Encrypted but Not Private: Application Identification from Traffic Metadata at Scale

Farzam Rezaei, Jorge E. López de Vergara, Luis de Pedro and Iván González
Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
{farzam.rezaei, jorge.lopez_vergara, luis.depedro, ivan.gonzalez}@uam.es

***Abstract*—It is well established that TLS and QUIC protect payload confidentiality but leave packet-level metadata exposed, enabling traffic analysis. However, the practical extent to which application identities can be inferred at backbone scale using only such metadata remains less thoroughly quantified. In this paper, we present a large-scale empirical study of application identification on the CESNET-TLS-Year22 dataset (10.7M flows) under a strict metadata-only setting, relying on packet sizes, inter-packet times, directions, and early packet sequences (PPI), with and without TCP PSH flags. We systematically compare tree-based models (Random Forest, LightGBM, XGBoost) and a GPU-accelerated MLP, evaluating both classification accuracy and inference throughput. Our results confirm and quantify the limited application-level privacy provided by encryption: a PPI-based MLP achieves 95–96% accuracy (macro F1 $>$ 0.94) across hundreds of services while processing millions of flows per second. TCP PSH flags consistently improve performance, whereas combining histogram and sequence features yields marginal gains.**

***Index Terms*—Encrypted traffic classification, TLS privacy, traffic fingerprinting, CESNET-TLS-Year22, PHIST histograms, PPI packet sequences, TCP PSH flags, GPU MLP, application identification, large-scale benchmark.**

I. Introduction

The increasing adoption of TLS and QUIC has reduced network visibility by preventing payload inspection, challenging traffic engineering, monitoring, and security enforcement [1], [2]. While encryption protects payload content, it is well known that packet-level metadata remains exposed, enabling traffic analysis [3], [4]. However, the practical extent to which backbone traffic can be fingerprinted at scale—using only metadata and without handshake identifiers—remains less comprehensively benchmarked. Existing studies typically focus on individual methods or feature sets without systematic comparison under unified conditions [5]–[8].

In this paper, we present a large-scale empirical evaluation of application identification on CESNET-TLS-Year22 [7], systematically comparing histogram-based features and early packet sequences (PPI, with and without TCP PSH flags) across tree-based models and a GPU-accelerated MLP. Our results confirm that application identification remains highly accurate (95–96%) using only packet dynamics, quantifying the limited application-level privacy provided by current encryption. Inference throughput is reported as a measure of deployment feasibility, assessing whether such identification is practical in real-time settings. We emphasize that our contribution is primarily empirical, and that identifying applications from traffic metadata does not by itself constitute a privacy violation without linking connections to specific user identities.

The paper provides the following contributions: (i) a systematic evaluation of histogram and PPI features (including PSH flags) on a large-scale dataset, (ii) a comparison of classical and neural models under a unified setup, and (iii) a discussion of the scope and limitations of metadata-based identification with respect to application-level privacy. To the best of our knowledge, no prior work evaluates these matters at this scale.

The rest of the paper is structured as follows: Section II reviews related work. Section III describes the methodology. Section IV presents experimental results. Section V discusses privacy implications. Finally, Section VI concludes the paper.

II. Related Work

Network traffic classification has evolved from payload-based inspection to metadata-driven approaches due to the widespread adoption of encryption. Early techniques such as port-based identification and DPI are now ineffective or impractical in TLS-dominated environments, motivating the shift toward flow-based methods that rely on observable traffic characteristics (e.g., packet sizes, timing, direction) [5], [6], [9].

Apart from these common features, NEMEA [10] traffic analysis tool also provides two feature families. Histogram-based features (PHIST) summarize packet sizes and inter-packet times into compact distributions, offering efficiency and robustness [8]. Packet Payload Information (PPI) captures the sequence of early packets in a flow, encoding temporal behavior and improving discriminative power, sometimes augmented with TCP flags [8], [11]. These representations are widely supported in CESNET datasets [12], which have become standard benchmarks for realistic evaluation.

A variety of machine learning models have been applied to these features. Tree-based methods such as Random Forest and gradient boosting (e.g., LightGBM) remain strong baselines for tabular data, while neural models (CNNs, RNNs, MLPs) can better exploit sequential patterns but introduce higher computational cost [13]–[15].

Despite avoiding payload inspection, metadata-based classification raises privacy concerns. It is well established that encryption does not hide traffic metadata, enabling the inference of visited websites or user activities from traffic patterns alone [3], [4], [16]. Our work builds on this understanding

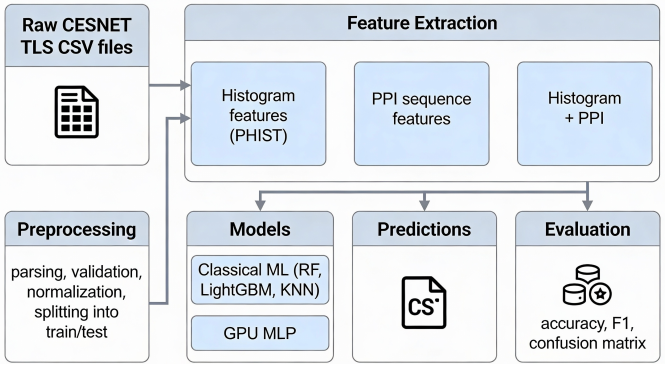

Fig. 1. Overview of the proposed methodology on CESNET-TLS: preprocessing → feature construction (PHIST, PPI, combined) → model training (RF/LightGBM/XGBoost/MLP) → prediction and evaluation.

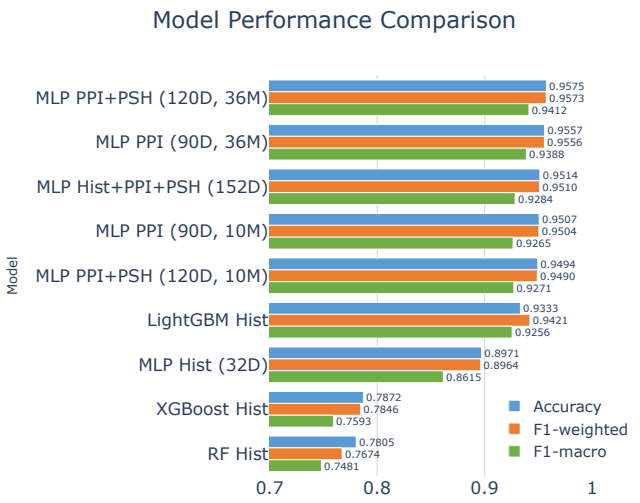

Fig. 2. Overall model performance comparison on CESNET-TLS-Year22 in terms of Accuracy, weighted F1-score, and macro F1-score (higher is better). Numbers in parentheses show the dimension of the input and the number of flows in the training.

by providing a systematic, large-scale benchmark comparing PHIST and PPI features across classical and neural models on CESNET-TLS-Year22, quantifying the extent of application-level information leakage under a strict metadata-only setting.

III. Methodology

We evaluate application identification from encrypted traffic using CESNET-TLS-Year22 under a strict metadata-only setting (packet sizes, timings, directions, and TCP flags). The task is multi-class classification of services (180 classes) using temporally separated train/test splits to reflect realistic deployment, as shown in Figure 1.

A. Features

We consider two standard feature families available in CESNET-TLS: (i) histogram features (PHIST, 32D), which summarize packet sizes and inter-packet times per direction, and (ii) packet sequences (PPI), encoding up to the first $K = 30$ packets using sizes, directions, inter-packet times, and optionally TCP PSH flags (90D without PSH, 120D with PSH). We also evaluate their concatenation (152D) to test complementarity.

B. Preprocessing

Flows are parsed from CSV files, features are constructed (PHIST, PPI, or combined), and labels are encoded from training data. Histograms are L1-normalized; all features are standardized for classical models. No class rebalancing is applied to preserve realistic distributions. Preprocessed datasets are cached to enable reproducible large-scale experiments.

C. Models

We compare tree-based ensembles (Random Forest, LightGBM, XGBoost) and a GPU-based MLP. Tree models are tuned with lightweight search over depth and number of estimators. The MLP consists of two hidden layers with ReLU, dropout, and AdamW optimization; hyperparameters are selected using validation-based tuning. The final model is exported for efficient batch inference.

D. Evaluation

Performance is measured using accuracy and macro F1-score, aggregated across temporally separated test splits. We also report inference throughput (flows/s) to assess deployment feasibility. Experiments are executed on a multi-GPU server to enable large-scale evaluation.

Inference throughput is not merely a systems metric: it directly determines whether metadata-based application identification constitutes a *practical* privacy concern. A classifier that is accurate but very slow for real-time use poses only a theoretical risk; one that processes millions of flows per second can be deployed transparently at backbone scale, making the privacy threat concrete and immediate.

IV. Results

We report accuracy, macro F1, and inference throughput on CESNET-TLS-Year22 under a metadata-only setting. The key questions are: (i) which feature representation is most effective, and (ii) whether models scale to backbone traffic.

A. Main Results

Table I summarizes the best configurations. PPI-based MLPs clearly outperform histogram-only approaches, reaching 95–96% accuracy and macro F1 above 0.93 across hundreds of classes. The best trade-off is achieved by PPI+PSH (120D), which combines high accuracy (∼95.7%) with very high throughput (∼7M flows/s).

Histogram-only models remain competitive but clearly weaker (∼90% accuracy). Among them, LightGBM is the strongest baseline (∼93%), while XGBoost and Random Forest perform significantly worse (∼78%). Figure 2 visualizes the overall accuracy, weighted F1, and macro F1 across the evaluated models.

TABLE I
OVERALL PERFORMANCE AND EFFICIENCY OF THE EVALUATED MODELS ON CESNET-TLS-YEAR22.

Model	Features (dim)	Flows (train / test)	Accuracy	F1 (weighted)	F1 (macro)	Flows/s	Inference time
MLP (2-hidden-layer)	Hist + PPI + PSH (152D)	1 Week / 1 Week	0.9514	0.9510	0.9284	3.30M	3.24 s
MLP (2-hidden-layer)	PPI + PSH (120D)	1 Year (Sampled) / 1 Week	0.9575	0.9573	0.9412	7.09M	1.51 s
MLP (2-hidden-layer)	PPI only (90D); no PSH	1 Year (Sampled) / 1 Week	0.9557	0.9556	0.9388	4.40M	2.43 s
MLP (2-hidden-layer)	PPI + PSH (120D)	1 Week / 1 Week	0.9494	0.9490	0.9271	6.60M	1.62 s
MLP (2-hidden-layer)	PPI only (90D); no PSH	1 Week / 1 Week	0.9507	0.9504	0.9265	5.52M	1.94 s
MLP (2-hidden-layer)	Hist only (32D)	1 Week / 1 Week	0.8971	0.8964	0.8615	1.19M	9.02 s
LightGBM	Hist (32D)	1 Year (Sampled) / 1 Week	0.9333	0.9421	0.9256	0.08M	14.4 s
XGBoost	Hist (32D)	1 Year (Sampled) / 1 Week	0.7872	0.7846	0.7593	≈0.07M	10.76 s
Random Forest	Hist (32D)	1 Year (Sampled) / 1 Week	0.7805	0.7674	0.7481	≈0.005M	20.35 s

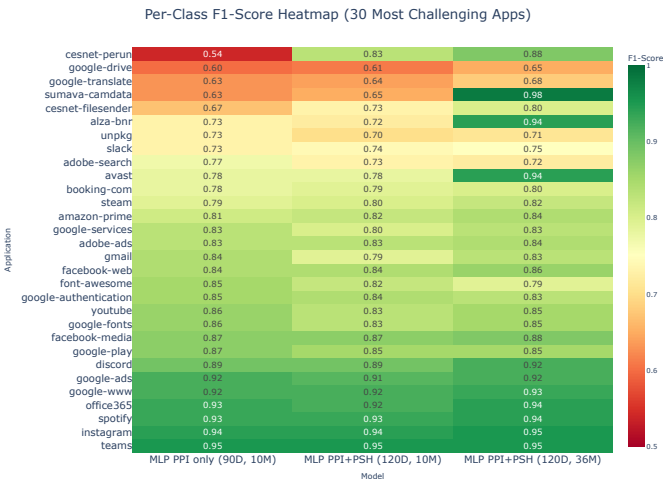

Fig. 3. Per-class F1-score heatmap for the 30 most challenging application classes, comparing representative PPI-based MLP variants. Rows are selected by lowest per-class F1 under the PPI-only (90D, no PSH, 10M) model.

B. Impact of Features

Moving from histograms to packet sequences provides the largest performance gain (∼+6% accuracy). Adding TCP PSH flags yields a smaller but consistent improvement (∼+0.2%). In contrast, combining histograms with PPI provides only marginal benefits, suggesting that early packet dynamics already capture most discriminative information. Figure 3 reports per-class F1-scores for the most challenging classes across representative PPI-based models.

C. Scalability

All MLP models are executed on GPU using TorchScript. The best configuration (PPI+PSH) processes more than 7 M flows/s, enabling real-time backbone deployment. In contrast, tree-based models on CPU remain two orders of magnitude slower (∼70–80k flows/s).

The experiments were run on a high-performance server with two AMD EPYC 7F72 CPUs (24 cores/48 threads each at 3.2 GHz), 1 TB DDR4 RAM, and four NVIDIA A100-SXM4 GPUs (40 GB HBM2 each); throughput values correspond to single-GPU inference with large batches.

Overall, results show that encrypted traffic still leaks sufficient information for accurate application identification, and that this identification is feasible in real time, strengthening the practical relevance of the observed privacy–visibility trade-off.

V. DISCUSSION

Encryption ensures payload confidentiality but does not prevent application-level fingerprinting from traffic metadata. Using only flow-level features, our best MLP (PPI+PSH) achieves ∼95.7% accuracy (macro F1 ∼0.94) across hundreds of classes at multi-million flows/s. These results empirically confirm and quantify a well-known limitation of transport-layer encryption: while payload content is protected, packet sizes, timings, directions, and TCP flags remain highly discriminative [3], [4].

It is important to note, however, that identifying application names or service categories from flow metadata does not, by itself, constitute a direct violation of user privacy. A critical additional step—linking identified connections to specific user identities—would be required for such identification to translate into meaningful privacy harm. Our results therefore characterize the information leakage from encrypted metadata rather than demonstrate a complete privacy attack.

From a practical standpoint, inference throughput is relevant because it determines whether metadata-based identification can be deployed at backbone scale in real time. GPU-based MLPs are both more accurate and orders of magnitude faster than CPU tree ensembles, making such deployment feasible. This has implications both for network operators seeking visibility and for privacy researchers evaluating the real-world threat of traffic analysis. Figure 4 highlights the corresponding relationship of inference speed and macro F1.

Notably, approaches that remove handshake identifiers (e.g., ECH) are unlikely to prevent such inference, as our models rely purely on traffic dynamics. Future defenses such as TLS-level padding [17] or traffic shaping should be evaluated against the baselines established here.

VI. CONCLUSIONS

We presented a large-scale empirical evaluation of metadata-only application identification on CESNET-TLS-Year22, systematically comparing histogram features and early packet sequences (PPI, with/without TCP PSH) across tree-based models and a GPU-accelerated MLP. PPI+PSH with MLP provides the best trade-off, achieving ∼95–96% accuracy with multi-million flows/s throughput.

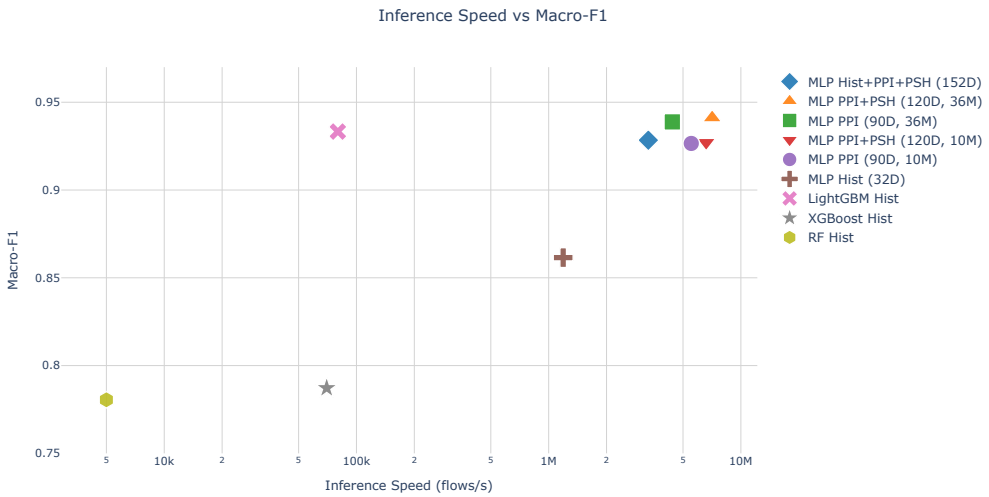

Fig. 4. Inference speed (flows/s, log scale) vs. Macro-F1 across evaluated models, illustrating scalability under high traffic volumes.

Our results confirm and quantify the known limitation that encryption protects payloads but not application identities: early encrypted traffic dynamics suffice for accurate service fingerprinting at backbone scale. While this does not constitute a direct privacy violation—as linking connections to individual users requires additional steps—it highlights a measurable information leakage that merits attention from the privacy community.

Future work should assess the importance of K in the PPI, cross-dataset generalization, robustness to protocol-level defenses (e.g., TLS padding), and lightweight interpretability to understand which features drive predictions. Code and implementation details are publicly available at: GitHub[1].

Acknowledgements

This work is partially funded by a grant from the Dept. of Electronics and Communication Technologies at Universidad Autónoma de Madrid, as well as by an R&D activity program with reference TEC-2024/COM-504 and acronym RAMONES-CM, granted by the Comunidad de Madrid, Spain, through the Directorate General for Research and Technological Innovation via Order 5696/2024.

References

[1] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine learning-powered encrypted network traffic analysis: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 409–454, 2022.

[2] I. A. Alwhbi, C. C. Zou, and R. N. Alharbi, "Encrypted network traffic analysis and classification utilizing machine learning," *Sensors*, vol. 24, no. 11, p. 3509, 2024.

[3] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, and K. Wehrle, "Website fingerprinting at internet scale," in *Proceedings of the 2016 Network and Distributed System Security Symposium (NDSS)*, 2016.

[4] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton, "Peek-a-boo, i still see you: Why efficient traffic analysis countermeasures fail," in *2012 IEEE symposium on security and privacy*. IEEE, 2012, pp. 332–346.

[1]https://github.com/farzamrezaei/Traffic-Classification-of-CESNET-TLS-Year22

[5] A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. ACM, 2005, pp. 50–60.

[6] P. Velan, M. Čermák, P. Čelada, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management*, vol. 25, no. 5, pp. 355–374, 2015.

[7] K. Hynek, J. Luxemburk, J. Pešek, T. Čejka, and P. Šiška, "CESNET-TLS-Year22: A year-spanning tls network traffic dataset from backbone lines," *Scientific Data*, vol. 11, p. 1156, 2024.

[8] J. Luxemburk and T. Čejka, "Fine-grained TLS services classification with reject option," *Computer Networks*, vol. 220, p. 109467, 2023.

[9] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 5, p. 5–16, Oct. 2006. [Online]. Available: https://doi.org/10.1145/1163593.1163596

[10] T. Cejka, V. Bartos, M. Svepes, Z. Rosa, and H. Kubatova, "Nemea: A framework for network traffic analysis," in *12th International Conference on Network and Service Management (CNSM 2016)*, 2016. [Online]. Available: http://dx.doi.org/10.1109/CNSM.2016.7818417

[11] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine learning-powered encrypted network traffic analysis: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791–824, 2023.

[12] J. Luxemburk and K. Hynek, "Datazoo: Streamlining traffic classification experiments," in *Proceedings of the 2023 Workshop on Explainable and Safety Bounded, Fidelitous, Machine Learning for Networking (SAFE '23)*. New York, NY, USA: ACM, 2023, pp. 3–7.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 2017, pp. 3149–3157.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[16] E. Papadogiannaki and S. Ioannidis, "A survey on encrypted network traffic analysis applications, techniques, and countermeasures," *ACM Comput. Surv.*, vol. 54, no. 6, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3457904

[17] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," RFC 8446, Aug. 2018. [Online]. Available: https://www.rfc-editor.org/info/rfc8446