Novel Attack Detection in IoT Network Intrusion Detection

Chiao-Hsi Joshua Wang (Student No. 46965611)

Supervised by Dr. Dan Kim

TABLE OF Contents

01 Project Purpose

02 Project Goals

03 Project Background

04 Methodology

05 Results

06 Conclusions

DID YOU KNOW?

There was a **400%** increase in malicious cyberattacks on Internet of Things (IoT) devices from 2022 to 2023

(Knowles, 2023)

Project Purpose

- Successful cyber-attacks can result in data breaches, privacy violations and service disruptions
- Network intrusion detection systems (NIDS) help to protect IoT ecosystems by identifying potential security threats
- This project addresses the limitations of current methods, specifically their inabilities to adapt to changing IoT environments and identifying specific attack types whilst recognising new attacks

Project Goals

1. Classify IoT network traffic as **either attack or benign traffic** (binary classification)

2. Identify the **type of traffic** IoT network traffic belongs to by classifying as **either benign traffic or the attack type** (multi-class classification)

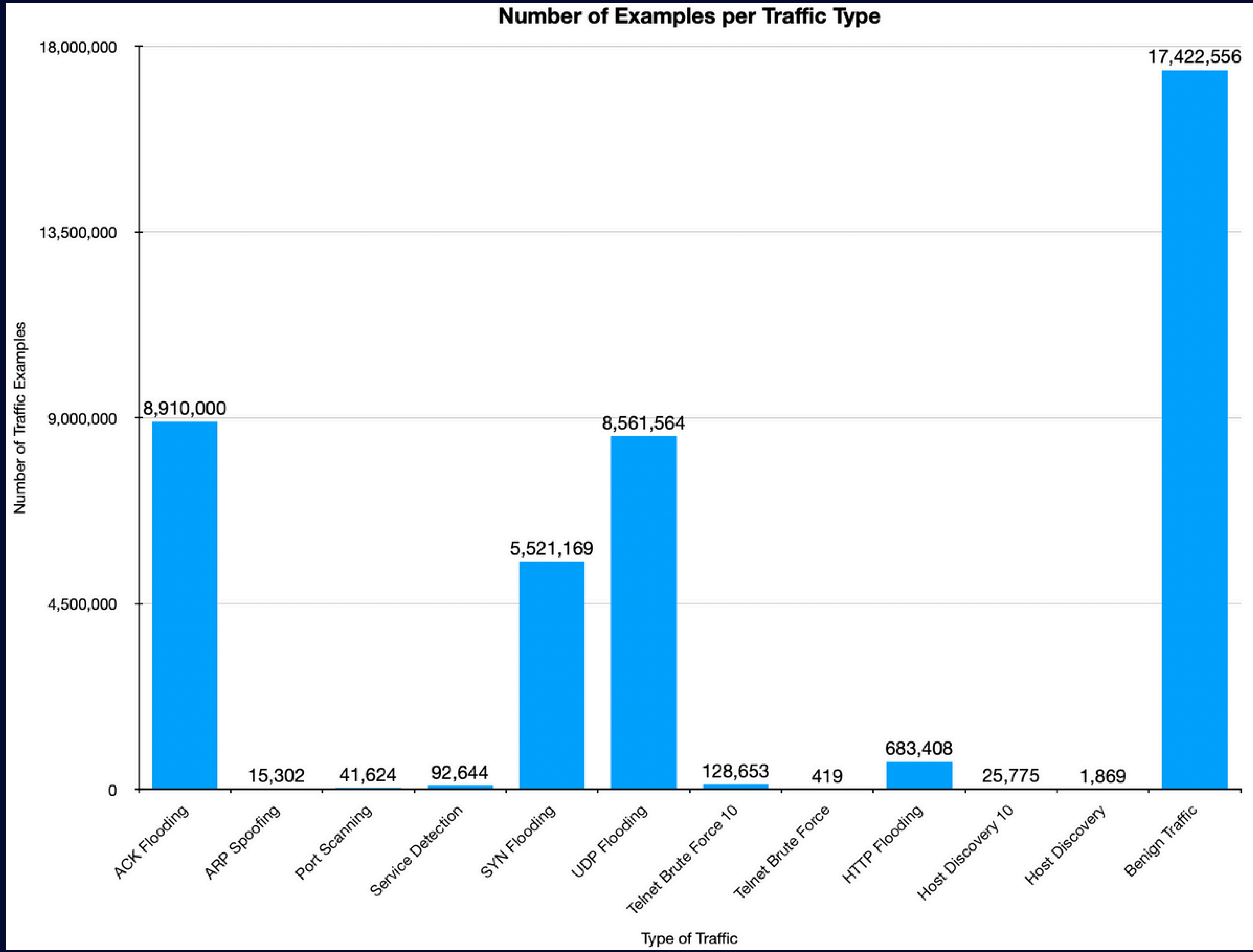
3. **Identify and classify novel attack types** accurately upon initial exposure

Project Background

- Most supervised-learning IoT network intrusion detection systems struggle to adapt to new types of attacks (Al Lail et al., 2023; Dong et al., 2016; Kim et al., 2017)
- State-of-the-art models such as the Kitsune model are trained in a semi-supervised manner and can be applied to all attacks, but only to decide whether the traffic is benign or attack (Mirsky et al., 2018)
- This project will create a model capable of identifying new attacks whilst also being able to identify the type of attack found (whether it is a known attack or new type of attack)

Methodology: Data Preprocessing

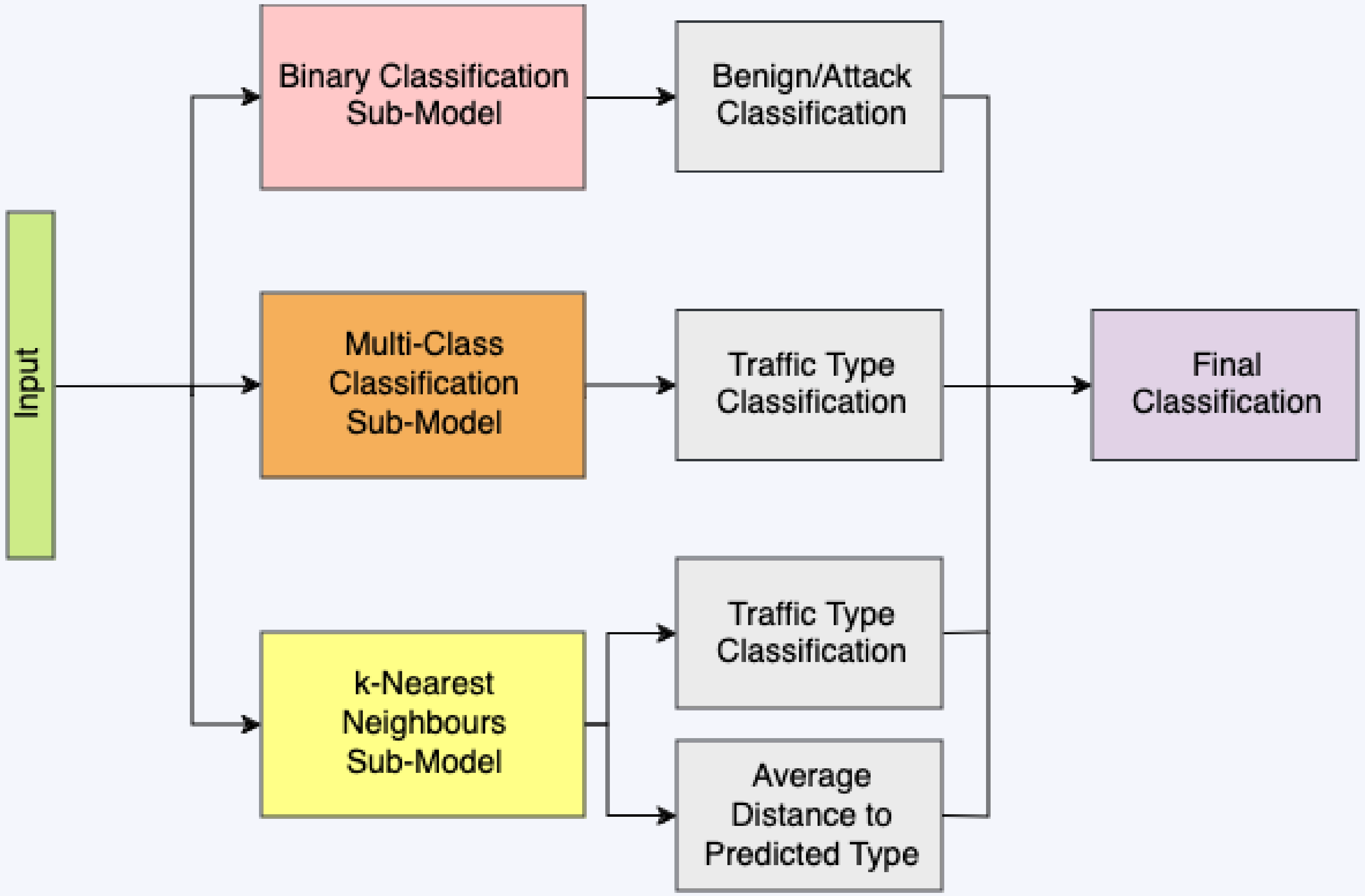
1. Feature Extraction (Kitsune)
2. Sampling
   a. 70-15-15 train/val/test for over sampling minority classes
   b. 5-5-5 train/val/test for under sampling majority classes
3. Leave one attack out of training set to simulate "unknown" attack
4. Min/max scaling
5. Principal Component Analysis



Number of samples available in the UQ IoT IDS Dataset for each type of traffic.

Methodology: Model

- Processed inputs used to train three sub-models
- Final outcome determined through voting from the three sub-models
- Voting mechanism considers labels generated by each of the three sub-models, as well as distance thresholds for k-Nearest Neighbours model



Model Architecture

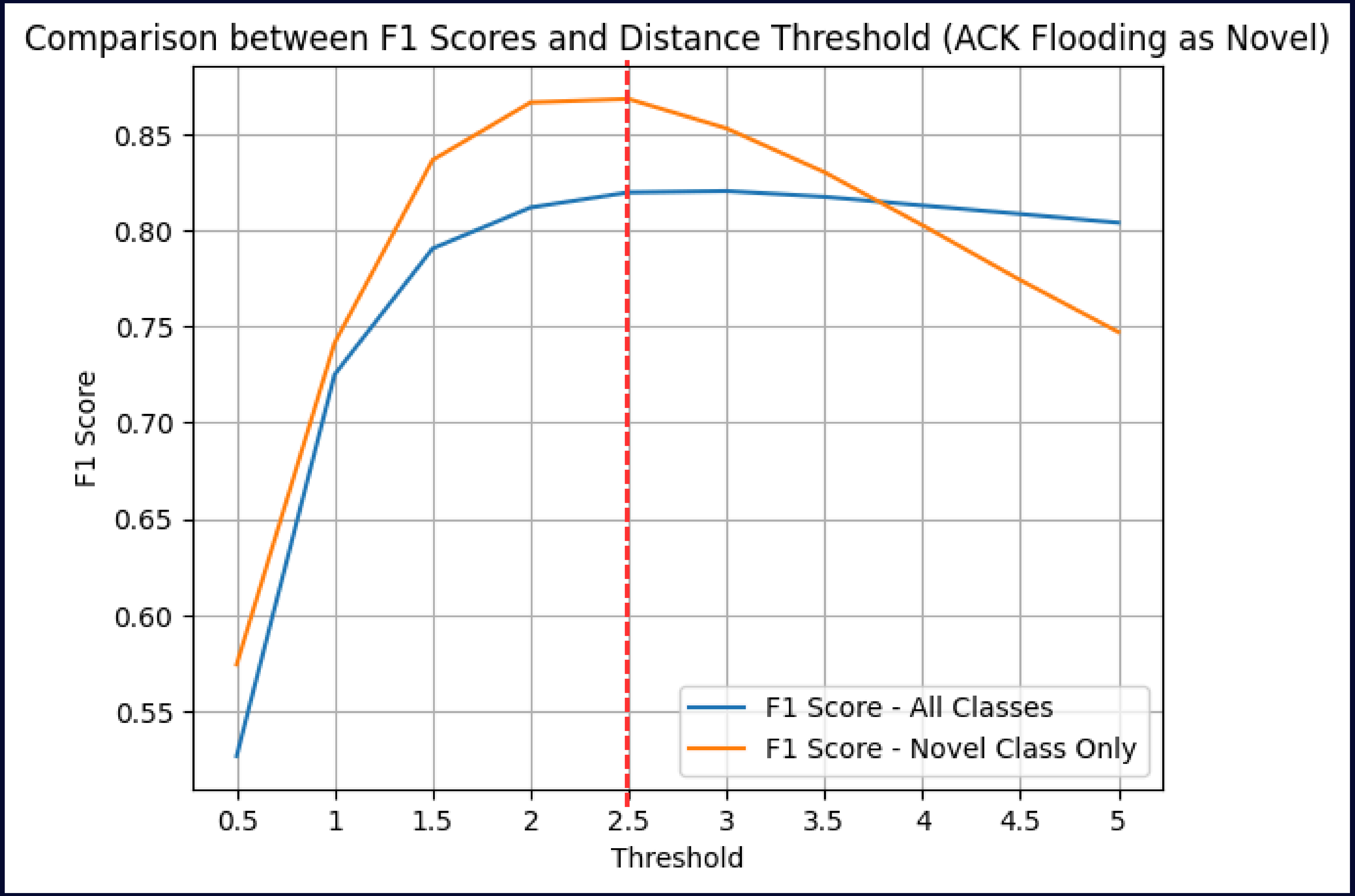
Chiao-Hsi Joshua Wang

Methodology: Voting Algorithm

```
if ((multi_label == 0 and binary_label == 0 and knn_label == 0)
   or (multi_label == 0 and binary_label == 0 and knn_label > 0)
   or (multi_label > 0 and binary_label == 0 and knn_label == 0)
):
   final_labels.append(0) # Predict benign
else:
   # Determine what attack class to predict
   if data_mean_dist > class_avg_distances.get(knn_label) * threshold:
      final_labels.append(-1)  # Predict novel
   else:
      final_labels.append(knn_label)  # Predict known-class attack
```

- data_mean_dist: mean distance between point being evaluated and its closest 5 neighbours
- class_avg_distances: dictionary of average distances between training data points of the same class
- threshold: Decision threshold value

Methodology: Distance Threshold Hyperparameter



- Optimal threshold value found through evaluation on validation set
- 2.5 is optimal for all combinations of "unknown" attack

Results - Easily Distinguishable Classes

- Performs well on most classes when classifying "known" attacks and identifying "unknown" attacks
- Average accuracy of 0.92, average F1 score of 0.8 for ACK Flooding as "unknown"

Packet Type	Precision	Recall	F1
Benign	0.99	0.98	0.99
ARP Spoofing	0.49	0.53	0.51
Port Scanning	0.88	0.82	0.85
Service Detection	0.91	0.78	0.84
SYN Flooding	0.81	0.94	0.87
UDP Flooding	1.00	0.95	0.98
HTTP Flooding	1.00	0.94	0.97
Telnet Brute Force	0.83	0.77	0.80
Host Discovery	0.71	0.42	0.53

Performance on Benign and Known Attacks
(ACK Flooding as Unknown)

Packet Type	Precision	Recall	F1
ACK Flooding	0.88	0.85	0.87

Performance on Unknown Attack
(ACK Flooding as Unknown)

Results - Similar Classes

- Model struggles on attacks that follow similar patterns (e.g. Port Scanning and Service Detection)
- Average accuracy of 0.89, macro F1 score of 0.71 for Port Scanning as "unknown"

Packet Type	Precision	Recall	F1
Benign	0.99	0.98	0.99
ACK Flooding	0.98	0.93	0.96
ARP Spoofing	0.52	0.45	0.48
Service Detection	0.70	0.83	0.76
SYN Flooding	0.98	0.92	0.95
UDP Flooding	1.00	0.96	0.98
HTTP Flooding	1.00	0.94	0.97
Telnet Brute Force	0.83	0.77	0.80
Host Discovery	0.71	0.44	0.55

Performance on Benign and Known Attacks
(Port Scanning as Unknown)

Packet Type	Precision	Recall	F1
Port Scanning	0.01	0.12	0.02

Performance on Unknown Attack
(Port Scanning as Unknown)

Overall Results and Comparisons

Comparison of Models for Multi-Class Classification

Model	Accuracy	Macro F1
Proposed Model	95.01%	79.65%
Random Forest	98.86%	88.30%
Convolutional Neural Network	89.25%	64.87%

Comparison of Models for Attack Detection

Model	Accuracy	Macro F1
Proposed Model	98.83%	98.79%
Kitsune	94.99%	92.21%

Conclusions and Future Work

- Developed a model capable of strong results compared to existing solutions
- Future work:
  - Improving performance on classes which tend to be similar to each other - feature engineering
  - Testing on different datasets
  - Giving model ability to learn from detected novel classes

References

Al Lail, M., Garcia, A., & Olivo, S. (2023). Machine learning for network intrusion detection—a comparative study. Future Internet, 15(7), 243. https://doi.org/10.3390/fi15070243

Dong, B., & Wang, X. (2016). Comparison deep learning method to traditional methods using for network intrusion detection. 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN). https://doi.org/10.1109/iccsn.2016.7586590

He, K., Kim, D., Zhang, Z., Ge, M., Lam, U. & Yu, J. (2022). UQ IoT IDS dataset 2021. The University of Queensland. (Dataset) doi: 10.48610/17b44bb http://dx.doi.org/10.48610/17b44bb

Kim, J., Shin, N., Jo, S. Y., & Kim, S. H. (2017). Method of intrusion detection using Deep Neural Network. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). https://doi.org/10.1109/bigcomp.2017.7881684

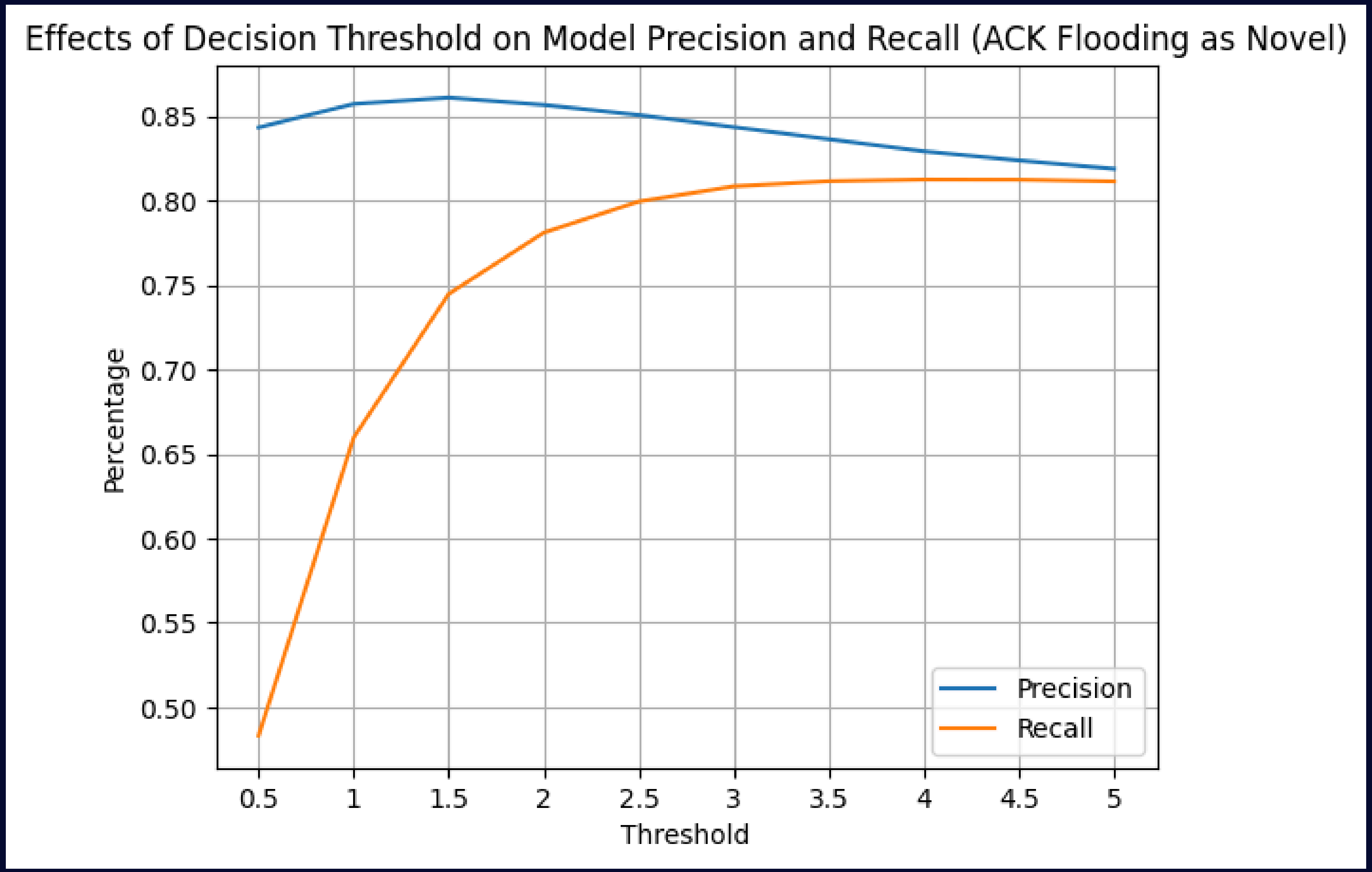
Knowles, C. (2023, October 26). Manufacturing sector hit hardest by 400% rise in IOT malware attacks. SecurityBrief Australia. https://securitybrief.com.au/story/manufacturing-sector-hit-hardest-by-400-rise-in-iot-malware-attacks

Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. Proceedings 2018 Network and Distributed System Security Symposium. https://doi.org/10.14722/ndss.2018.23204

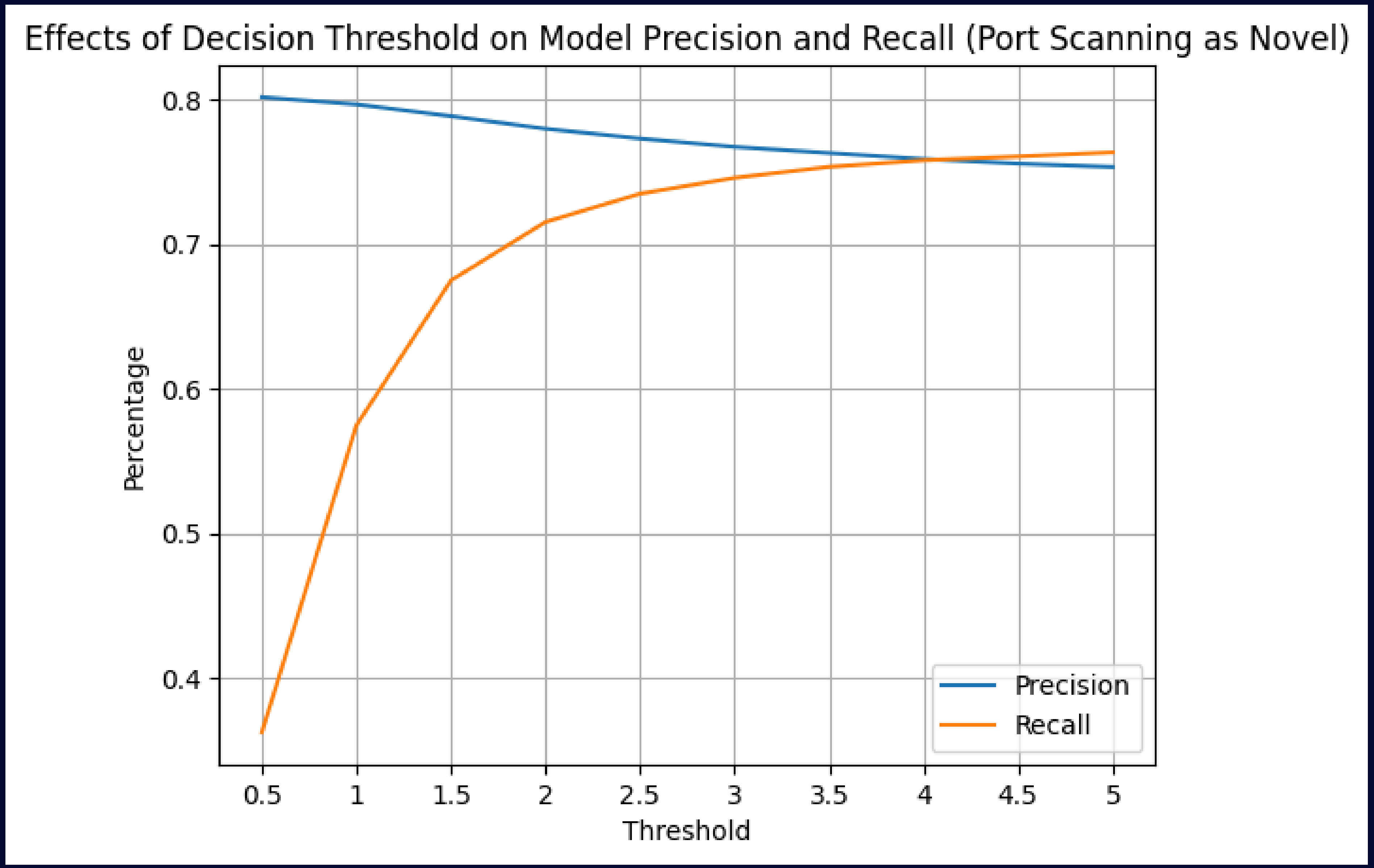
Appendix A: Model Architecture/Parameters

- Multi-class and Binary Classification Models:
  - Input: Tensor of size 12
  - Layers: [512, 128, 64, 16]
  - Outputs: 9 (multi-class); 1 (binary)
  - Criterion: CrossEntropyLoss (multi-class); BCEWithLogitsLoss (binary)
  - Optimizer: Adam, default learning rate of 0.001
- k-Nearest Neighbours:
  - k = 5

Appendix B: Distance Threshold Precision and Recall



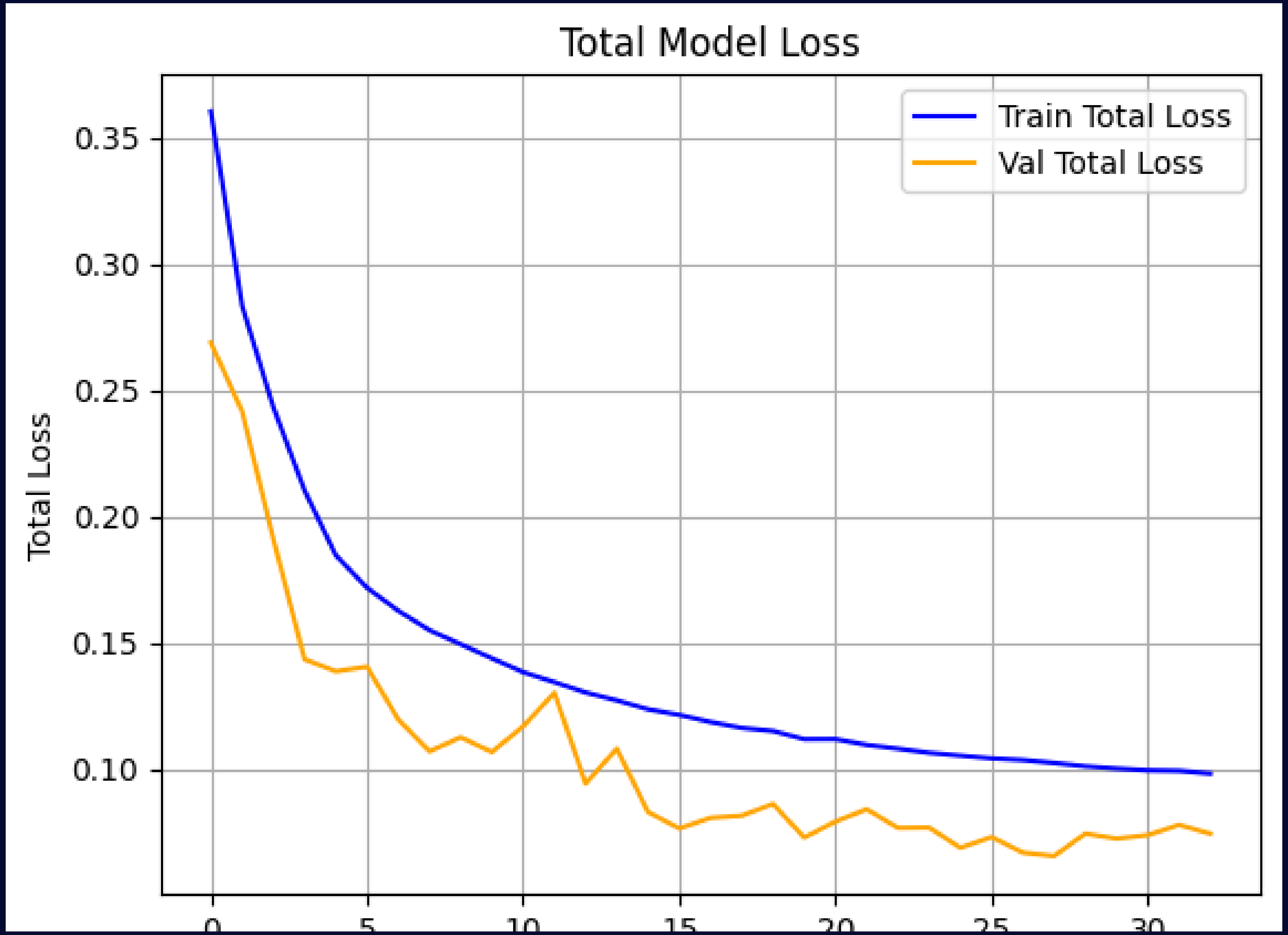
F1 Scores for Different Distance Thresholds with ACK Flooding as Novel Attack



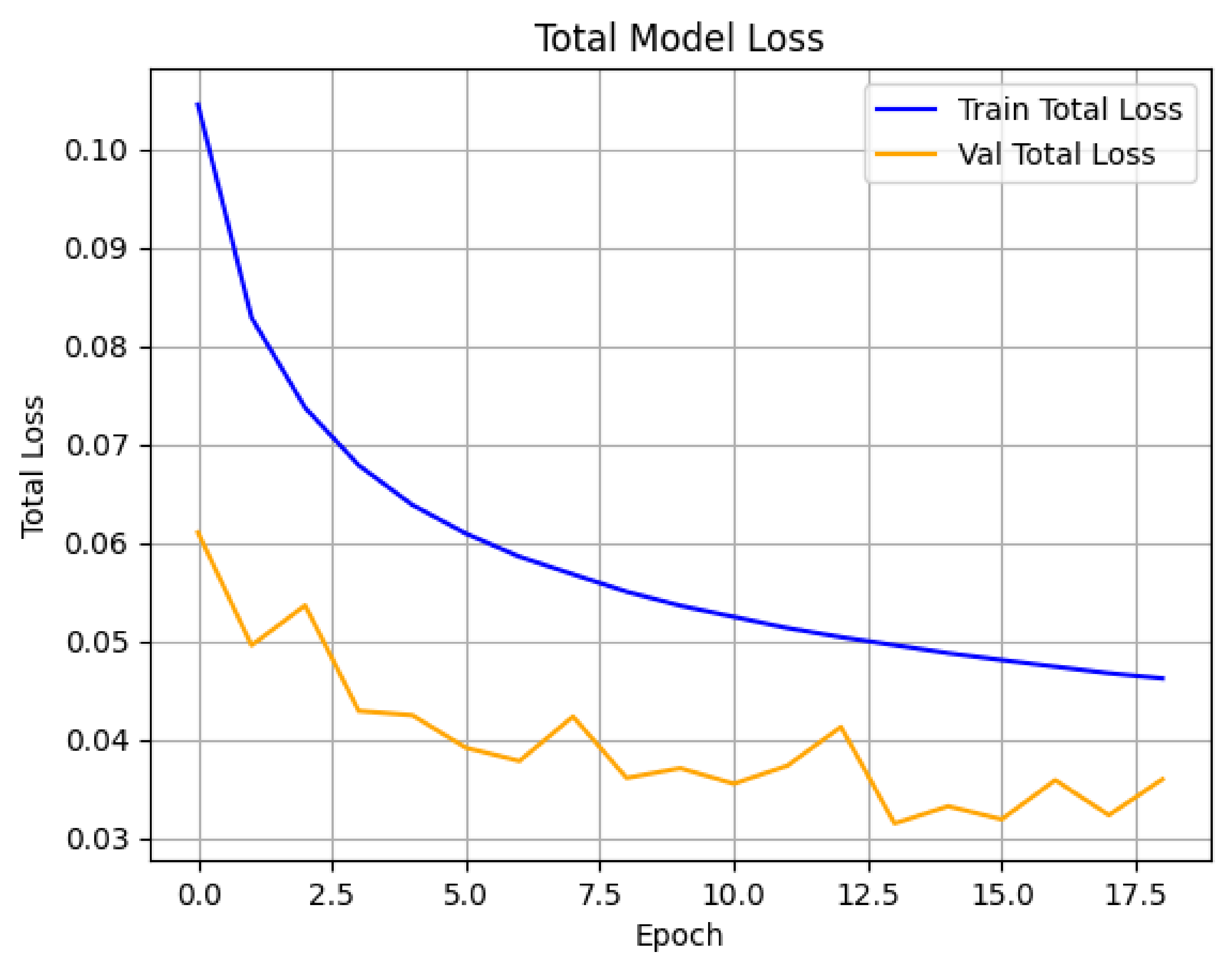
F1 Scores for Different Distance Thresholds with Port Scanning as Novel Attack

Chiao-Hsi Joshua Wang

Appendix C: Model Training/Validation Loss



Multi-Class Model Training/Validation Loss
with ARP Spoofing as Unknown Attack



Binary Class Model Training/Validation Loss
with ARP Spoofing as Unknown Attack