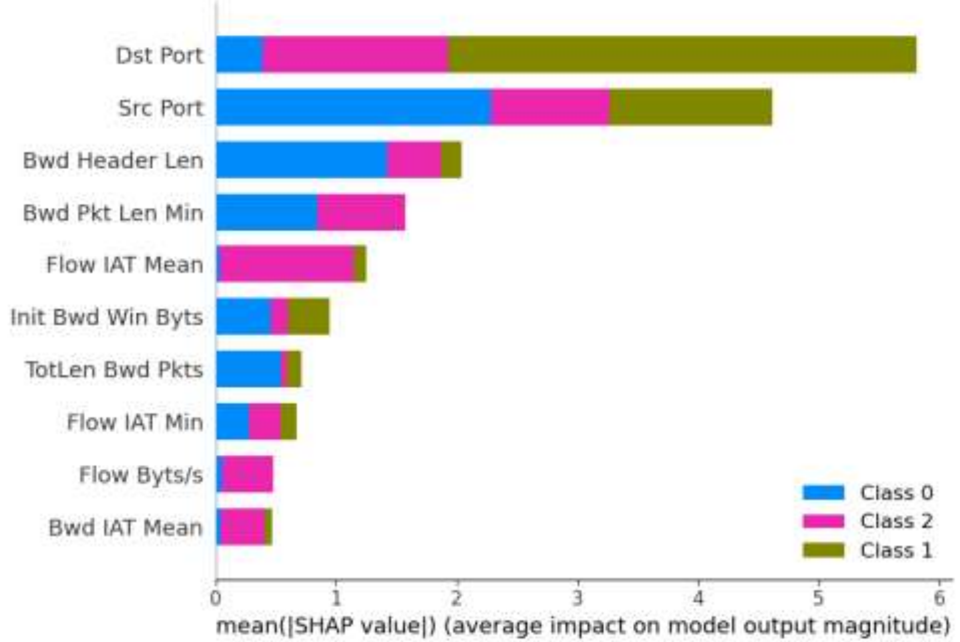


SHAP Waterfall Plot

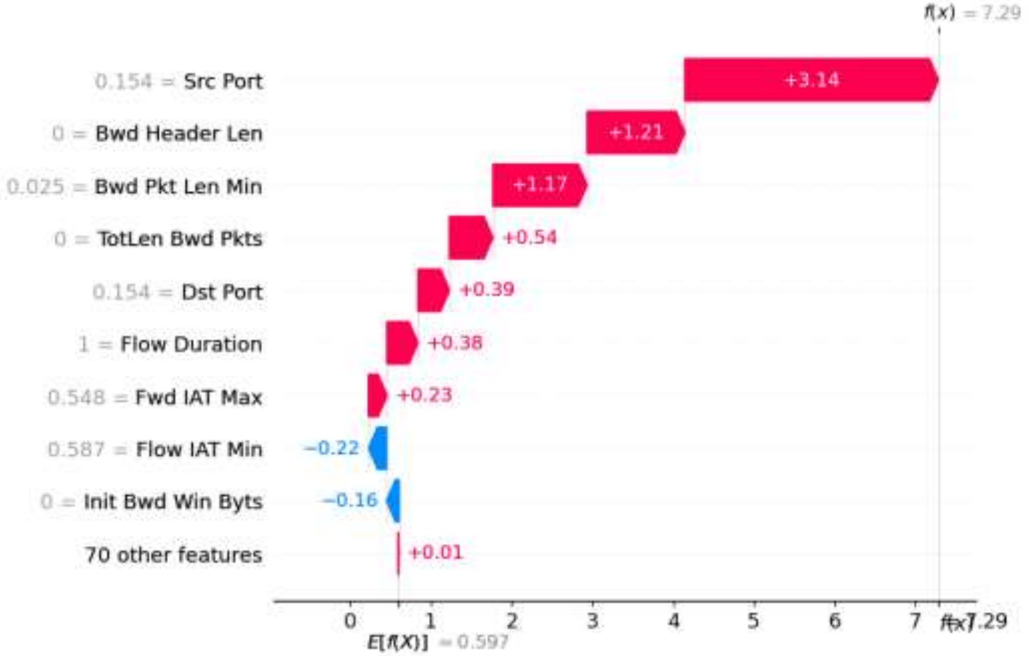


Explainability results
CIC IoT dataset 2022 - CICFlow

SHAP Summary Plot



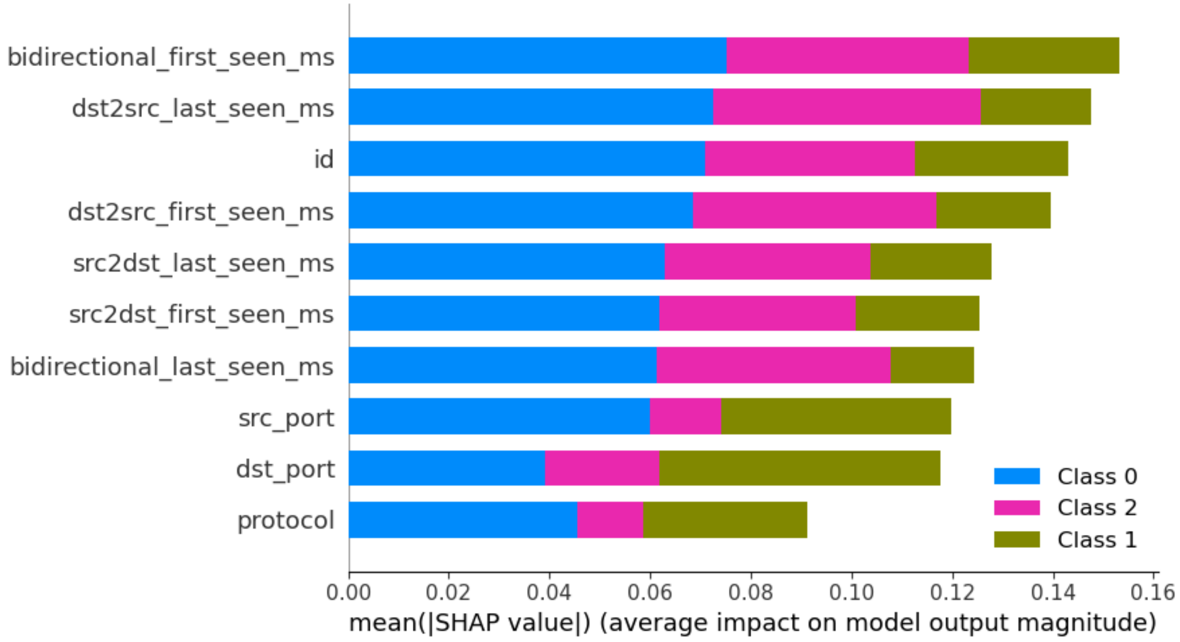
SHAP Waterfall Plot



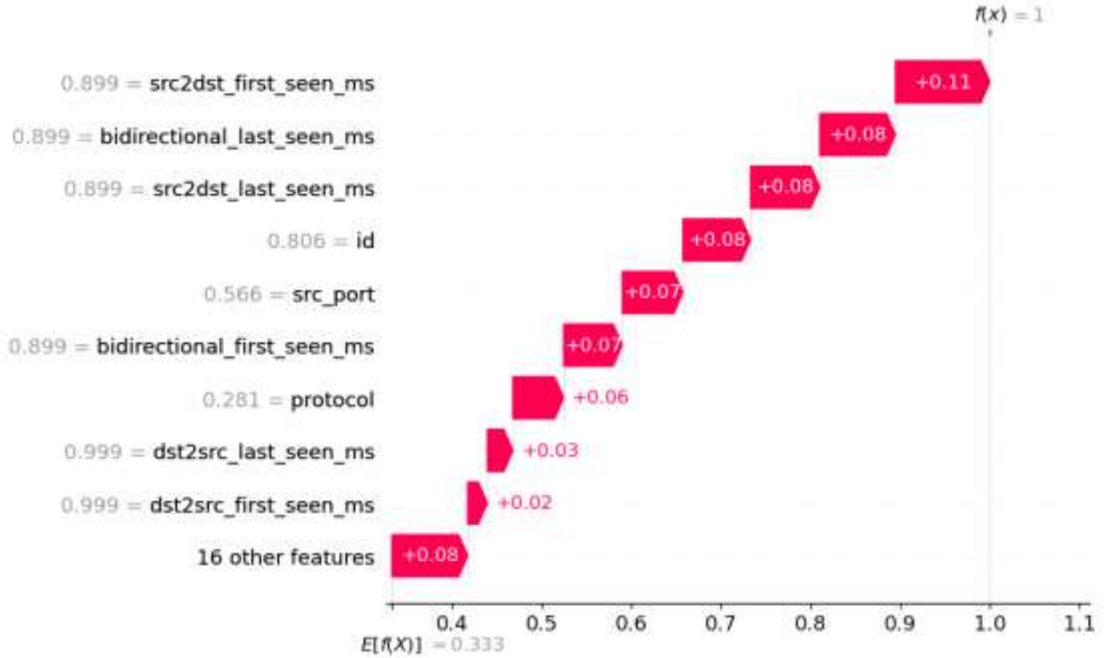
ARES conference

Explainability results
CIC IoT dataset 2022 - NFStream

SHAP Summary Plot



SHAP Waterfall Plot



ARES conference

Conclusions

- The role of IDS is crucial in detecting potential cyber-attacks and unknown anomalies.
- AI-powered IDS has shown promise in detecting threats; however, they still face challenges like false alarms and explainability issues
- Introduced an AI-powered IDS designed for IoT, including XAI functions.
- The proposed IDPS is effective in detecting malicious activities in IoT and IEC 60870-5-104 IIoT environments.
- The SHAP-based XAI functions provide feature importance for each decision, enhancing understanding and trust for security administrators and cybersecurity analysts.

ÁRES conference

Thank You!

M. Siganos et al.

K3Y LTD

msiganos@k3y.bg

K3Y R&D AND CYBER SECURITY

- **Explainable AI-based Intrusion Detection in the Internet of Things**

TRUSTEE