Lightweight Traffic Classification: A Simple Baseline Matching Deep Learning Performance

J. Pesek, J. Luxemburk, and K. Hynek

FIT at Czech Technical University in Prague and CESNET

Introduction

Network traffic classification (TC) has advanced significantly due to machine learning, but much of this progress relies on outdated, lab-generated datasets like ISCXVPN2016, CIC-IDS-2017, and CTU-13. In our previous study, we showed that a simple k-nearest neighbors classifier achieved unexpectedly high accuracy on such datasets, attributing this to redundancy—many flows are nearly identical, and random data splits often place similar samples in both training and test sets, inflating performance metrics. Building on this, we evaluated eight widely used TC datasets using a basic 1-nearest neighbor (1-NN) classifier based solely on packet sequences and achieved results on par with or better than state-of-the-art methods. By progressively subsampling the training data, we further confirmed that redundancy significantly skews evaluation outcomes. Our findings suggest that perceived advances in TC may be overstated, highlighting the need for deeper investigation into dataset redundancy and more realistic benchmarking practices.

The Baseline

We adopted the approach proposed in our previous paper, which they describe as a simple baseline. The method forms a feature vector using the first n packets and their corresponding inter-packet times. In our study, we focus specifically on the first **10 packets**, encoding each using three types of features: packet sizes s (clipped to the range $[0, 1500]$), packet directions d (with values ± 1.0), and inter-packet arrival times i (clipped to $[0, 1000]$ ms and scaled by a factor of 0.1). These values are concatenated to form a feature vector. Formally, the feature extraction function Φ, that maps each flow $\mathcal{F}$ to a real-valued vector of length 30 is defined as $\Phi(\mathcal{F}) = (s_1, \ldots, s_{10}, d_1, \ldots, d_{10}, i_1, \ldots, i_{10})$, where all elements have been preprocessed as described. This feature representation is then used for classification via the 1-NN algorithm, which assigns each flow the class label of its closest neighbor in the feature space.

State-of-the-art comparison

We evaluated our approach on multiple popular network traffic datasets. All of the datasets are highly cited and used as a benchmark for evaluating the results of proposed models. As SOTA results, we chose recent papers with reported accuracy to compare our baselines. Except for [1] and [2], no picked paper made any effort to reduce the variance and reported only one number as a final result. Moreover, the compared papers predominantly employ deep learning or related techniques.

Table 1. Comparison of classification performance across datasets and tasks. We report the state-of-the-art accuracy for each dataset, our baseline accuracy (mean $\pm$ standard deviation), the difference in percentage points, and macro-averaged recall.

Dataset	SOTA[%]	Baseline [%]	Δ [pp]	Macro Rec. [%]
ISCXTor2016	100.00 [1]	99.63 ± 0.02	-0.37 ± 0.02	95.54 ± 0.48
USTC-TFC2016	98.30 [3]	96.14 ± 0.64	-2.16 ± 0.64	90.40 ± 0.76
VNAT	98.03 [4]	99.91 ± 0.01	1.88 ± 0.01	98.97 ± 0.34
CTU-13	99.30 [5]	99.46 ± 0.08	0.16 ± 0.08	98.37 ± 0.50
MQTTSet	99.90 [6]	100.00 ± 0.01	0.10 ± 0.01	99.95 ± 0.08
ISCXVPN App	79.92 [2]	72.54 ± 1.05	-7.38 ± 1.68	62.65 ± 2.84
ISCXVPN Traffic	81.71 [2]	74.39 ± 1.35	-7.32 ± 1.85	73.87 ± 1.90
CIC-DoHBr	99.99 [7]	99.96 ± 0.00	-0.03 ± 0.00	99.91 ± 0.02
CIC-IDS-2017	95.79 [8]	99.81 ± 0.02	4.02 ± 0.02	88.01 ± 0.65

Sampling evaluation

We investigated data-sample redundancy in the datasets by randomly subsampling. We iteratively sampled 1%, 5%, 10%, and 50%. It can be observed that using as little as 1% of the training data is sufficient to achieve performance comparable to that obtained with the full dataset.

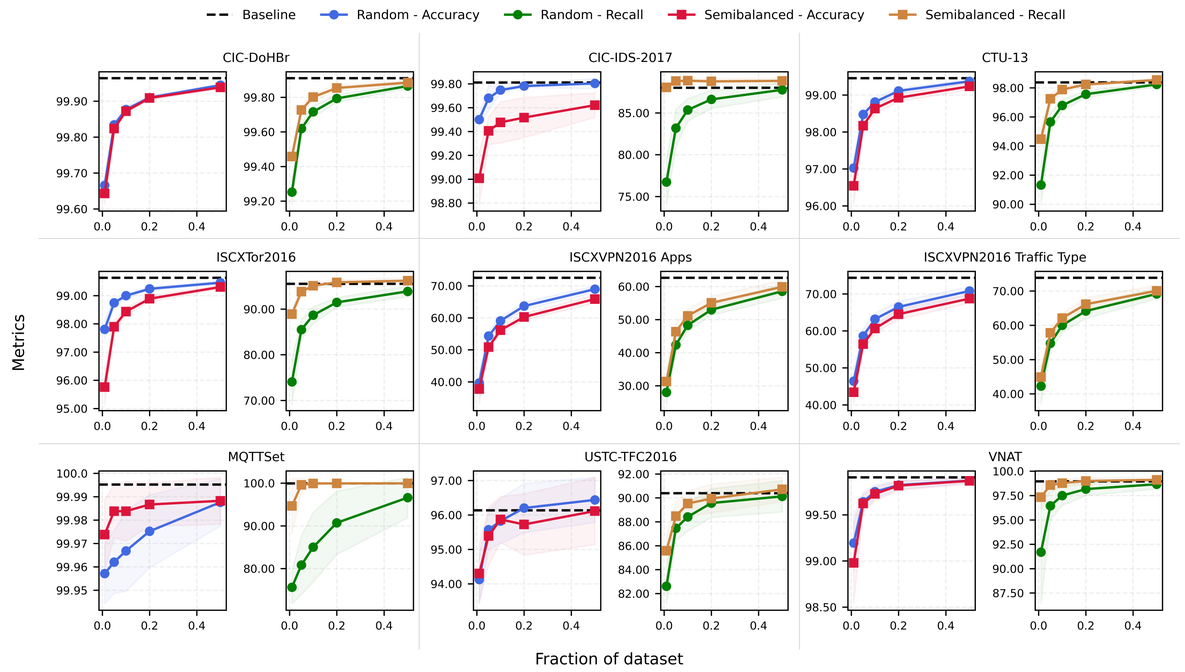

Figure 1. Comparison of classification accuracy (left subplot of each group), weighted accuracy (macro-averaged recall; right subplot of each group), versus the fraction of training data used across multiple network traffic datasets. The black dashed lines indicate the baseline using the full training set.

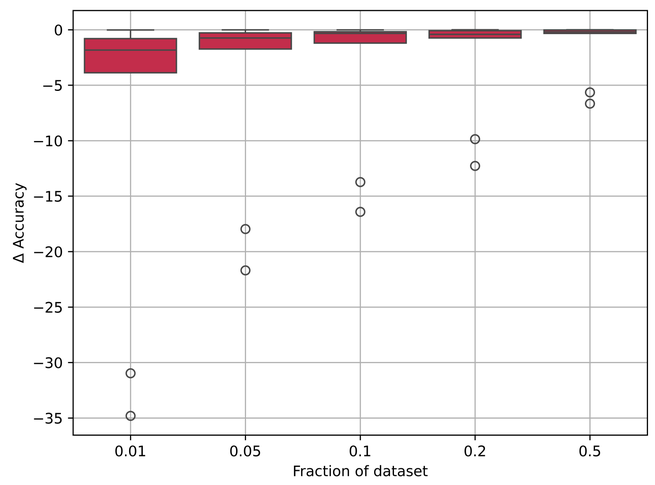

Figure 2. Aggregated accuracy drop relative to the full-data baseline when training on fractions of the dataset (semibalanced sampling). Each box-and-whisker plot pools results across all datasets; open circles mark outliers, which correspond to the two ISCXVPN2016 variants in every case.

References

[1] J. Dai, X. Xu, and F. Xiao, "GLADS: A global-local attention data selection model for multimodal multitask encrypted traffic classification of IoT," vol. 225, p. 109652. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138912862300097X

[2] A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, and A. Pescapé, "Improving performance, reliability, and feasibility in multimodal multitask traffic classification with XAI," vol. 20, no. 2, pp. 1267–1289. [Online]. Available: https://ieeexplore.ieee.org/document/10049138/

[3] Y. Hu, X. Duan, Y. Chen, and Z. Zhao, "Effect analysis of malicious flow classification model based on representation learning on network flow anomaly detection."

[4] R. J. Babaria, M. Lyu, G. Batista, and V. Sivaraman, "FastFlow: Early yet robust network flow classification using the minimal number of time-series packets."

[5] A. Pektaş and T. Acarman, "Deep learning to detect botnet via network flow summaries," vol. 31, no. 11, pp. 8021–8033. [Online]. Available: https://doi.org/10.1007/s00521-018-3595-x

[6] U. Do, L. Lahesoo, R. M. Carnier, and K. Fukuda, "Evaluation of XAI algorithms in IoT traffic anomaly detection," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, ISSN: 2831-6983.

[7] M. T. Jafar, "Analysis and investigation of malicious DNS queries using CIRA-CIC-DoHBrw-2020 dataset," vol. 2, no. 1, number: 1. [Online]. Available: https://www.mjaias.co.uk/mj-en/article/view/24

[8] Y. N. Kunang, S. Nurmaini, D. Stiawan, and B. Y. Suprapto, "Attack classification of an intrusion detection system using deep learning and hyperparameter optimization," vol. 58, p. 102804.

jaroslav.pesek@fit.cvut.cz