Appendix — Rules Refine the Riddle: Global Explanation for Deep Learning-Based Anomaly Detection in Security Applications

DONGQI HAN, Tsinghua University and Zhongguancun Laboratory, China
ZHILIANG WANG, Tsinghua University and Zhongguancun Laboratory, China
RUITAO FENG, Singapore Management University, Singapore
MINGHUI JIN, State Grid Shanghai Municipal Electric Power Company, China
WENQI CHEN, Tsinghua University and Zhongguancun Laboratory, China
KAI WANG, Tsinghua University and Zhongguancun Laboratory, China
SU WANG, Zhongguancun Laboratory, China
JIAHAI YANG, Tsinghua University and Zhongguancun Laboratory, China
XINGANG SHI, Tsinghua University and Zhongguancun Laboratory, China
XIA YIN, Tsinghua University and Zhongguancun Laboratory, China
YANG LIU, Nanyang Technological University, Singapore

A GEAD TREES AND RULES

This appendix provides the tree plots used in the evaluation in this paper. In each plot, a grey node represents a non-leaf node (feature), and blue and red leaves respectively represent normal or abnormal decisions by the **GEAD** (also the original model if the fidelity is high enough). The grey lines mean the decision path that we do not focus on, while red lines are highlighted for being part of certain rules (the edges of the nodes are also highlighted as red for certain decisions we focus on). For the detailed feature descriptions in each application/dataset, please refer to their papers.

Here is a list for quick reference:

- Fig. 1 and Fig. 2 are used in the §4.3 for Usage 1 (Inaccurate Feature);
- Fig. 3, Fig. 4, Fig. 5 are used in the §4.3 for Usage 1 (Mistaken Label);
- Fig. 6 is used in the §4.3 for Usage 2 (Hyperparameter Selection);
- Fig. 7 is used in the §4.3 for Usage 2 (Training Data Selection);
- Fig. 8 is used in the §4.3 for Usage 2 (Training Data Selection);
- Fig. 9 and Fig. 10 are used in the §4.3 for Usage 2 (Verication of Unlearning);
- Fig. 11 is used in the §4.4 for Insight 1;
- Fig. 12 is used in the §5.

2024. XXXX-XXXX/2024/6-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

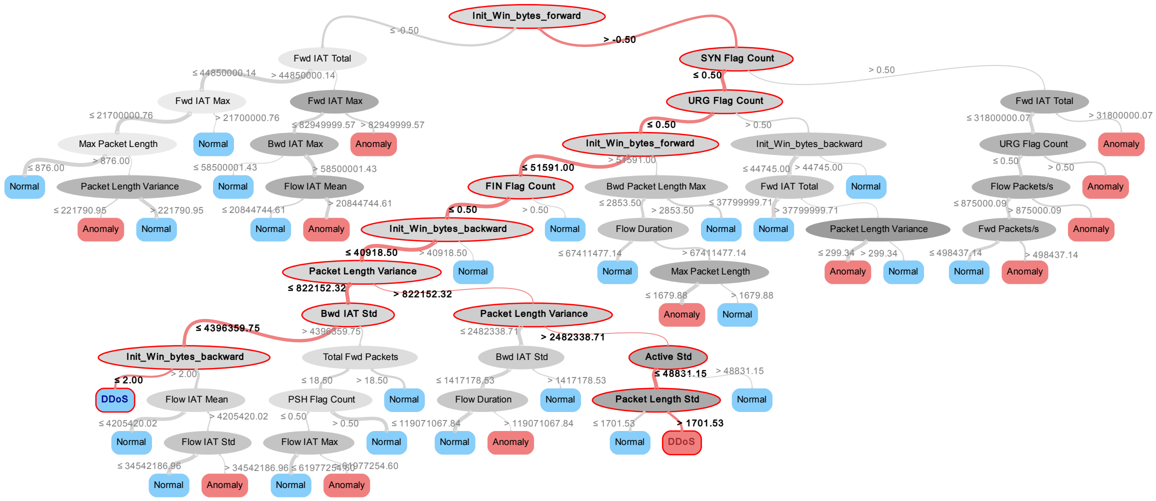


Fig. 1. Autoencoder trained with CICIDS2017 (highlight DDoS rules)

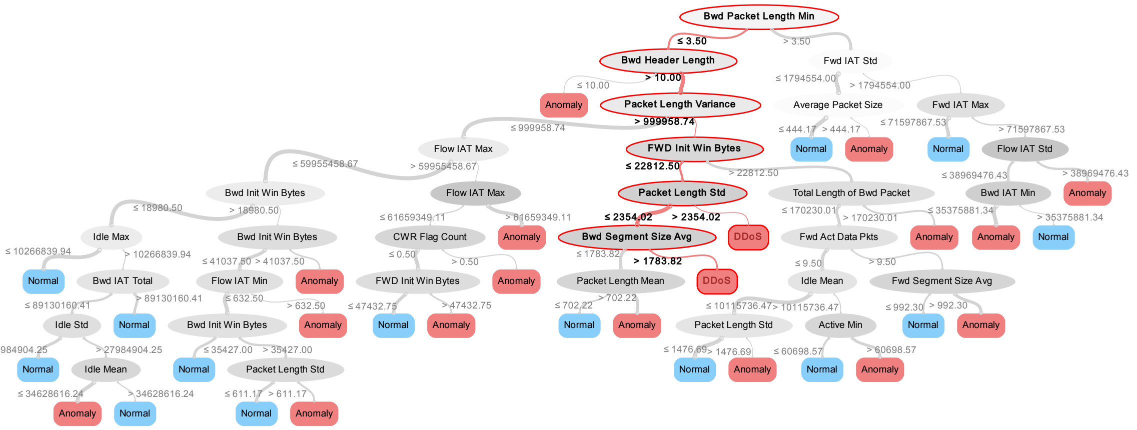


Fig. 2. Autoencoder trained with Improved CICIDS2017 (highlight DDoS rules)

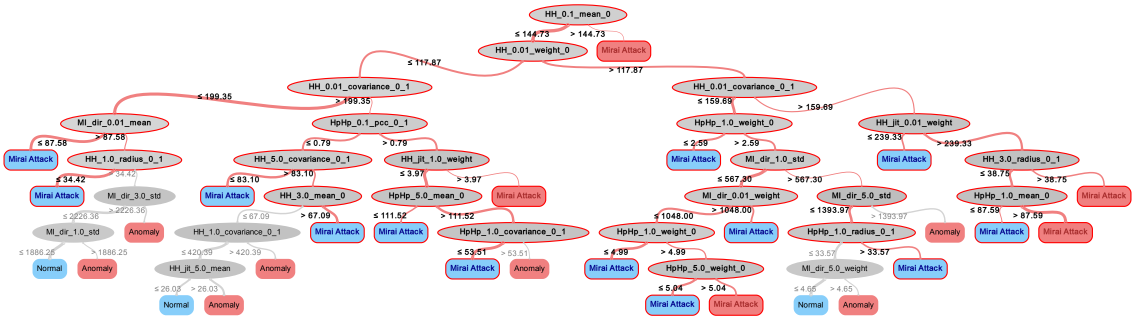


Fig. 3. Autoencoder trained with Kitsune before correcting labels (highlight Mirai rules). Note that many of “Mirai” rules are confirmed with the wrong labels (they are actually normal samples).

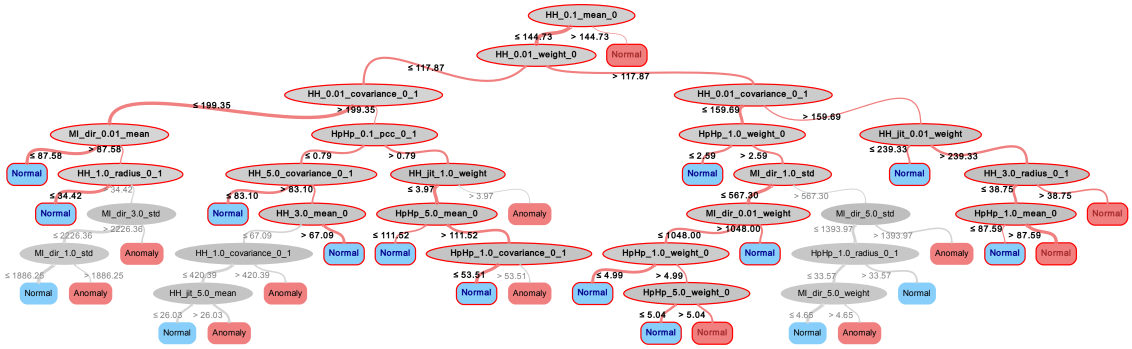


Fig. 4. Autoencoder trained with Kitsune before correcting labels (highlight normal rules)

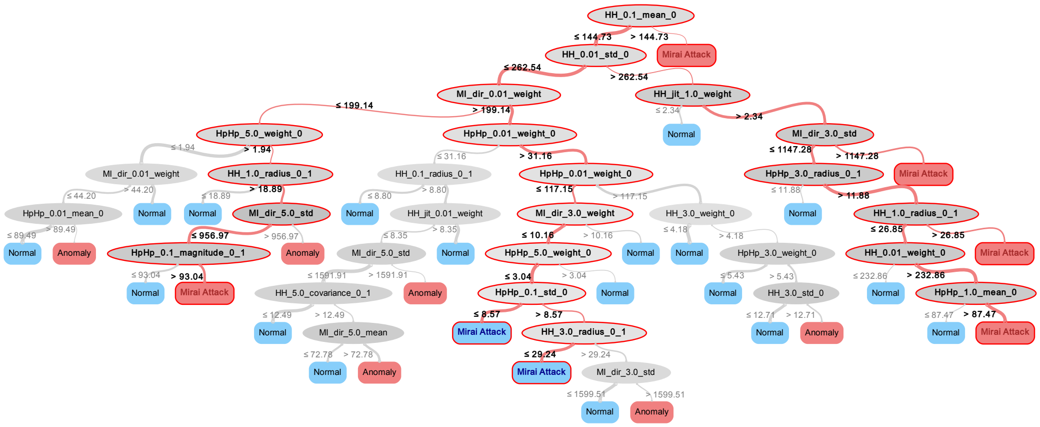
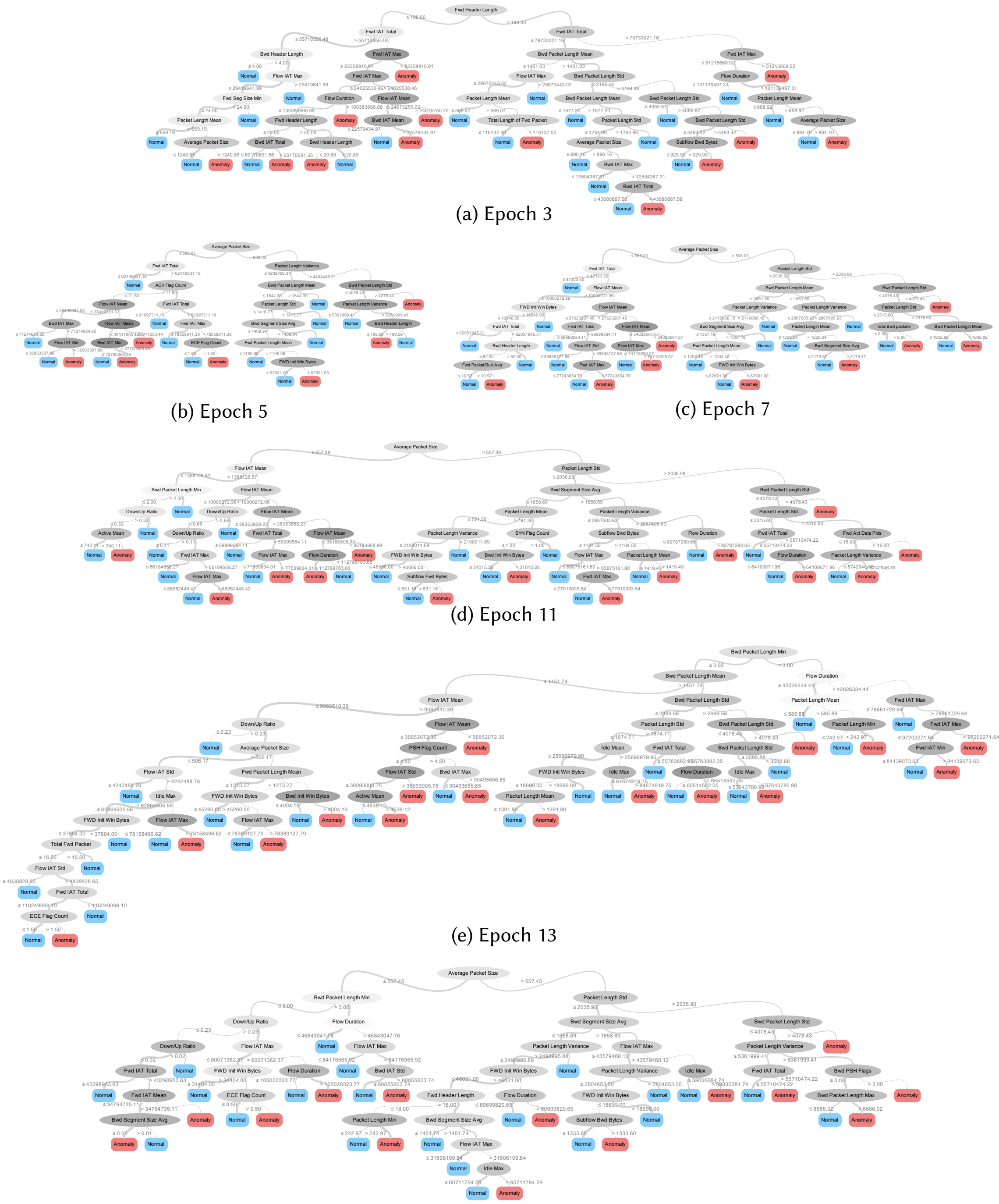


Fig. 5. Autoencoder trained with relabeled Kitsune after correcting labels (highlight Mirai rules)



(a) Epoch 3

(b) Epoch 5

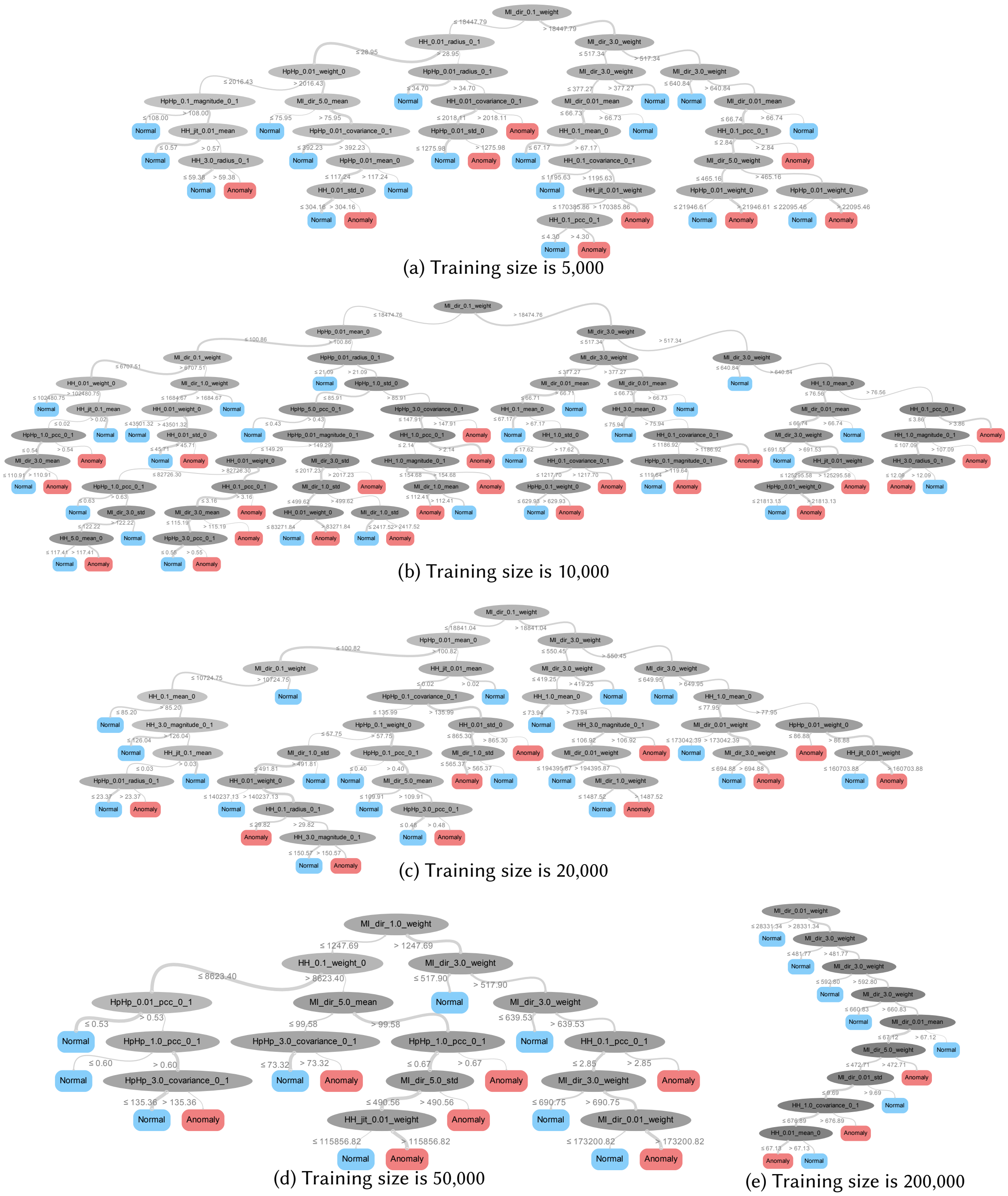
(c) Epoch 7

(d) Epoch 11

(e) Epoch 13

(f) Epoch 15

Fig. 6. Rules of models at different epochs during training.



(a) Training size is 5,000

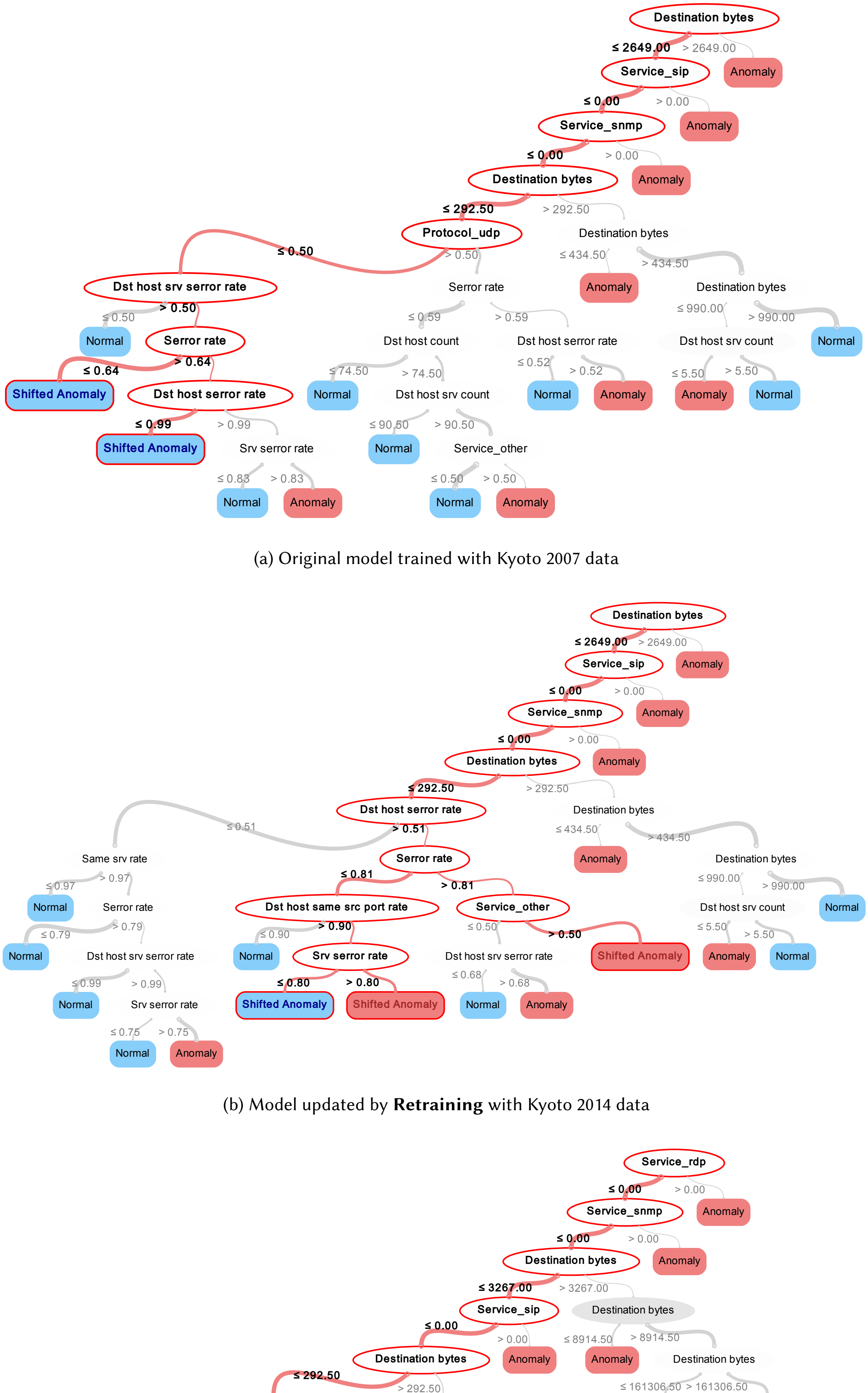
(b) Training size is 10,000

(c) Training size is 20,000

(d) Training size is 50,000

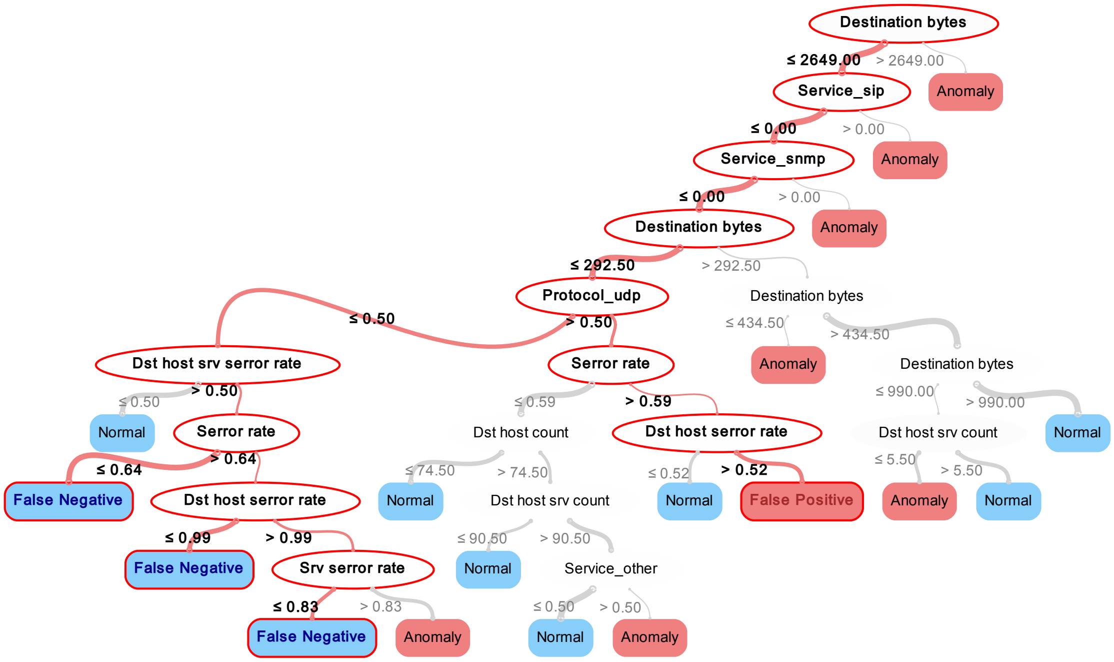
(e) Training size is 200,000

Fig. 7. Rules of models trained with different training sizes.

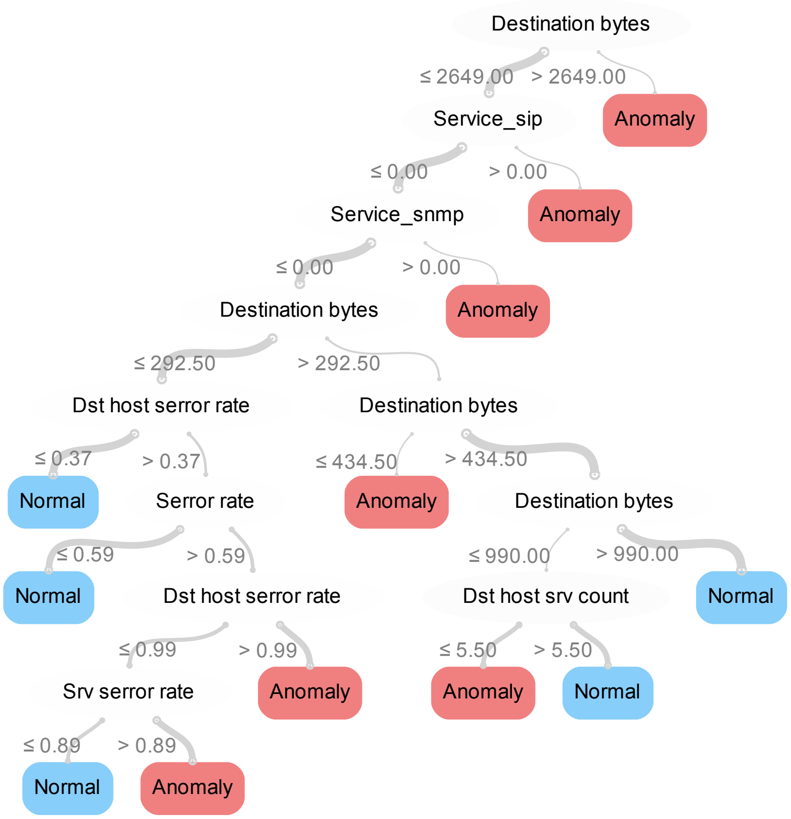


(a) Original model trained with Kyoto 2007 data

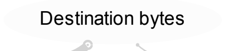
(b) Model updated by **Retraining** with Kyoto 2014 data

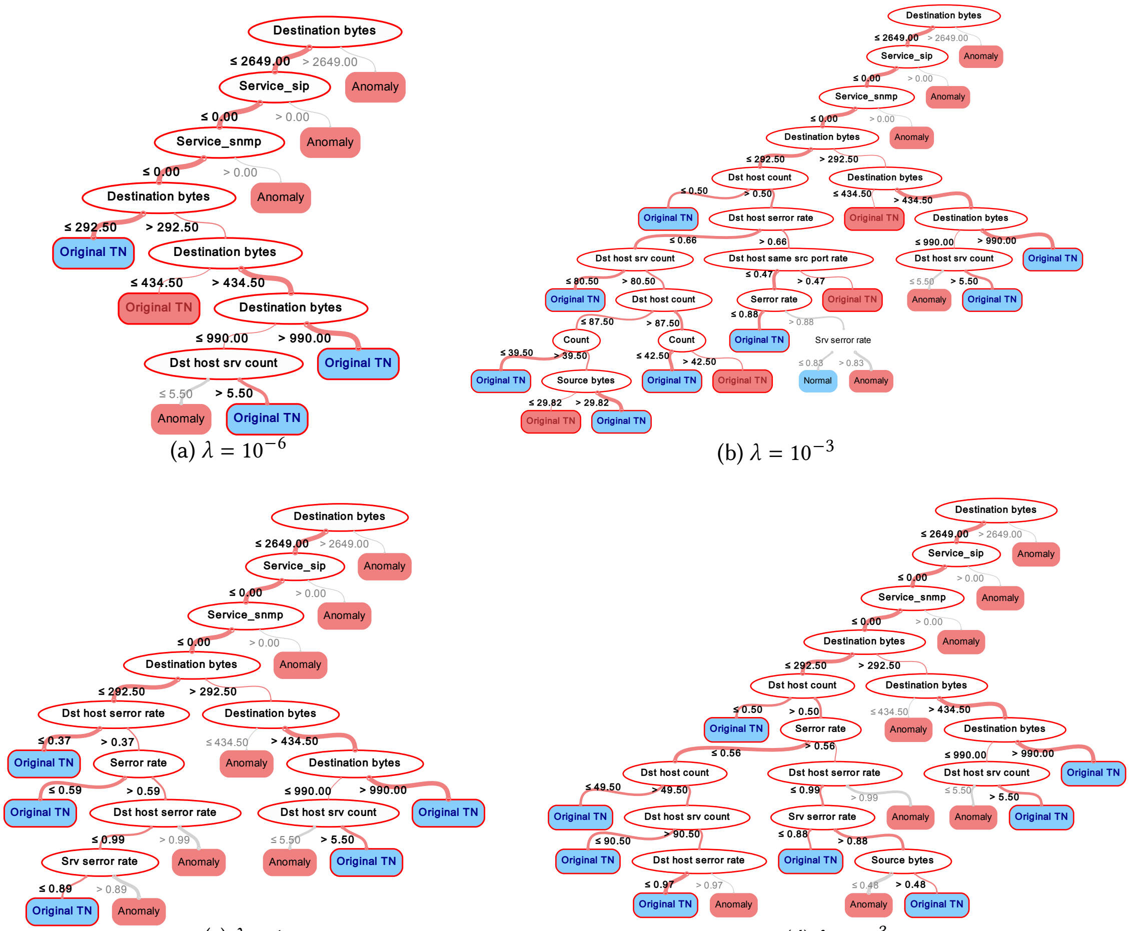


(a) Original model trained with Kyoto 2007 data



(b) Model updated by **UNLERAN** with **FP**s in Kyoto 2014 data





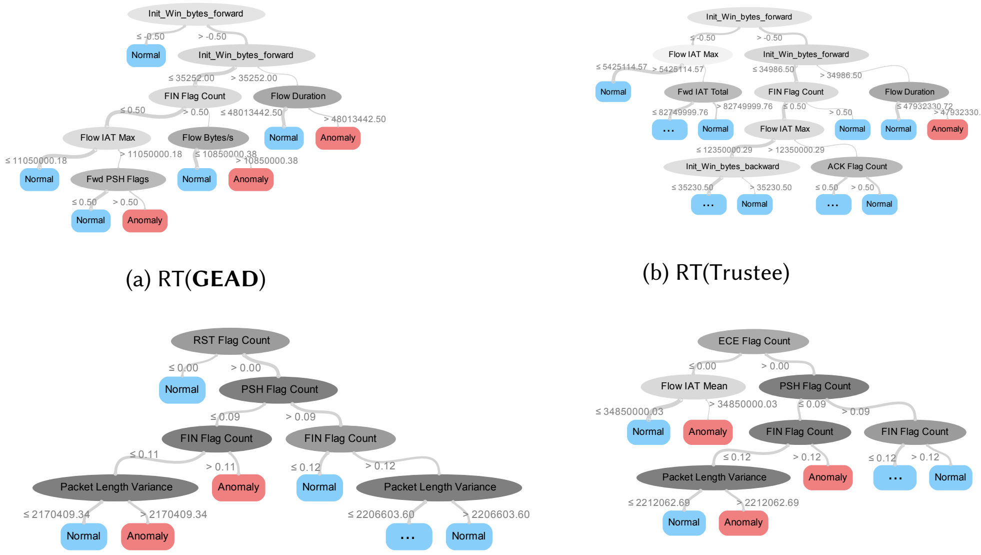
(a) $\lambda = 10^{-6}$

(b) $\lambda = 10^{-3}$

(c) $\lambda = 1$

(d) $\lambda = 10^{3}$

Fig. 10. Model Rules of UNLERAN with different λ.



(a) RT(**GEAD**)

(b) RT(Trustee)

(c) DT(IID+OOD)

(d) DT(Trustee)

Fig. 11. Comparison of rules of different explanations.

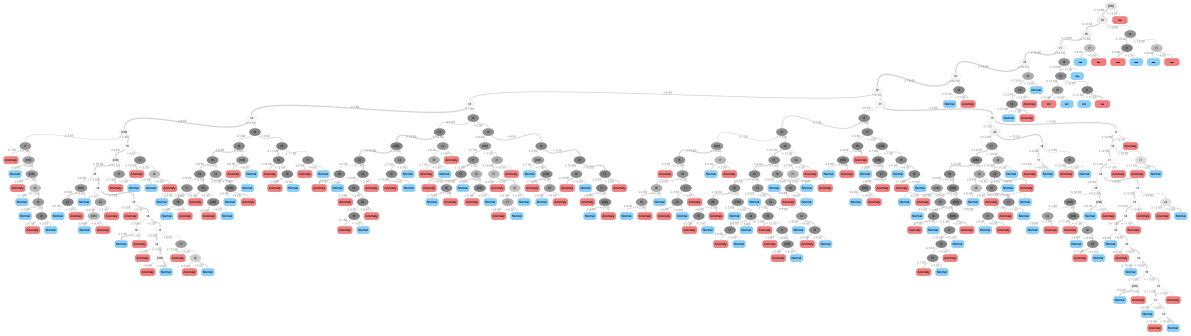


Fig. 12. Desensitized rule tree of DeepLog (device 2) in the real-world use case. The log key is desensitized due to data compliance and privacy restrictions.