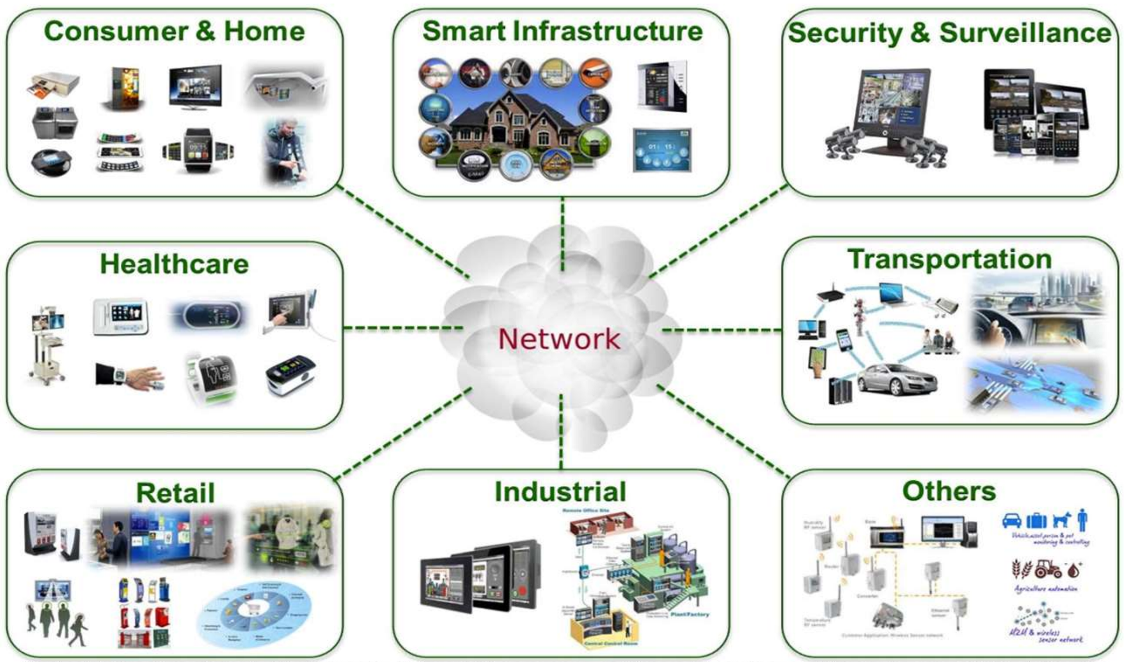
Performance Improvement of Anomaly Detection on Internet of Things Network



Latha Suryavanshi Karakos
http://www.iotcream.com/

Centre for Reverse Engineering and Assured Microelectronics

Morgan State University

Agenda

- Introduction
- Background
- Research problem
- Research goals
- Methodology
- Results
- References
- Acknowledgement
- Questions and Discussion

Cyberattacks on IoT devices

- IoT devices are prone to a variety of cyberattacks

- List of Cyberattacks on IoT devices
  1. DoS
  2. Data Sniffing/Snooping/Eavesdropping
  3. Buffer Overflow
  4. Firmware Hijack
  5. Identity and data theft
  6. Spoofing
  7. Ransomware
  8. Man-in-the Middle
  9. Password attacks
  10. Botnets ….. just to name a few

Botnet Detection

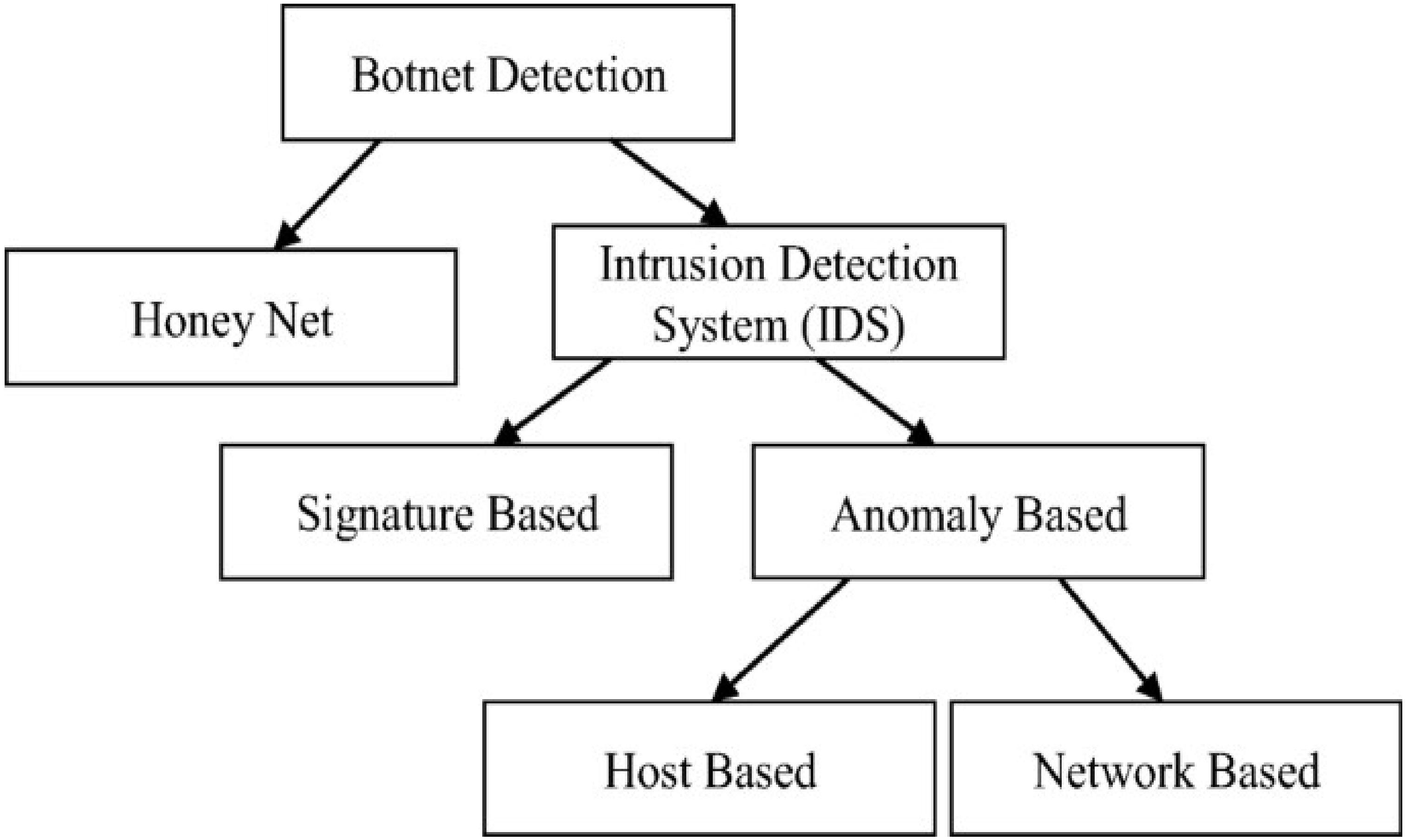


Figure 3b

Adapted from **Mimicking attack by botnet and detection at gateway**
V.Ramakrishna and R.Subhashini ,Springer
https://link.springer.com/article/10.1007/s12083-019-00854-9

IoT Device Industry Challenges

- IoT Device vendors are under time to market pressure

- Device security is not given consideration it deserves !

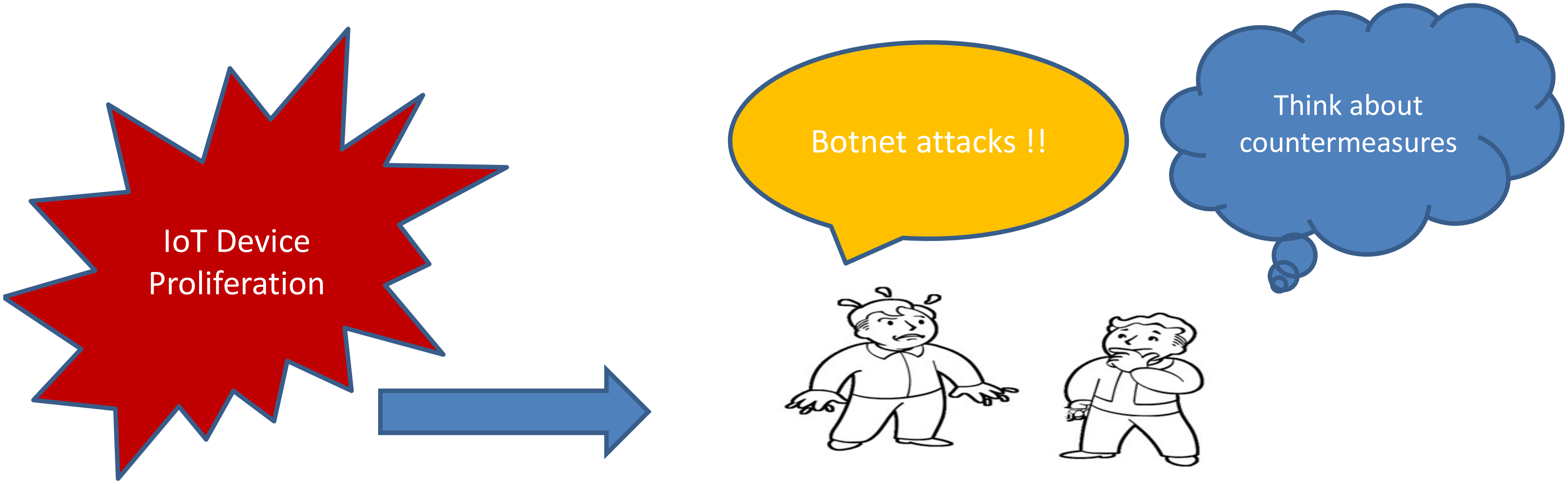
  Why???

- Huge market for cheap devices
- Cost increases due to security feature implementation
- Delays in product releases

It Means

Loss of early profit and even market share !

Research Problem



"This research work is about statistical and machine learning based countermeasures for Botnet attacks on IoT devices"

Botnet Countermeasures

- Anomaly detection has received a lot of attention
  - Several statistical and machine learning models and techniques have been studied

- Decision Tree is one such model. It offers:
  - fast prediction speed
  - fast training speed
  - small memory usage
  - suitable for deployment on small form factor devices

Novelty of this work

- Novel labeling method
- Incremental training
- Three new predictive models
  - o For detection of three attack vectors on IoTID20
    1) Mirai-Ack Flooding
    2) Mirai-HTTP Flooding
    3) Mirai-UDP Flooding attacks
- Analysis of performance characteristics as a function of data size

F-Score

$$\text{Accuracy} = \frac{\text{TP+TN}*}{\text{TP+TN+FP+FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-Score} = 2 * \frac{Precision*Recall}{Precision+Recall}$$

*TP – True Positive, TN- True Negative, FP –False Positive, FN -False Negative

Reference Model

- Built and validated decision tree model
  - IoTID20 dataset Ullah *et al.* (2020)

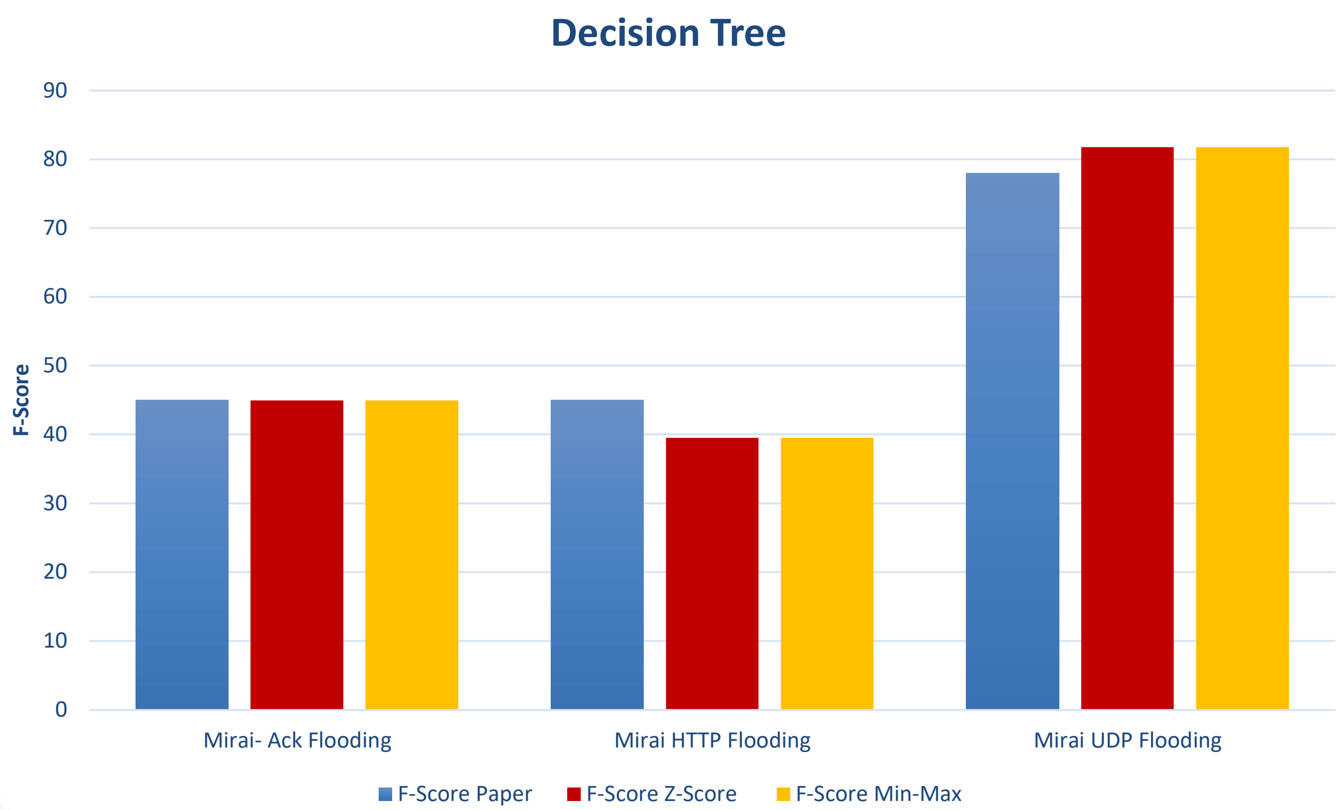


Figure 32

Novel Labeling Method and Training

- Efficient and cost-effective
  - Useful when manual labeling effort is limited
- Incremental Training
- Focus on attack vectors of interest

  Mirai Ack Flooding
  Mirai HTTP Flooding
  Mirai UDP Flooding
- Others

  Normal
  DoS Synflooding
  MITM ARP Spoofing
  Scan Hostport
  Scan Port
  Mirai Hostbruteforce

Rationale behind this study

- IoTID20 dataset - 625k observations split by Ullah into
  - Training set - 70% - 438k
  - Test set - 30% - 175k

- In a real-world scenario:
  - IoT Edge device companies are small to mid-sized
  - Working with such huge datasets is not economically viable

- Why?
  - Human labor to label the datasets is expensive
  - 438k is too high a number

Rationale behind this study

Example:

Assume that the time taken by a person to annotate 1 data point = 1 minute

> 438000 observations/60 minutes = 7,300 hours
> 7,300 hours = 912, 8-hour working days
> 912, 8-hour working days = 2.5 years to finish data annotation
> Cost: 0.25-0.5 million dollars at current data labeling rates

Proposed Solution:

- Use *smaller labeled* training data sets
- Improve performance using unlabeled data with self-learning, specialized learning
- Use of Incremental training

Rationale behind this study

How big should the dataset size be?

"A size that small–mid sized companies can afford"

Example:

a) **one day (very small)**

8 hours of data labeling yields 480 labeled data points

b) **one week (small)**

5 days of data labeling is 2400 labeled data points

c) **one month (medium)**

4 weeks of data labeling is 9600 labeled data points

Experiments

- A set of 16 experiments was conducted
- 4 different sizes of dataset and 4 different predictive models
- Features varied from 10 to 70
- Tree depth varied from 2 to 20
- Dataset incremented from 10% to 90% in steps of 10%
- Sizes of dataset
    - Very Small – 0.11% of total labeled training data of Ullah (2020)
    - Small – 0.6%
    - Medium – 2.3%
    - Large – 90%-100 %
- Performance characteristics
    - Accuracy / F-Score

Self-Labeling Classifier

- Novelty: It’s the first time it is applied to IoTID20
- It is also referred to as self-training or decision-directed learning, hybrid learning method
- Computing resources are leveraged to automatically label a large amount of unlabeled data in lieu of human labor
- Reduces labeling cost significantly

Specialized Classifier

- Specialized classifier is a special case of the reference classifier

- Attacks of interest:
  Mirai-Ack Flooding, HTTP Flooding and UDP flooding

- The training dataset treats sub-category of attacks that are not of interest as 'OTHERS'

- Faster labeling => Cost Reduction
  - Manual labor not used for annotation

Combined Methods

- This is a combination of Self-Labeling and Specialized Classifiers
- Attack vectors not of interest fall into the 'OTHERS' sub-category
- Further reduces labeling cost significantly
  - One reduction comes from using small labeled data and large self-labeled data
  - Another reduction comes from not labeling the "other" sub-categories

Results – Large dataset

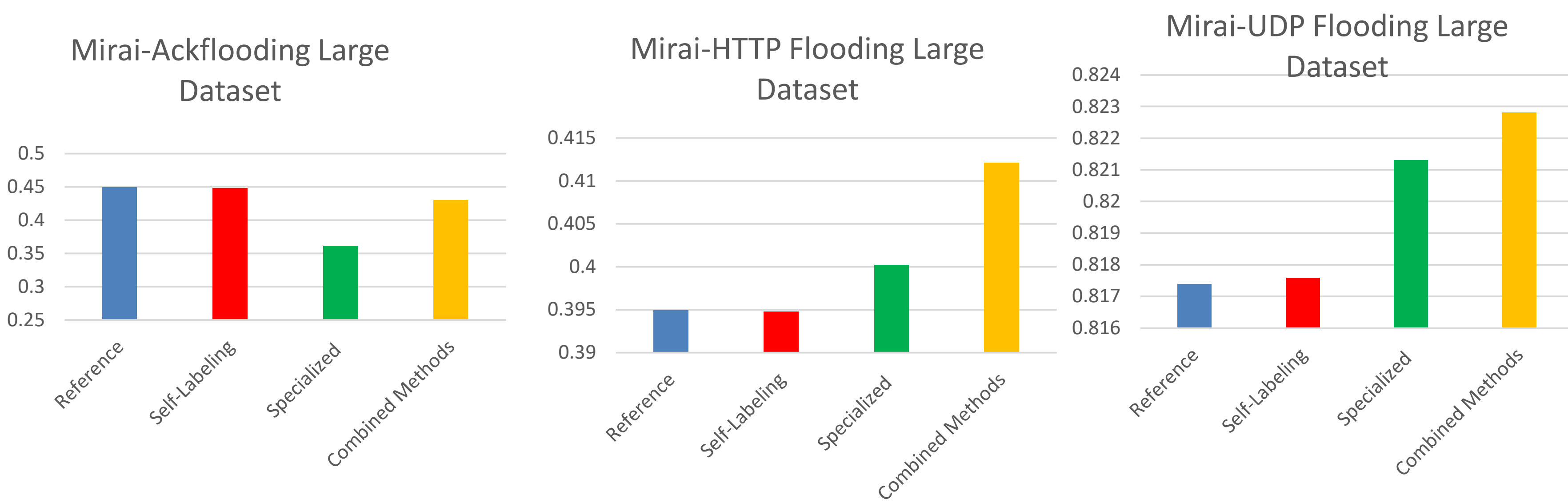


Figure 33

Results – Medium dataset

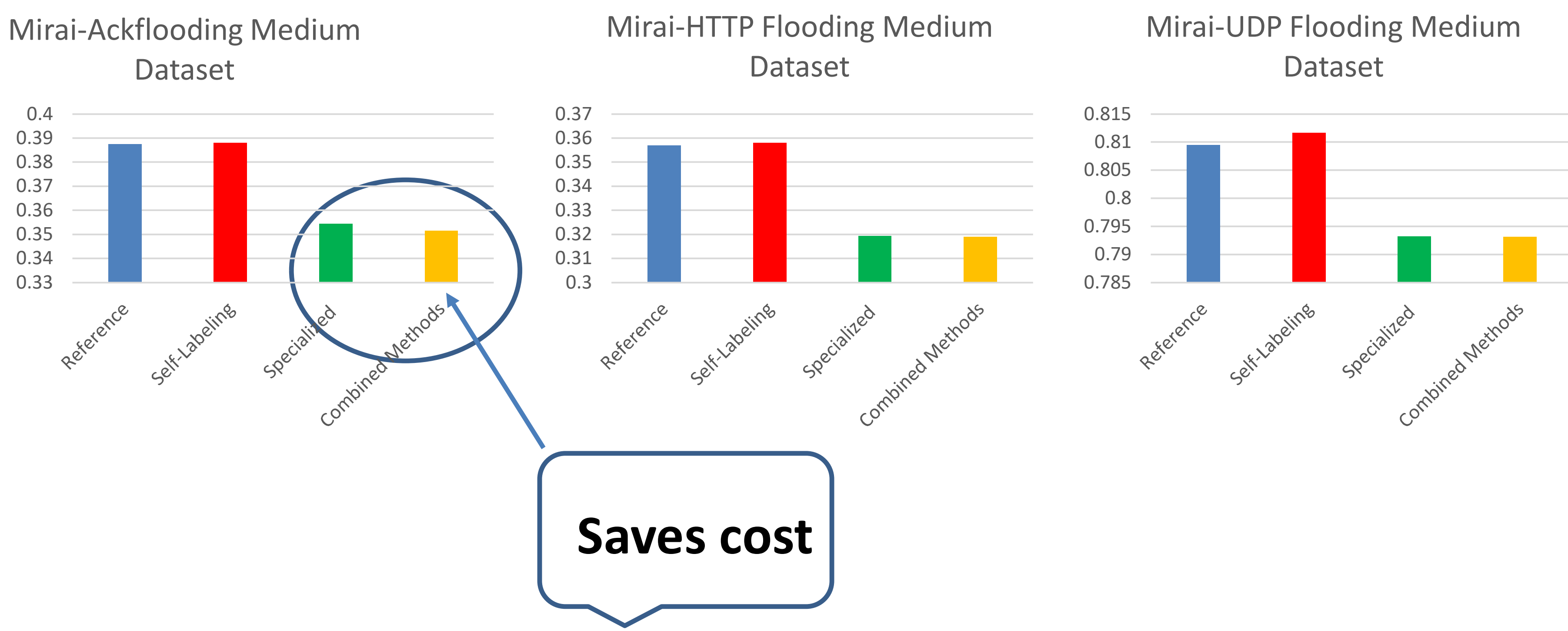


Figure 34

Results – Small dataset

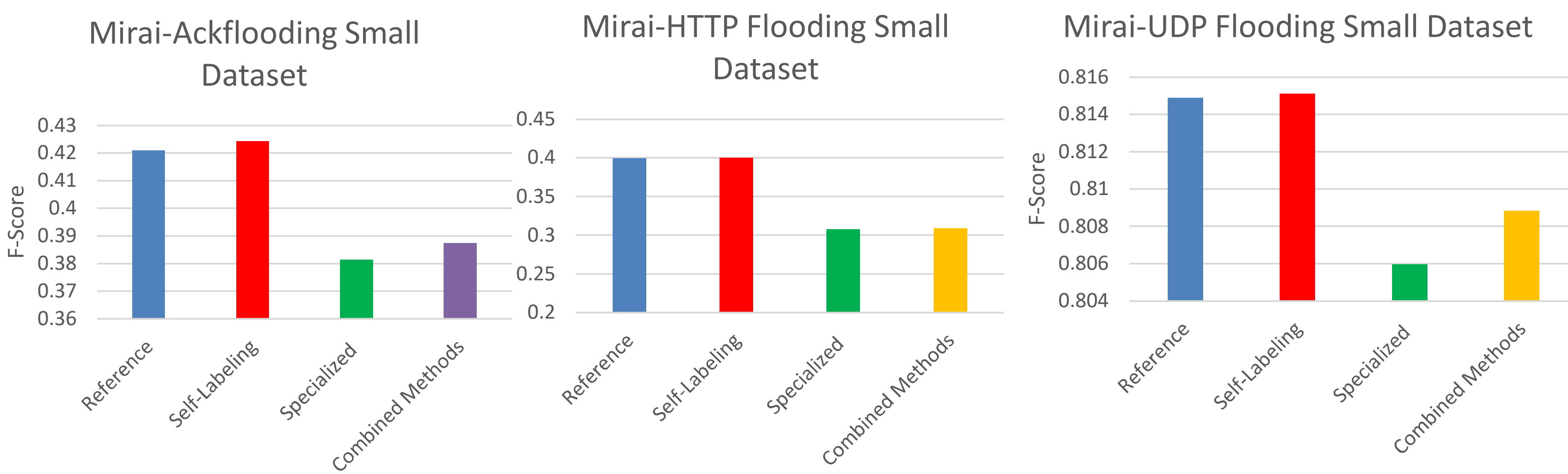


Figure 35

Results – Very Small dataset

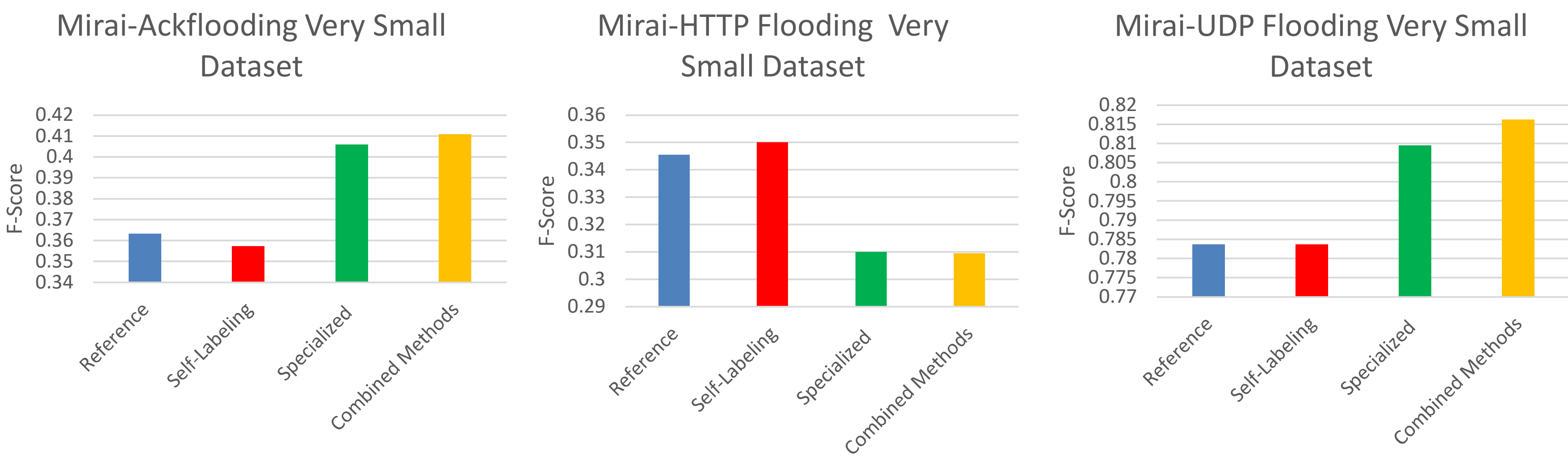


Figure 36

Observations and Analysis

Large dataset:

- Up to 4.3% improvement in F-Score in detecting Mirai-HTTP Flooding compared to Reference model
- For 2 out of 3 labels, Combined Methods performs the best

Medium dataset:

- Self-Labeling classifier performs nearly the same or better for all three attack vectors
- Self-Labeling classifier gives 1.1% improvement in F-Score for Mirai UDP Flooding

Observations and Analysis (cont.)

Small dataset:

- Self-Labeling classifier performs nearly the same or better for all three attack vectors

Very small dataset:

- Specialized classifier and combined methods perform best in detecting Mirai-Ack Flooding (13% gain over Reference) and Mirai-UDP Flooding (4% gain)

- Self-learning classifier performed better for Mirai-HTTP Flooding

Conclusions

It can be concluded that

- Small to very small datasets perform as well as medium to large datasets in terms of F-Score while detecting the three attack vectors

- Smaller dataset sizes save costs

- Self-Labeling predictive models are faster to label and train, and perform well with all attack vectors (for very small training dataset)
    - They prove to be the most cost effective

- Although the F-Score of the specialized classifier does not match the reference model, it offers important benefits:
    - Less intensive for human labeler
    - Time and cost savings

Acknowledgement

Advisers

Dr. Jumoke Ladeji-Osias

Dr. Kevin Kornegay

Dr. Kofi Nyarko

Funding

- Special Thanks to Northrop Grumman for supporting this work

Questions and Discussion