

# Feature Selection Improves Tree-based Classification for Wireless Intrusion Detection

Shilpa Bhandari<sup>1</sup>, Avinash K Kukreja<sup>1</sup>, Alina Lazar<sup>1</sup>,  
Alex Sim<sup>2</sup>, Kesheng Wu<sup>2</sup>

<sup>1</sup>Department of Computer Science and Information Systems, Youngstown State University

<sup>2</sup>Scientific Data Management Research Group, Lawrence Berkeley National Laboratory



# Introduction and Motivation

- 5G wireless technologies and IoT grow in size and complexity
- Robust network security systems, such as intrusions detection systems (IDS) become important
- Passive wireless traffic monitoring tools collect huge amounts of data
- Machine learning are black boxes
- SHapley Additive exPlanations (SHAP) to identify and rank important features



# 5G and IOT

- Sensors and wireless devices interconnected through the Internet are becoming extremely important
- Cyberattacks are threatening banking, online shopping, e-health and other digital services

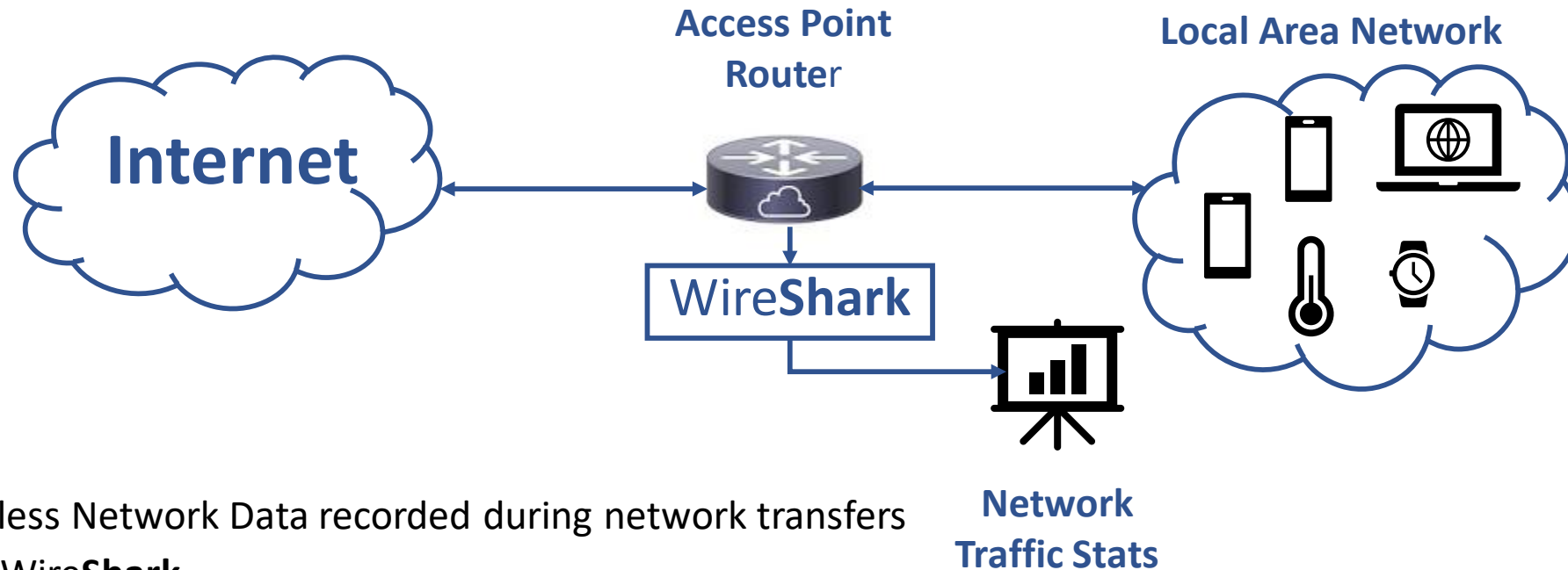


# Wireless Network Intrusion

---

- The 802.11 protocol is commonly used to implement WiFi networks.
  - Wireless local area networks connect not only computers and cell phone, but also personal devices and IoTs.
  - Security based on WEP and WPA/WPA2 protocols.
- New penetration tools make it easier to automate network attacks.
  - Wireless networks are more vulnerable compared to wired networks since they are open.
  - Depending on the security protocol many types of attacks are possible.

# AWID Simulated Datasets



- Wireless Network Data recorded during network transfers
  - WireShark
- Research Question: **Using large wireless network datasets can we identify minimal sets of features that correctly discriminate between the “normal” versus “attack” transfers?**
- Typical categories of network attacks include:
  - **Impersonation**
  - **Flooding**
  - **Injection**

# Current Drawbacks

## Problems:

- Not real-time, only checked when something is wrong
- Large datasets to analyze

## Ideal case: Automate the detection of wireless intrusions and raise alerts

- Wireless network traffic data is high volume, heavy stream of high dimensionality data
- Few training (labelled) datasets available
- Data does not follow the same distribution
- Machine learning models are black boxes

# Main Contributions

## Statistical Analysis:

- To extract throughput threshold values to label the time intervals as 'low' versus 'normal'.

## Supervised Machine Learning:

- Classification experiments performed on the AWID wireless network data using several tree-bases classification methods..

## SHapley Additive exPlanations (SHAP):

- To be used to select consistent and small feature subsets to reduce the execution time and improve classification accuracy.

## Evaluation:

- Checking the results of this supervised learning approach, especially the false positives and false negatives, using throughput plots.

# Previous Work

In 2015, Koliadis collected a set of wireless intrusion detection datasets and made them publicly available for research.

A curated subset of 36 features was used by Rezvy et al. to improve the overall prediction process, however, no details of the feature selection process were given in the paper.

Aminanto proposed a feature ranking and selection procedure, based on stack autoencoder networks and tested it in conjunction with three classification methods to improve the prediction of only the "**impersonation**" attack class.





# Deep Learning Methods

Thing was one of the first to proposed using SAEs for intrusion detection and ran the experiments on the AWID dataset. Optimal results obtained using the Parametric Rectified Linear Unit (PRelu) activation function.

Wang et al. analyzed and compared multiple architectures SAEs and CNNs, using the PRelu activation function, for wireless intrusion detection.

The ladder model proposed by Ran can be categorized as a semi-supervised deep learning approach that combines supervised and unsupervised training.



# Random Forest

A overview survey of Random Forest (RF) applications for IDSs was presented by Resende.

RF deal well with imbalanced datasets, large number of features and categorical features as well as numerical features.

An advantage of RF is that models can be trained in shorter amount of time compared to deep learning.

Another advantage is the ability to easily perform feature ranking and selection.

# Interpretable Machine Learning Approach

---

Research Question:

**Using large wireless network datasets such as AWID can we identify minimal sets of features that correctly discriminate between the “normal” versus “attack” transfers?**

Tree-based Methods

SHAP Feature Ranking

Classification on the  
Reduced Feature Set

# Tree-based Methods

XGBoost

LightGBM

CatBoost



Feature Condition



Normal



Normal



Attack

# SHAP – Dimensionality Reduction

The tree-based methods need to rank the features in order to decide which one provides the best split.

The SHAP method is based on game theory.

Shapley values work for both classification and regression and provides a consistent method to rank the features used to build the models.

The ranking can be used for feature selection, while the Shapley values can help practitioners decide the cut off point.

The SHAP dependence plots capture the impact of one feature, on the final classification task.

# AWID (The Aegean WiFi Intrusion Dataset)

AWID is a publicly available collection of datasets, containing both "normal" and "attack" real network flows.

The tabular dataset includes 154 features, plus the class.

There are 4 classes represented, that include "normal" and 3 types of attacks "injection", "impersonation" and "flooding".

The features mainly store MAC layer information collected from network traces using WireShark.

# AWID (The Aegean WiFi Intrusion Dataset)

- The data is already divided into two datasets on called training (AWID-CLS-R-Trn) and one called testing (AWID-CLS-R-Tst).
- The two datasets were collected at different times. The distribution of the 4 classes in the training and testing dataset is shown in Table 1.
- The ratio between the number of **"normal"** and **"attack"** instances is 10:1 in the training dataset and 12:1 for the testing dataset.
- This shows the class imbalance, that occurs when classes are not equally represented in the dataset.

	<b>Normal</b>	<b>Injection</b>	<b>Impers.</b>	<b>Flooding</b>
AWID-R-Trn	1,633,190	65,379	48,522	48,484
AWID-R-Tst	530,785	16,682	20,079	8,097
Total	2,371,281	82,061	68,601	56,581

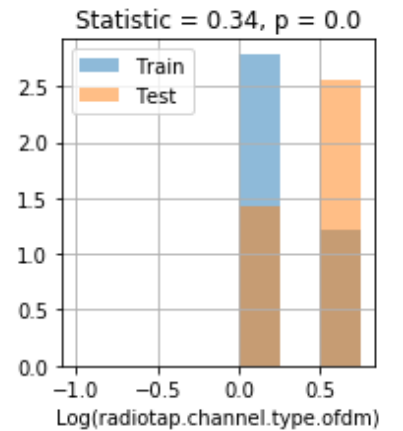
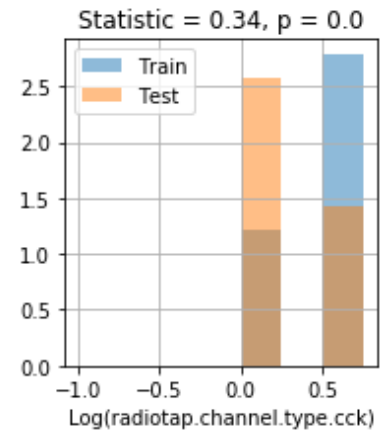
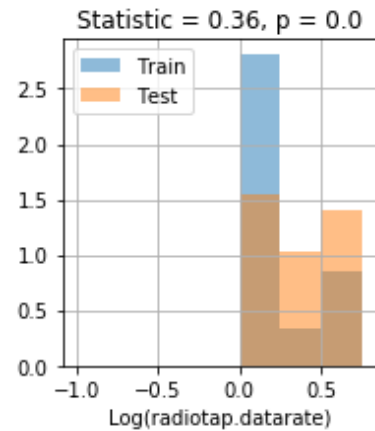
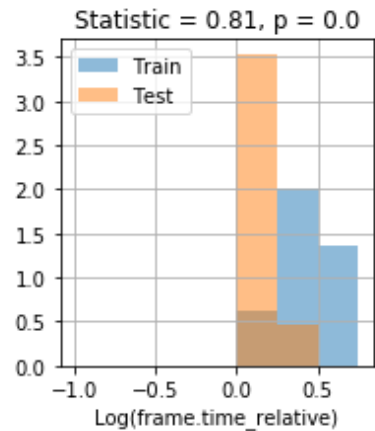
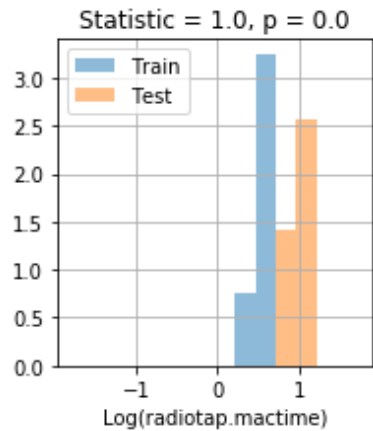
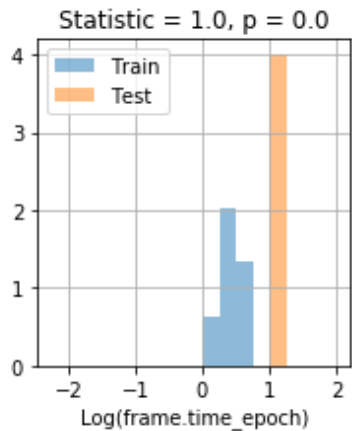
# Data Preprocessing

- The AWID datasets contain multiple features with different data types and value ranges.
- There is only one string feature, namely SSID, and all the other ones have numeric or nominal values.
- Features that represent MAC addresses are stored as hexadecimal values and need to be converted before the analysis.
- Particularly, in the training dataset there are many missing values and after converting these to zeros, many features have more than 99% zero values. After removing the mostly "zero" features, 100 features were left.
- A typical MAC address takes values in the  $[-2^{31}, 2^{31} - 1]$ , the typical value of subtypes (feature wlan.fc.subtype) is an integer between 0 and 12.





# Summary Statistics



# Experiments and Results

- All features are normalized using the MinMax Scaling procedure.
- To alleviate the large imbalanced problem of 10:1 ratio, we only select a number of "normal" instance equal with the number of all "attacks".
- Multi-class classification using Random Forest, XGBoost, CatBoost and LightGBM is performed.
- SHAP is used to rank and select the first 15 most important features. Classification and ranking is performed again on the reduced feature set with similar accuracy.
- SHAP dependence plots show the effect of each individual variable on the prediction outcome.

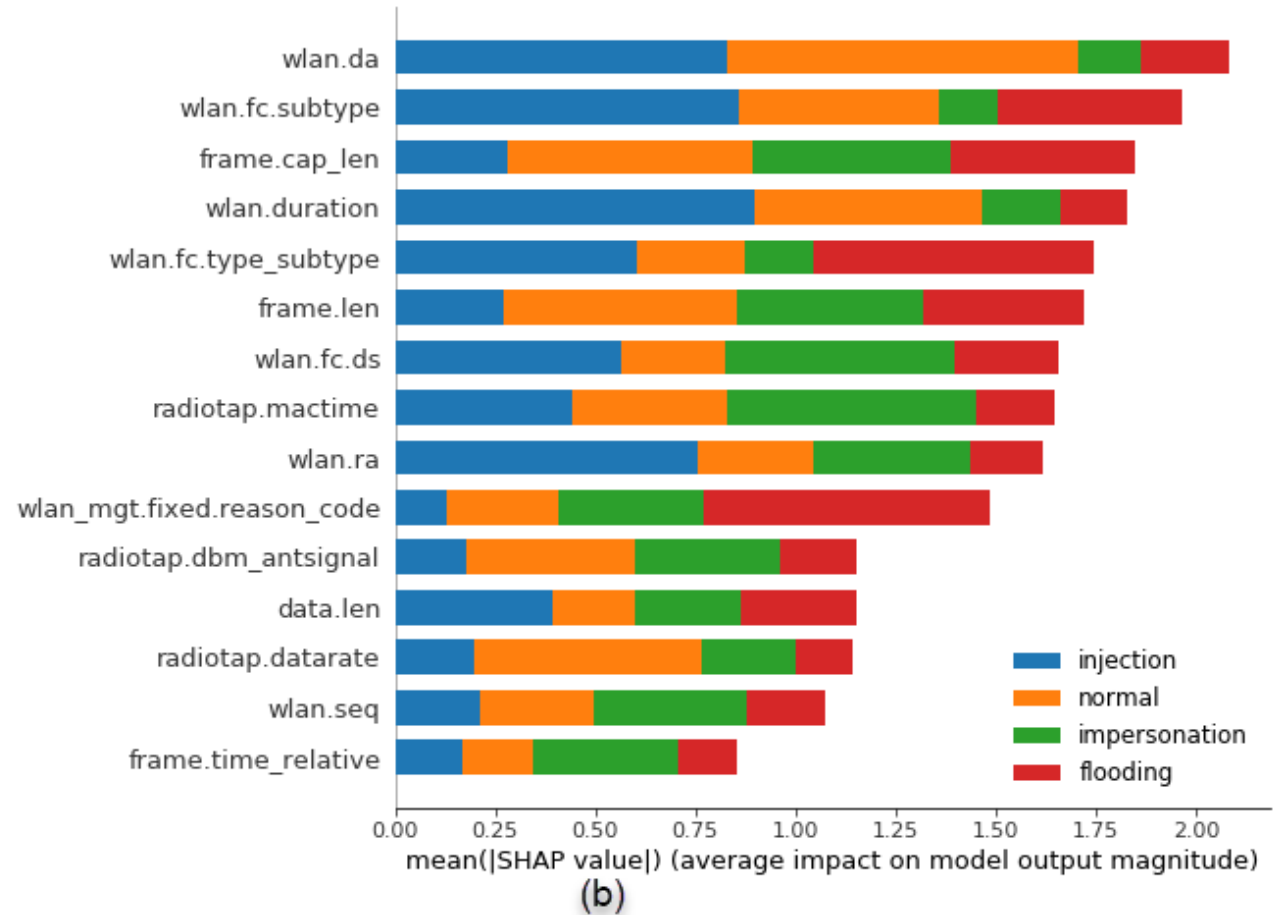
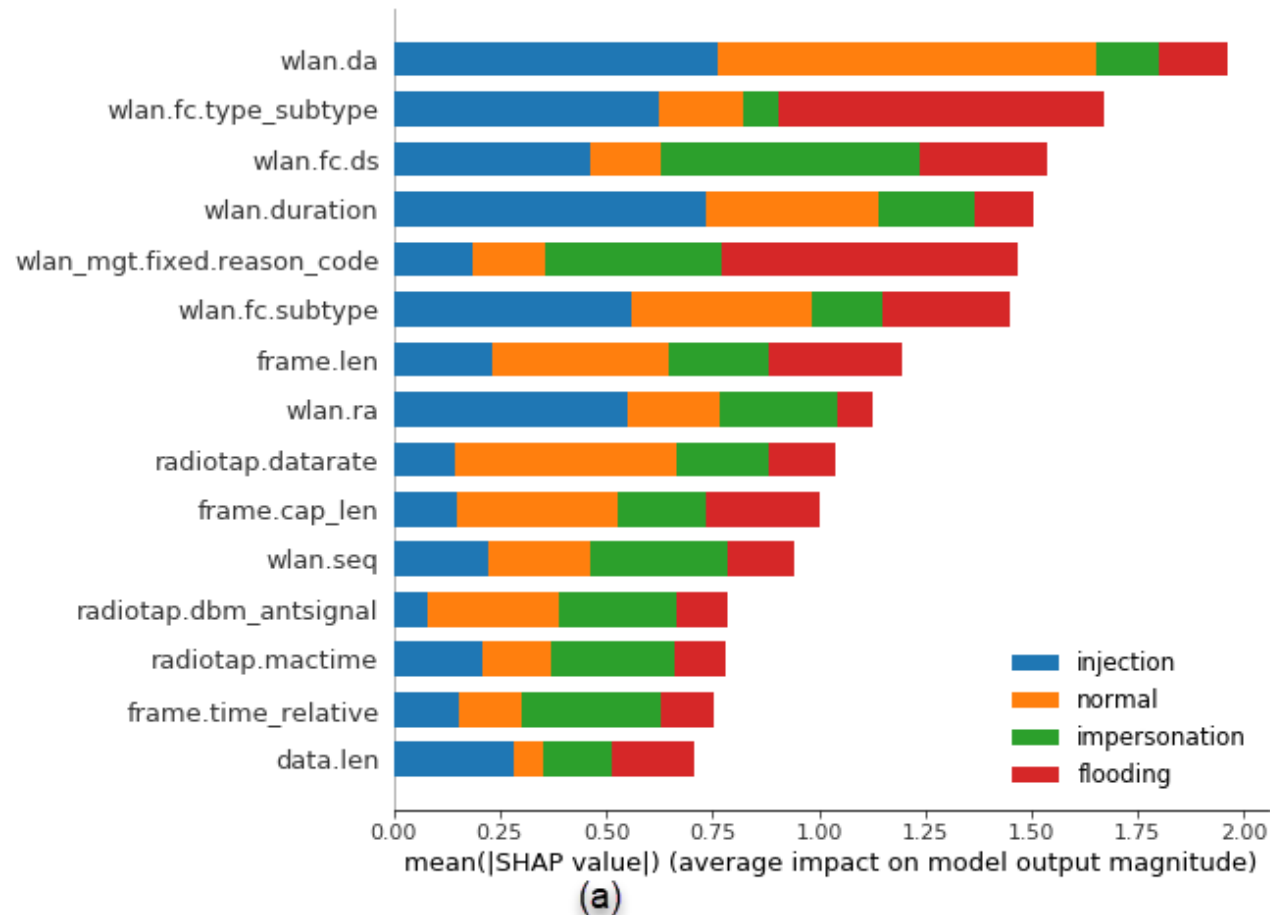
# Accuracy for the Initial and Reduced Feature Sets

---

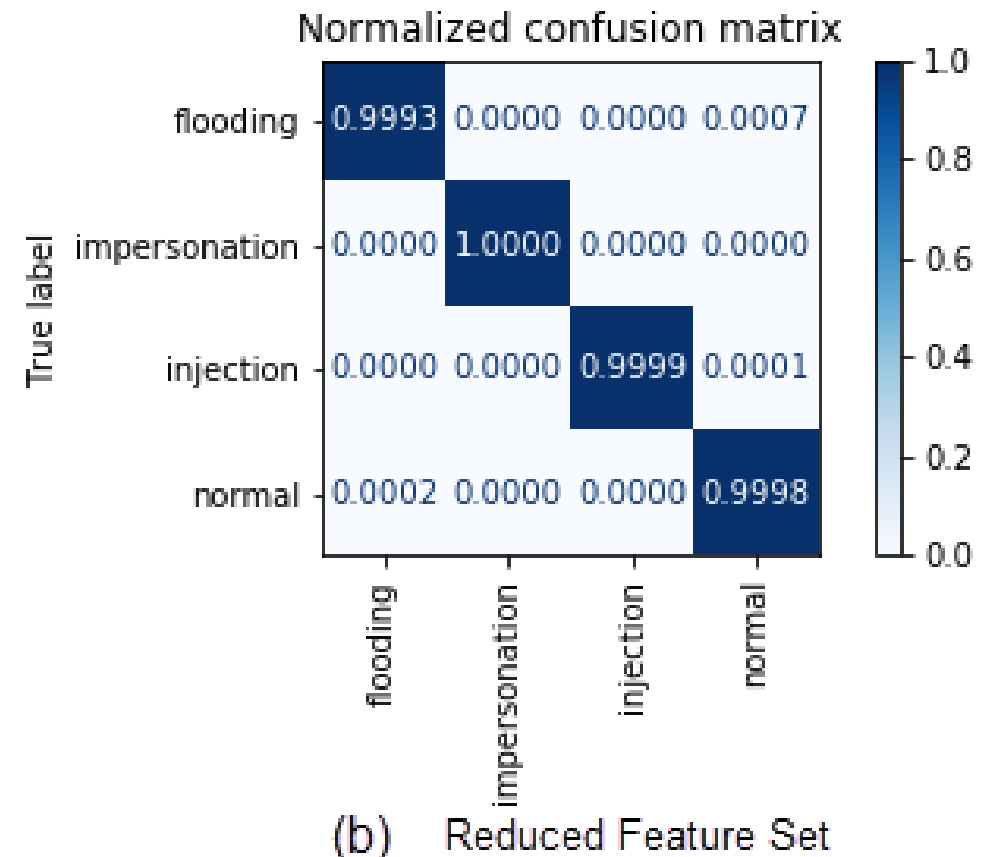
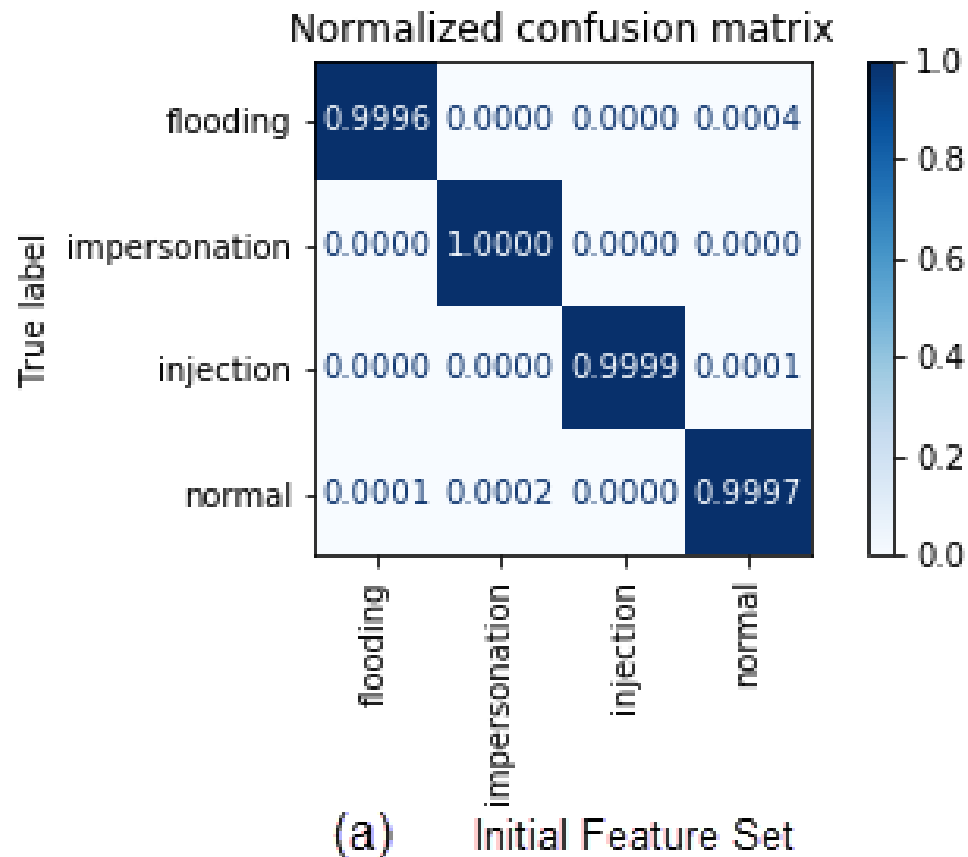
	Acc.	Prec.	Recall	F1	Time(s)
<b>GaussNB</b>	76.51	63.42	91.19	65	0
<b>RF</b>	100	99.94	100	99.97	27.1
<b>XGBoost</b>	99.8	98.25	99.88	99.05	61
<b>LGBM</b>	99.99	99.87	100	99.93	7.36
<b>CatBoost</b>	99.98	99.72	99.98	99.85	31.2

	Acc.	Prec.	Recall	F1	Time(s)
<b>GaussNB</b>	84.37	61.92	82.96	65.17	0
<b>RF</b>	99.99	99.91	100	99.95	50.5
<b>XGBoost</b>	99.85	98.63	99.94	99.28	241
<b>LGBM</b>	99.99	99.87	100	99.93	15.7
<b>CatBoost</b>	99.98	99.74	99.98	99.86	47.6

# SHAP Feature Importance Results

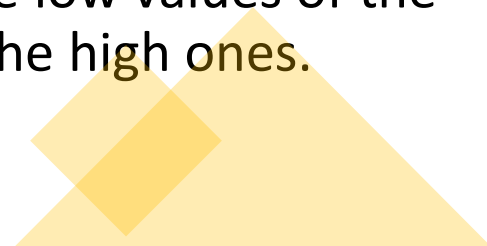


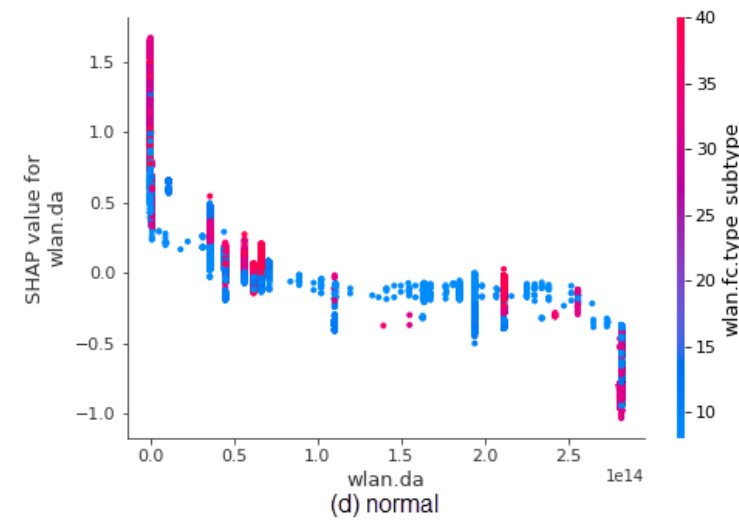
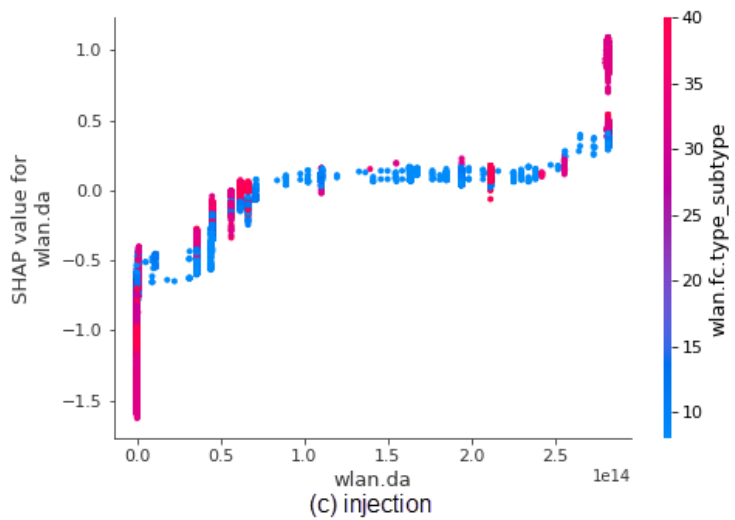
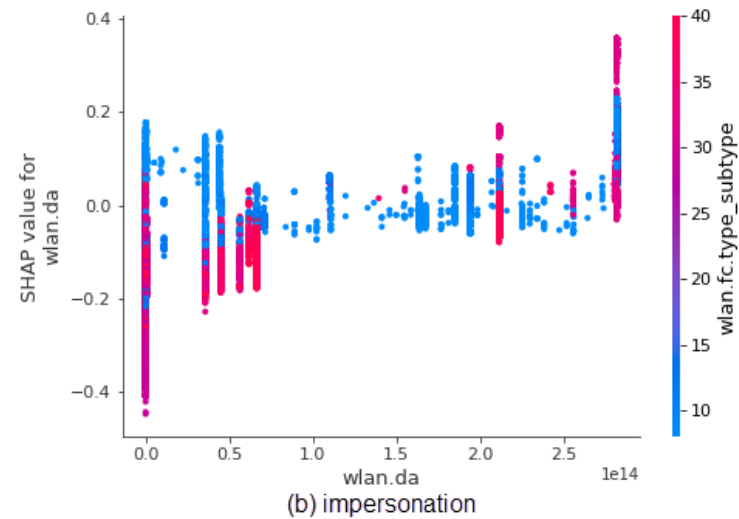
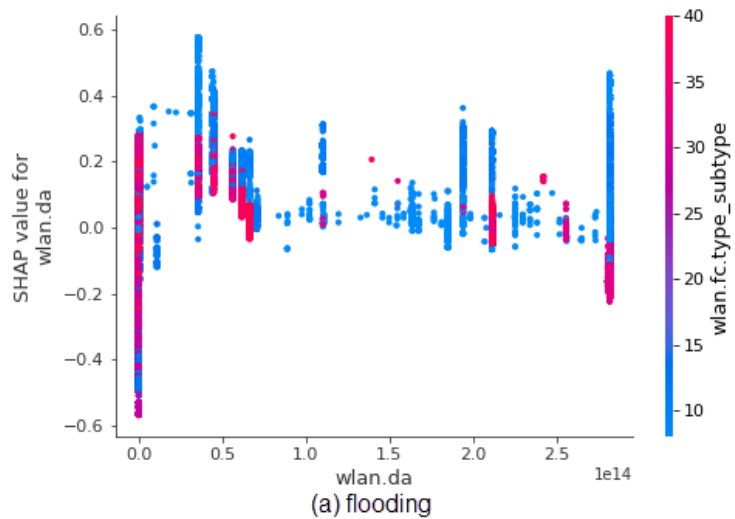
# Confusion Matrices for Initial and Reduced Feature Sets

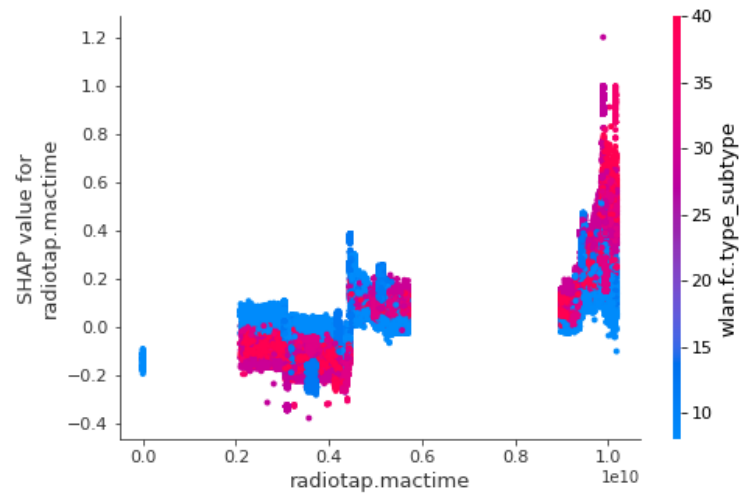




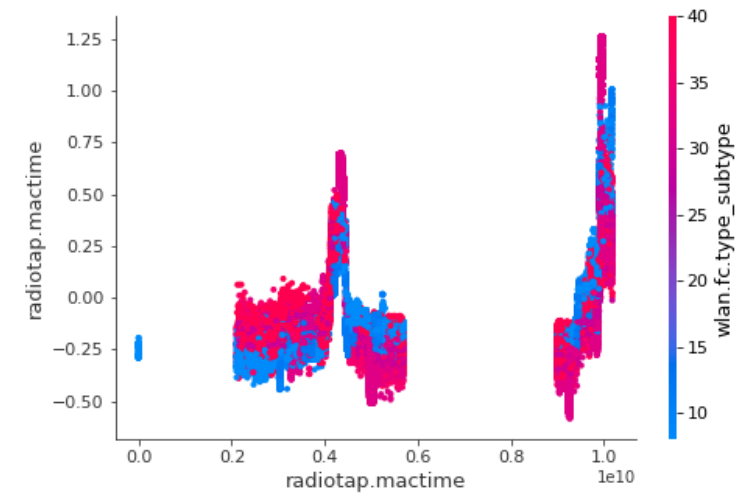
# Feature Dependence Plots

- We choose the most important feature **wlan.da** and the most important time related feature **radio.tap.mactime** and plotted the partial dependence plots for each class.
  - These figures show the marginal effect that each individual variable has on the prediction outcome.
  - The x-axis represents the value of one feature versus the SHAP values on the y-axis.
  - Each dot on the plot denotes an instance. They are colored by the values on another feature, **wlan.fc.type\_subtype** in this case.
  - Blue represents the low values of the feature while red the high ones.
- 

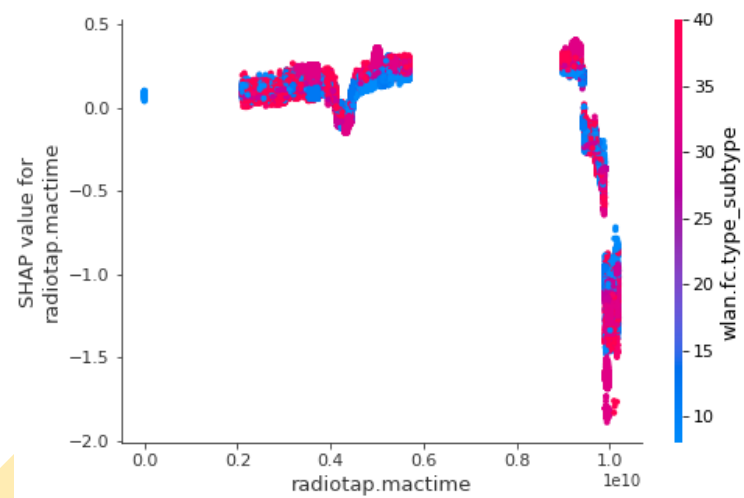




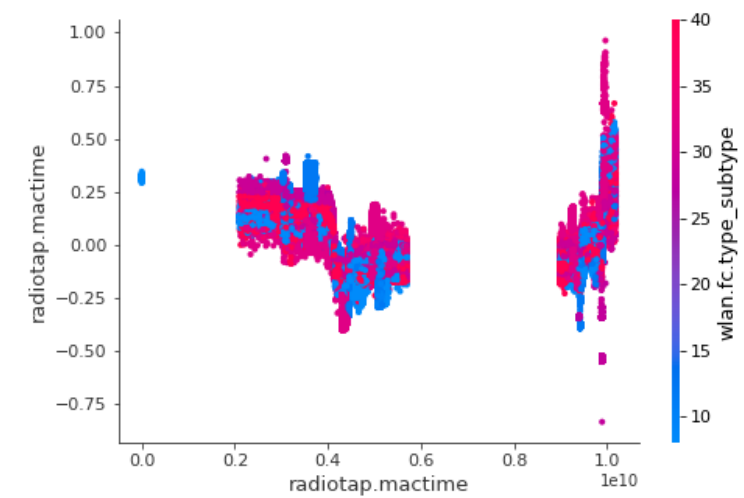
(a) flooding



(b) impersonation



(c) injection



(d) normal



# Conclusions

This paper presents a supervised data analytics system that effectively mines wireless network traffic data to discriminate between “normal” and “attack” transfers. The proposed methodology is based on a two-phase approach that

- uses tree-bases binary classification methods to classify wireless network transfers as “normal” versus “attack”.
- builds a feature ranking and selection process to remove all unnecessary features and plot the correlations between features and the final classification result.

# Conclusions

The tests on the proposed method showed its ability to accurately identify large windows of low throughput, but also showed problems in case of isolated or alternating intervals.

In future, we plan to extend the current system:

- To use domain adaptation to attenuate the distribution differences between the training and the testing datasets
- To check this method using the larger AWID training and testing datasets.
- To apply the same approach to other datasets to investigate the generalization capabilities of the presented method.



# Acknowledgements

*This work was supported by:*

- *The Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.*
- *This research used resources of the National Energy Research Scientific Computing Center.*