European Journal of Computer Science and Information Technology, 14(5),46-69, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Category-Invariance-Guided Feature Selection (CIFS): Diagnosing and Mitigating the Generalization Gap in Flow-Based VPN Traffic Classification

Rabby Zil Jalal Kanon
Nanjing University of Post and Telecommunication (NJUPT), China

Md Shoriful Islam Fahim
Nanjing University of Post and Telecommunication (NJUPT), China

Khaledujjaman Sohan
Nanjing University of Post and Telecommunication (NJUPT), China

doi: https://doi.org/10.37745/ejcsit.2013/vol14n54669 Published August 23, 2026

Citation: Kanon R.Z.J., Fahim M.S.I., Sohan K. (2026) Category-Invariance-Guided Feature Selection (CIFS): Diagnosing and Mitigating the Generalization Gap in Flow-Based VPN Traffic Classification, *European Journal of Computer Science and Information Technology*, 14(5),46-69

Abstract*: Prior studies on VPN traffic classification using the ISCX VPN-nonVPN dataset report high accuracy (often above 90%) under conventional random train-test splits. This study investigates whether such performance reflects genuine generalization. After deduplication and 20-seed application-disjoint evaluation, Random Forest's accuracy falls from 92.3% to between 35% and 67% when tested on withheld, unseen application categories, and for four of seven categories falls at or below a trivial majority-class baseline. Three independent sanity checks confirm this gap is genuine rather than a data-leakage artifact. A feature ablation study shows the gap is driven by category-specific feature interactions rather than a single overfitting mechanism.As a secondary, partial mitigation, we propose Category-Invariance-Guided Feature Selection (CIFS), a lightweight feature-selection method that removes application-revealing signal from features. On Random Forest, CIFS significantly improves accuracy in two of seven categories after Bonferroni correction (BROWSING, MAIL). However, an identical evaluation on Decision Tree shows no significant benefit and one significant harm, indicating CIFS's benefit is classifier-specific rather than universal.*

Keywords: VPN traffic classification, generalization gap, feature selection, Random Forest, machine learning.

INTRODUCTION

The widespread adoption of Virtual Private Networks (VPNs) has made encrypted traffic a significant challenge for network monitoring, intrusion detection, and policy enforcement, since traditional deep packet inspection cannot access encrypted payloads. To address this, prior work has proposed flow-based statistical features, such as packet size, inter-arrival time, and throughput, as a basis for machine learning-based VPN detection. Using the ISCX VPN-nonVPN (CIC-VPN2016) dataset (Draper-Gil et al., 2016), several studies have reported classification accuracies exceeding 90% using models such as

Decision Tree, Random Forest, and Support Vector Machine (SVM) evaluated under random train-test splits.

However, random splitting does not guarantee that the test set is truly independent of the training set at the application level, since flows from the same application category (e.g., the same chat or streaming service) can appear in both partitions. This raises the question of whether reported accuracies reflect a model's ability to detect VPN usage in general, or whether they partly reflect memorization of application-specific traffic patterns. This study addresses this question in two stages. First, and centrally, it evaluates VPN traffic classifiers under an application-disjoint, unseen-category protocol, in which entire application categories are withheld from training and used only for testing, with duplicate flows removed and every evaluation repeated across 20 independent random splits to ensure the resulting estimates are reliable. Second, having established and diagnosed the resulting generalization gap, it proposes and evaluates Category-Invariance-Guided Feature Selection (CIFS) as a secondary, partial mitigation — a method that directly targets one plausible mechanism behind the gap: features that inadvertently encode application identity rather than VPN status.

The contributions of this study are:

- A rigorous, multi-seed, leakage-checked empirical demonstration of the generalization gap between conventional random-split evaluation and application-disjoint, unseen-category evaluation. This diagnosis, not the mitigation proposed afterward, is the central and most robust contribution of this study.
- A category-wise analysis, benchmarked against a majority-class baseline, identifying which application types are most affected and where the classifier is systematically biased rather than merely uncertain.
- A feature ablation study showing that feature-group importance for generalization is category-specific rather than universal, indicating the generalization failure is driven by category-specific feature interactions rather than a single overfitting mechanism.
- CIFS, a lightweight feature-selection method designed to be model-agnostic in principle, evaluated as a partial mitigation for the diagnosed gap. On Random Forest, it improves unseen-category accuracy in four of seven categories at the conventional $\alpha = 0.05$ level (two of seven — BROWSING and MAIL — after Bonferroni correction), with no significant harm. A Decision Tree extension (Extension to Decision Tree section) finds no significant benefit and one significant harm (CHAT), indicating the method's benefit has not yet been demonstrated to be model-agnostic in practice.
- Paired effect sizes (Cohen's d) and bootstrap 95% confidence intervals for all seven CIFS-vs-baseline comparisons, confirming the pattern of significance results and surfacing one rank-based/mean-based divergence (P2P) attributable to two outlier seeds rather than a genuine effect.
- A detection-efficiency analysis comparing the practical deployability of the evaluated models.
- A literature cross-check that identifies closely related independent work and honestly repositions our contribution around what is genuinely distinct, with every cited reference independently re-verified against its original source prior to submission.

- A transparent account of how earlier, less rigorous versions of this analysis produced conclusions that did not survive duplicate-removal, multi-seed verification, multiple-comparison correction, or cross-classifier extension, and how each was revised accordingly.

RELATED WORK

The ISCX VPN-nonVPN dataset (Draper-Gil et al., 2016) was introduced to support flow-based VPN traffic classification research, and has since been used in numerous studies applying conventional machine learning and deep learning models to distinguish VPN from non-VPN traffic, typically reporting accuracies above 90% under random train-test splitting. While these studies establish the feasibility of flow-based VPN detection, they generally do not assess whether high accuracy persists when the model is evaluated on application categories not represented during training, nor do they report data-cleaning steps (such as duplicate removal) or variability across repeated evaluations, leaving open the questions of real-world generalization and result stability addressed in this study.

Separately, the broader machine learning literature has long studied domain-invariant representation learning, in which models are trained to rely on signal that transfers across domains while suppressing domain-specific information, typically via adversarial training or distribution-alignment objectives in deep networks. Within encrypted-traffic analysis specifically, GETA (Gunasekara et al., 2026) applies meta-learning, embedding refinement, and self-attention to support few-shot adaptation to unseen domains, including VPN traffic classification, across nine public datasets. CIFS addresses a related underlying problem — generalizing to conditions not seen during training — but does so through a lightweight, interpretable feature-selection step compatible with simple classifiers (Decision Tree, Random Forest), rather than through a deep meta-learning representation, and specifically targets application-category generalization rather than cross-dataset or cross-domain adaptation.

Closely related work identified after initial drafting

During a literature cross-check requested by the supervisor, we identified BiasSeeker (Wang, Xie, Wang, & Cui, 2026), a framework that detects “shortcut” features in encrypted network traffic classification using Adjusted Mutual Information (AMI) to rank features by how strongly they correlate with the class label, on the premise that highly label-correlated but semantically task-irrelevant features are more likely to be dataset-specific shortcuts. BiasSeeker is evaluated across 19 datasets and three classification tasks, including VPN detection on ISCXVPN2016. This is conceptually close to the dual-scoring step in CIFS (Method Section), and we report the overlap transparently rather than claiming the underlying feature-scoring principle as novel.

TABLE 1. Comparison between BiasSeeker and CIFS.

Aspect	BiasSeeker (Wang et al., 2026)	CIFS
Core mechanism	Adjusted Mutual Information (AMI) ranks features by label-correlation	Mutual-information ratio: VPN-relevance vs. category-relevance
Feature type	Raw packet-level fields (IP, port, TCP timestamp, etc.)	Flow-level statistical features (packet size, timing, throughput)
Dataset overlap	Includes ISCXVPN2016 (1 of 19 datasets, 3 tasks)	ISCXVPN2016 (CIC-VPN2016) only
Evaluation protocol	Occlusion-based mitigation on random train/test splits; no category holdout	Application-disjoint (leave-one-category-out) generalization + majority-baseline comparison
Feature-count selection	Manual top-k inspection guided by domain knowledge	Adaptive per-category k, chosen via leakage-free internal validation split (plus two always-retained throughput features, Method section)
Model scope	Not restricted to a specific classifier family	Designed to be model-agnostic; validated on Random Forest, with a Decision Tree extension (Extension to Decision Tree section) showing no significant benefit

The two methods differ in several respects that we believe are material rather than cosmetic. BiasSeeker operates on raw packet-level protocol fields and mitigates shortcuts via occlusion (zero-padding, relative transformation, random masking) evaluated on conventional random train/test splits; it does not evaluate generalization to an entirely unseen category held out from training. CIFS instead operates on flow-level statistical features and is evaluated specifically under an application-disjoint, leave-one-category-out protocol with a majority-class baseline, and selects its feature-count threshold adaptively per held-out category via a leakage-free internal validation split — a step with no counterpart in BiasSeeker's manual top-k inspection. We therefore position the contribution of this study not as the feature-scoring principle itself, but as (a) the diagnosis of a category-level generalization gap using a rigorous, majority-baseline-referenced, multi-seed protocol, and (b) an adaptive application of category-invariant feature selection to directly mitigate that specific, diagnosed gap — with (a) being the primary and (b) the secondary contribution.

Beyond BiasSeeker and GETA, several other recent works bear directly on the evaluation protocol and mechanism used in this study. Pekár, Plný, and Hynek (2026) provide a tutorial-style treatment of flow-based traffic classification methodology that explicitly defines Random, Disjoint, and Out-of-Distribution (OOD) split strategies, describing the OOD split as a specific case of the disjoint split in which the test set contains categories entirely unseen during training. This terminology aligns with, and lends methodological grounding to, the application-disjoint protocol used throughout Result section.

Huang, Li, and Yang (2026) study a related but distinct failure mode in encrypted traffic pretraining: they argue that byte-level pretrained representations for traffic classification, evaluated in part on ISCX-VPN, are undermined by dataset-specific, unlearnable fields being treated as reconstruction targets during pretraining. While their setting (self-supervised byte-level pretraining) differs from ours (supervised classification on flow-level statistical features), their diagnosis is conceptually consistent with this study's finding that certain features encode dataset- or category-specific artifacts rather than transferable signal.

Elshewey and Osman (2025) report that a stacked deep ensemble (combining CNN, RNN, DNN, LSTM, and GRU) retains high macro-F1 and macro ROC-AUC under temporal and GroupKFold out-of-distribution splits on the CIC/ISCX VPN-nonVPN benchmark, among other datasets. This is evidence that at least one prior study has evaluated OOD generalization on the same benchmark family, but using a deep ensemble architecture rather than a feature-selection intervention on classical classifiers; a direct comparison between the two approaches under an identical held-out-category protocol is left for future work.

Two further works illustrate alternative strategies for improving generalization in encrypted/network traffic classification, against which CIFS can be contrasted. Qiu et al. (2023), in 3D-IDS, use a non-parametric, mutual-information-based optimization to disentangle flow features for intrusion detection, aiming to automatically separate general attack-indicative features from attack-specific ones — a feature-disentanglement goal related in spirit to CIFS's category-invariance scoring, though applied to a different task (multi-class intrusion detection rather than binary VPN detection) and via a different mechanism (disentangled representation learning rather than discrete feature subset selection). Zhan, Yang, Jia, and Fu (2025), in EAPT, instead pursue generalization through adversarial pre-training of a transformer model, reporting improved generalization ability from a training-time adversarial objective rather than a feature-selection step. Together, these works indicate that improving generalization in this space is an active area, with prior approaches spanning representation disentanglement, adversarial training, and meta-learning; CIFS contributes a lightweight, feature-selection-based alternative applicable to simple, non-deep classifiers, evaluated under an explicit application-disjoint protocol with a majority-baseline reference.

METHODOLOGY

Dataset

This study uses the ISCX VPN-nonVPN (CIC-VPN2016) dataset (Draper-Gil et al., 2016), which provides pre-extracted flow-based statistical features (e.g., forward/backward inter-arrival time statistics, flow duration, packets/bytes per second, and active and idle time statistics) for network flows labeled by application category (Browsing, Chat, File Transfer, Mail, P2P, Streaming, VOIP) and by VPN status (VPN or non-VPN). The raw file contains 59,706 flows.

Preprocessing

Rows containing sentinel/invalid values (−1, used by the feature extractor to denote an unmeasured active/idle period) were removed, reducing the dataset from 59,706 to 25,648 flows.

Exact duplicate rows were then removed, a step absent from an earlier version of this analysis. This removed a further 4,623 duplicate rows, for a final cleaned dataset of 21,025 flows. Duplicate records, if left in place, can inflate conventional (random-split) accuracy by allowing near-identical rows to

appear in both the training and test partitions; their removal is therefore treated as a required preprocessing step rather than an optional one.

A binary VPN-status label (VPN = 1, Non-VPN = 0) was derived from the traffic_type field. The original application-category label was retained separately to enable application-disjoint splitting. Feature scaling (StandardScaler) was fit only on the training partition and applied to the test/held-out partition, to avoid preprocessing leakage.

Splitting Strategy

Two evaluation protocols are used. (1) Random split (baseline): a conventional 80/20 stratified random split, consistent with prior literature, used as a reference point. (2) Application-disjoint / unseen-category split: for each of the seven application categories, all flows belonging to that category (VPN and non-VPN versions) are withheld entirely from training and used only as a held-out test set; the remaining six categories are split 80/20 (stratified) to train and validate the model under normal conditions. This is repeated once per category, so that every category is used as the unseen test set exactly once.

To ensure results reflect genuine generalization behavior rather than the outcome of a single, arbitrarily chosen split, every unseen-category evaluation reported for Random Forest (Category-Wise Analysis to Feature Ablation Study sections) was repeated across 20 independent random 80/20 splits of the training pool (seeds 0–19), keeping the held-out category's test set fixed. Results are reported as mean ± standard deviation. The three-model comparison (Model Comparison section) was repeated across 10 splits per category, a reduced count chosen to keep SVM training time tractable. For each held-out category, a majority-class baseline (the accuracy obtainable by always predicting the more frequent class within that category's test set) is also reported, to distinguish reduced accuracy from systematic classification bias.

A session-disjoint split, in which flows from the same capture session are constrained to appear only in the training set or only in the test set, was also considered. However, neither the processed flow-level dataset used in this study nor the corresponding official CIC-hosted flow-feature release (verified directly against the official dataset browser) retains session or flow identifiers (e.g., source/destination IP, port, or timestamp), which are typically present in the original ISCXFlowMeter/CICFlowMeter output. Session-disjoint evaluation was therefore not conducted; this is discussed further in the Limitations section.

Models and Evaluation Metrics

Three classifiers were evaluated: Decision Tree, Random Forest, and SVM (RBF kernel), consistent with common choices in prior literature, using default hyperparameters in all cases. Performance was measured using accuracy, together with a confusion matrix for a detailed error analysis of one representative case. A feature ablation study measured the effect of removing specific feature groups (timing-related, throughput-related, and active/idle-related features) on both normal and unseen-category accuracy, averaged over 20 random splits per configuration. A detection-efficiency analysis measured prediction time per model to assess practical deployability.

Sanity Checks

Before reporting the results below, three sanity checks were performed to rule out data leakage or pipeline bugs as an explanation for the observed conventional accuracy of approximately 92%. (1) Feature-column check: none of the 23 feature-column names contain the substrings 'vpn' or 'traffic', ruling out an accidental label proxy among the features. (2) Train/held-out overlap check: for each of the seven unseen-category evaluations, the set intersection between the traffic_type values present in the training pool and those in the held-out category was confirmed to be empty, ruling out category leakage across the split. (3) Label-permutation (shuffle) test: the vpn_status label was randomly permuted (breaking any real relationship between features and label) and the full training/evaluation pipeline was re-run unchanged. Table 2 reports the resulting accuracy for each category, alongside the majority-class baseline computed on the same shuffled test split.

Table 2. Label-permutation sanity check: accuracy with vpn_status randomly shuffled, compared to the majority-class baseline on the shuffled split.

Category	Shuffled-Label Accuracy	Baseline (shuffled split)
BROWSING	55.3%	61.1%
CHAT	52.6%	55.4%
FT	50.7%	52.6%
MAIL	51.8%	55.4%
P2P	52.4%	57.1%
STREAMING	51.7%	55.2%
VOIP	51.1%	52.7%

With labels shuffled, accuracy falls to between 50.7% and 55.3% across all seven categories — close to chance level for a binary classification task, and well below the approximately 92% conventional accuracy obtained with real labels. Had the pipeline contained a leakage bug (e.g., a label proxy among the features, or improper train/test separation), shuffled-label accuracy would be expected to remain high regardless of the label assignment. Its collapse to near-chance level, together with the two structural checks above, indicates that the conventional accuracy reported in this study reflects a genuine relationship between the flow-based features and VPN status, rather than a data-pipeline artifact.

RESULTS: CHARACTERIZING THE GENERALIZATION PROBLEM

Model Comparison (10-Seed, Deduplicated Data)

Table 3 presents the performance of the three classifiers, averaged across all seven held-out categories, based on 10 random splits per category.

Table 3. Model comparison: conventional vs. unseen-category accuracy, mean ± std across categories and 10 seeds.

Model	Normal Accuracy (mean)	Unseen-Category Accuracy (mean ± std)
Random Forest	92.2%	56.7% ± 13.1%
Decision Tree	88.3%	55.6% ± 11.0%
SVM	64.4%	43.3% ± 14.1%

Random Forest and Decision Tree achieve similar unseen-category accuracy on average, with Random Forest holding a clearer advantage in conventional accuracy; SVM performs markedly worse on both measures. All three classifiers show a substantial drop from conventional to unseen-category evaluation, indicating that the generalization challenge is not specific to one model architecture.

Category-Wise Analysis (Random Forest, 20-Seed, Deduplicated Data)

Table 4 reports, for Random Forest, the mean and standard deviation of unseen-category accuracy across 20 random splits, for each held-out application category, together with the majority-class baseline for that category's test set.

Table 4. Unseen-category accuracy by held-out category, Random Forest, 20-seed mean ± std, with majority-class baseline (deduplicated data).

Held-Out Category	Normal Acc. (mean)	Unseen Acc. (mean ± std)	Majority Baseline	Beats Baseline?
FT	92.4%	66.6% ± 1.6%	73.6%	No
BROWSING	92.8%	65.2% ± 1.3%	50.4%	Yes
CHAT	92.6%	63.8% ± 0.9%	54.7%	Yes
MAIL	92.1%	61.1% ± 1.0%	51.3%	Yes
VOIP	92.0%	51.7% ± 17.2%	76.1%	No
STREAMING	92.1%	50.9% ± 3.2%	57.6%	No
P2P	92.0%	35.0% ± 4.9%	88.9%	No

Three of seven categories (BROWSING, CHAT, MAIL) show unseen-category accuracy clearly above the majority-class baseline. The remaining four (FT, VOIP, STREAMING, P2P) fall at or below their respective baselines, indicating that for these categories the classifier's predictions are not merely less accurate but systematically unhelpful relative to a trivial always-predict-the-majority-class rule. This is most severe for P2P, where accuracy (35.0%) is 53.9 percentage points below the baseline (88.9%). VOIP shows by far the highest split-to-split variance (std 17.2%, roughly 4–19 times that of other categories), reflecting its comparatively small, imbalanced held-out set.

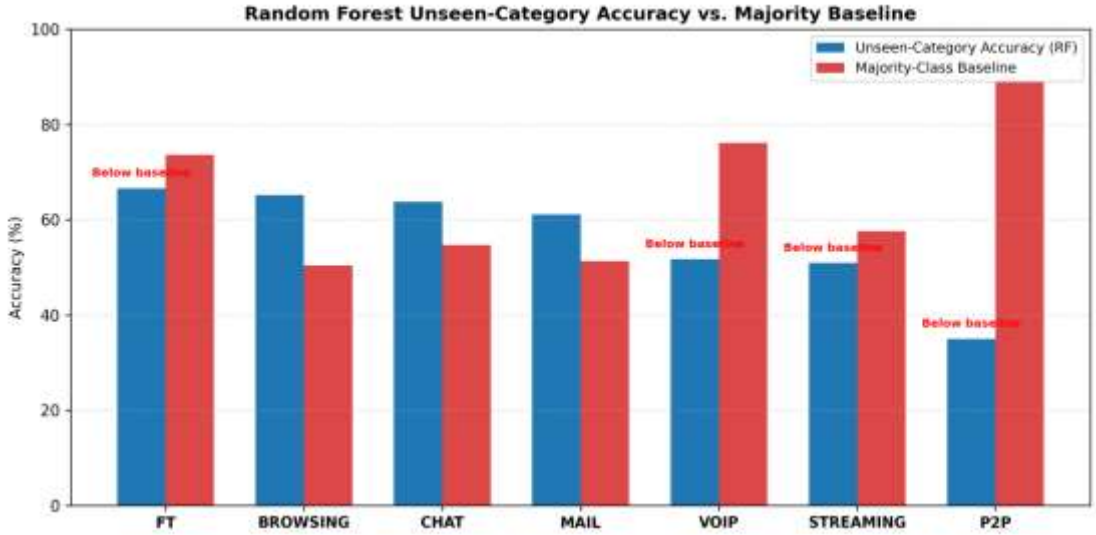

Figure 1. *Random Forest unseen-category accuracy versus the majority-class baseline, all seven held-out categories (Table 4). Bars marked "Below baseline" indicate categories where the classifier performs worse than trivially predicting the majority class.*

VOIP Case Study

A confusion matrix analysis of the Random Forest classifier on one representative held-out VOIP split (seed 42) is shown below. Of 2,332 total VOIP-category test flows, 1,775 were actually VPN traffic and 557 were actually non-VPN traffic (a 76.1%/23.9% split, consistent with the majority baseline in Table 4).

Table 4a. Confusion matrix, VOIP held out, seed 42 (rows: actual; columns: predicted).

	Predicted Non-VPN	Predicted VPN	Total (Actual)
Actual Non-VPN	151	406	557
Actual VPN	968	807	1,775

This split yields an overall accuracy of 41.1%, a VPN recall of 45.5%, and a VPN precision of 66.5%. Of all 2,332 predictions, 1,213 (52.0%) were predicted VPN and 1,119 (48.0%) were predicted non-VPN — a roughly balanced split. This is a materially different picture from an earlier, non-deduplicated analysis of this dataset, which reported that 80.4% of predictions were non-VPN and interpreted this as a systematic bias toward the non-VPN class. With duplicates removed, the classifier's predictions on unseen VOIP traffic are close to evenly split between the two classes; the shortfall relative to the majority baseline arises because the classifier fails to recognize a large fraction of actual VPN-VOIP flows (recall 45.5%) in a test set that is 76.1% VPN, rather than from a directional bias toward either label. We report this revision because it changes the mechanism, though not the overall conclusion (that unseen VOIP accuracy is well below the majority baseline).

Feature Ablation Study

Table 5 shows the 20-seed mean and standard deviation of unseen-category accuracy on the VOIP held-out set, under four feature configurations.

Table 5. Feature ablation results, VOIP held out, 20-seed mean ± std (deduplicated data).

Feature Configuration	Normal Accuracy (mean)	VOIP Unseen Accuracy (mean ± std)
Full (all) Features	92.0%	51.7% ± 17.2%
Without Timing Features	83.7%	25.6% ± 0.6%
Without Throughput Features	91.3%	37.2% ± 13.6%
Without Active/Idle Features	91.8%	42.3% ± 11.0%

For VOIP, all three feature-removal configurations reduce mean unseen accuracy relative to the full feature set, most severely for timing features (51.7% to 25.6%, with a comparatively narrow standard deviation, indicating a consistent effect) and least severely for active/idle features (51.7% to 42.3%). Removing active/idle features also reduces variance substantially (standard deviation 17.2% to 11.0%). This VOIP-specific result differs from an earlier, non-deduplicated analysis, which had found that removing active/idle features increased VOIP unseen accuracy; with duplicates removed, the direction of this effect reverses, underscoring the importance of the deduplication step for this particular finding.

To test whether feature-group effects generalize beyond VOIP, all four configurations were evaluated across all seven held-out categories. Table 6 focuses on the comparison between the full feature set and the set without active/idle features.

Table 6. Unseen-category accuracy (mean ± std across 20 splits) with and without active/idle features, all seven categories (deduplicated data).

Held-Out Category	Full Features (mean ± std)	Without Active/Idle (mean ± std)
BROWSING	65.2% ± 1.3%	65.9% ± 1.2%
CHAT	63.8% ± 0.9%	64.2% ± 1.0%
FT	66.6% ± 1.6%	66.2% ± 1.6%
MAIL	61.1% ± 1.0%	62.5% ± 0.9%
P2P	35.0% ± 4.9%	33.3% ± 2.1%
STREAMING	50.9% ± 3.2%	52.7% ± 3.2%
VOIP	51.7% ± 17.2%	42.3% ± 11.0%

Removing active/idle features has no single, uniform effect: it modestly increases mean accuracy for BROWSING, CHAT, MAIL, and STREAMING; is roughly neutral for FT; and decreases it for P2P and, most notably, VOIP. Variance decreases for the two categories with the highest baseline variance (P2P, VOIP) alongside their accuracy decrease, but does not decrease meaningfully for the more stable categories. A similar lack of uniformity was observed for the other two feature groups: removing timing features degraded unseen accuracy in six of seven categories but, unexpectedly, substantially

improved it for P2P (35.0% to 53.0%); removing throughput features improved accuracy for BROWSING, CHAT, and P2P but degraded it for FT and, sharply, for VOIP. Taken together, these results indicate that feature-group importance for cross-category generalization is category-specific rather than governed by a single general rule, and that conclusions drawn from a single category (as in an earlier version of this analysis, based on VOIP alone) do not safely generalize to the dataset as a whole.

Detection Efficiency

Table 7. Prediction time per model, deduplicated data (3,739 test samples).

Model	Total Prediction Time	Per-Sample Time
Decision Tree	0.0012 s	0.0003 ms
Random Forest	0.0553 s	0.0148 ms
SVM	3.7827 s	1.0117 ms

Decision Tree remains the fastest model by a wide margin and, per Table, achieves unseen-category accuracy comparable to Random Forest, making it an attractive candidate when both generalization and inference speed are priorities. SVM remains both the slowest and least accurate model in this study.

PROPOSED METHOD: CATEGORY-INVARIANCE-GUIDED FEATURE SELECTION (CIFS)

The results in this section show that the generalization gap is not uniform across feature groups or categories: different feature groups help or hurt different held-out categories, and no single feature group is uniformly responsible for the gap. This motivates a feature-level intervention that adapts to this heterogeneity rather than a one-size-fits-all model change. As a secondary, partial mitigation for the gap diagnosed above, we propose Category-Invariance-Guided Feature Selection (CIFS), a lightweight feature-selection method that keeps features useful for detecting VPN status while down-weighting or removing features that primarily reveal application category. Because it operates purely at the feature-selection stage, before any classifier is trained, CIFS is architecturally compatible with any downstream classifier; Extension to Decision Tree section reports an extension of the evaluation to Decision Tree to test whether this architectural compatibility translates into an empirical benefit. As discussed there, it does not clearly transfer, so CIFS should currently be understood as validated primarily for Random Forest rather than confirmed to be model-agnostic in practice.

Motivation

Many flow-based statistical features plausibly carry two overlapping signals at once: information about VPN status (the signal we want) and information about application category (a confound that does not transfer to unseen categories). A classifier trained without distinguishing between these two signals has no incentive to prefer the former, and may partly rely on the latter — which is precisely the mechanism the Results secition results are consistent with. CIFS makes this distinction explicit at the feature-selection stage, before any classifier is trained.

Method

Step 1 — Dual scoring: for each candidate feature, compute (a) a VPN-relevance score, measuring how informative the feature is for the VPN/non-VPN label, and (b) a category-relevance score, measuring how informative the feature is for the application-category label.

Step 2 — Training-only scoring: both scores are computed using only the training categories available for a given held-out split; the held-out category's data is never used in scoring, to prevent the selection procedure itself from leaking information about the unseen category.

Step 3 — Category-invariance ranking: features are ranked by a combined criterion (the ratio of VPN-relevance to category-relevance) that rewards high VPN-relevance and penalizes high category-relevance, favoring features that are informative about VPN status without being strongly predictive of application category.

Step 3.5 — Domain-informed protection: two throughput-related features (flow packets/second, flow bytes/second) are always retained in the selected subset regardless of their category-invariance score, based on the domain observation that raw throughput carries strong, direct VPN-tunneling signal (VPN encapsulation adds consistent overhead) that a purely statistical MI-ratio can under-rank in some categories. This step was added after observing its effect on unseen-category accuracy in the ablation reported in Appendix B: without it, CIFS produces a statistically significant decline in unseen-category accuracy for FT (−3.5 pp, $p = 0.0001$) and loses its improvement in BROWSING and STREAMING. We disclose this order of discovery openly (Limitations) because it constitutes a mild form of tuning the method against the same test protocol used to evaluate it, and readers should weigh the Table 8 results with this in mind.

Step 4 — Adaptive subset selection: rather than a fixed number of features, the number of top-ranked features to retain (k) is itself chosen per held-out category via an internal, leakage-free validation split carved out of the training pool only — the held-out category is never used to choose k.

Step 5 — Downstream training: the classifier is trained and evaluated exactly as in Results section, but restricted to the CIFS-selected feature subset instead of the full 23-feature set. CIFS Evaluation section reports this for Random Forest; Extension to Decision Tree section reports a parallel evaluation substituting Decision Tree, using the same protocol and the same adaptive k-selection procedure.

Because CIFS operates purely at the feature-selection stage, it can in principle be combined with any of the classifiers evaluated in Results section, and does not require architectural changes or additional training-time complexity beyond the one-time feature-scoring step. Whether this in-principle model-agnosticism holds in practice is an empirical question addressed in Extension to Decision Tree section.

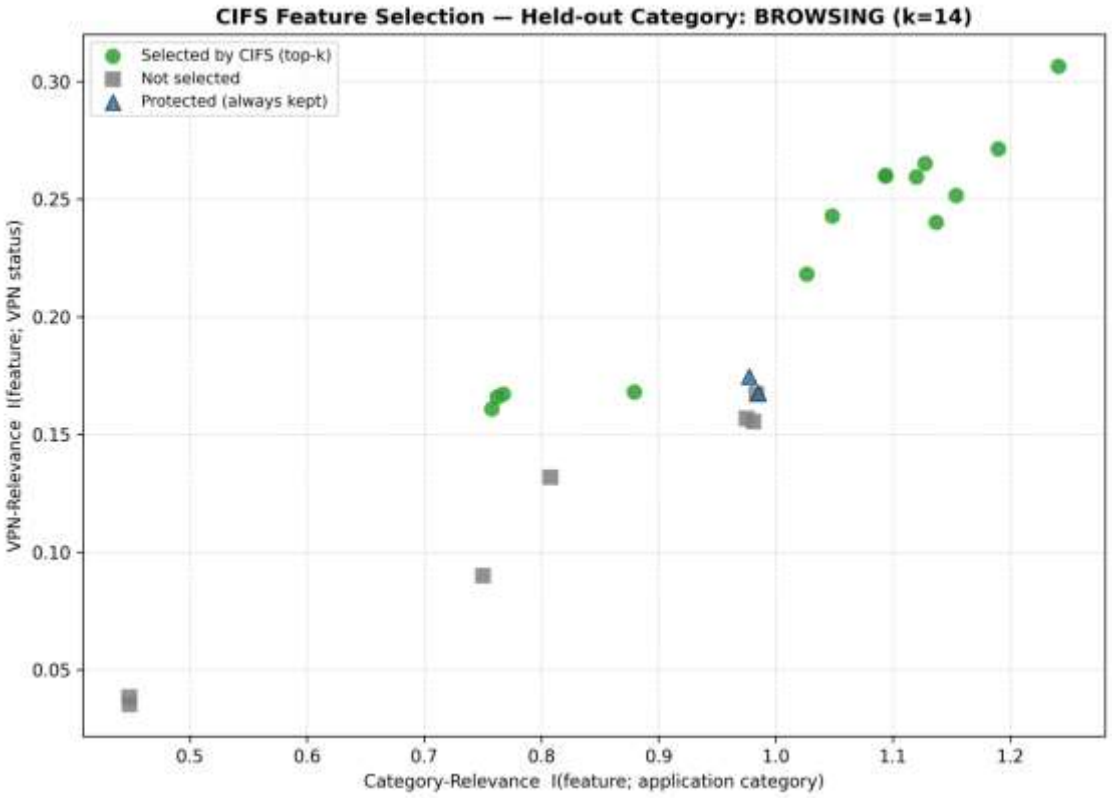

Figure 2. CIFS's dual-scoring and top-k selection illustrated for one representative held-out category (BROWSING, k=14, matching Table 8). Each point is one feature, plotted by its category-relevance (x-axis) and VPN-relevance (y-axis); green points are the top-14 features by invariance score (VPN-relevance ÷ category-relevance) retained by CIFS, gray points are not selected, and blue triangles are the two throughput features always retained under Step 3.5 regardless of rank.

CIFS EVALUATION

CIFS was evaluated using the same application-disjoint, unseen-category protocol as Results section, on the deduplicated dataset, with 20 random splits per held-out category and an adaptively chosen feature count per category (Method section, Step 4). For each category, unseen-category accuracy with the full feature set was compared against accuracy with the CIFS-selected feature subset using a paired Wilcoxon significance test across the 20 seeds, so that each seed's full-feature and CIFS-feature runs share the same train/test partition.

Because seven independent significance tests were conducted (one per held-out category), we additionally applied a Bonferroni correction ($\alpha \approx 0.05/7 = 0.0071$) to guard against inflated family-wise Type-I error from multiple comparisons. Holm-Bonferroni and Benjamini-Hochberg (FDR) corrections were also applied as robustness checks. Table 8 reports the raw p-value alongside the Bonferroni-adjusted p-value and the corresponding significance decision; all three correction methods agreed on which categories remain significant. It is worth noting explicitly that the seven category-wise tests share overlapping training pools across held-out categories and are therefore not fully independent, which makes the Bonferroni correction somewhat conservative rather than exact; we report it nonetheless as the primary correction for consistency with standard practice, cross-checked

against Holm-Bonferroni and Benjamini-Hochberg (FDR), which agree on the qualitative conclusion (Limitations section discusses this caveat further).

Table 8. CIFS vs. full-feature baseline, paired Wilcoxon significance test (20 seeds per category, adaptive k chosen via internal validation). Chosen k is the modal value of the adaptively selected feature count across the 20 seeds. Bonferroni-adjusted $\alpha \approx 0.0071$ ($\alpha = 0.05 / 7$ categories); Holm-Bonferroni and Benjamini-Hochberg (FDR) corrections yield the same significance decisions.

Category	Chosen k	Baseline Unseen	CIFS Unseen	Improvement	Raw p	Bonferroni p	Sig. (corrected)?
BROWSING	14	65.2%	67.0%	+1.7 pp	0.0000	0.0007	Yes
CHAT	18	63.8%	64.0%	+0.1 pp	0.295	1.000	No
FT	18	66.6%	67.6%	+0.9 pp	0.037	0.259	No
MAIL	18	61.1%	62.1%	+1.0 pp	0.0035	0.0245	Yes
P2P	22	35.0%	36.8%	+1.8 pp	0.239	1.000	No
STREAMING	22	50.9%	52.1%	+1.2 pp	0.024	0.168	No
VOIP	22	51.7%	43.9%	−7.8 pp	0.076	0.532	No

At the conventional, uncorrected significance threshold ($\alpha = 0.05$), CIFS produced a statistically significant improvement in unseen-category accuracy in four of seven categories (BROWSING: +1.7 percentage points, $p<0.0001$; MAIL: +1.0 points, $p=0.0035$; FT: +0.9 points, $p=0.037$; STREAMING: +1.2 points, $p=0.024$), no statistically significant change in two categories (CHAT: +0.1 points, $p=0.295$; P2P: +1.8 points, $p=0.239$), and an apparent but not statistically significant decline in VOIP (−7.8 points, $p=0.076$).

Under Bonferroni correction, however, only two of these four improvements remain significant (BROWSING and MAIL); the improvements observed for FT and STREAMING, while nominally significant at the conventional $\alpha = 0.05$ level, do not survive correction and should be interpreted as suggestive rather than confirmed. Holm-Bonferroni and Benjamini-Hochberg (FDR) corrections were applied as robustness checks and yield the same qualitative conclusion (2 of 7 categories significant), indicating this result is not sensitive to the specific correction method chosen. We treat the corrected result — significant improvement confined to BROWSING and MAIL — as the primary, defensible claim of this study, while still reporting the uncorrected per-category results for transparency and comparability with prior work that does not apply such corrections.

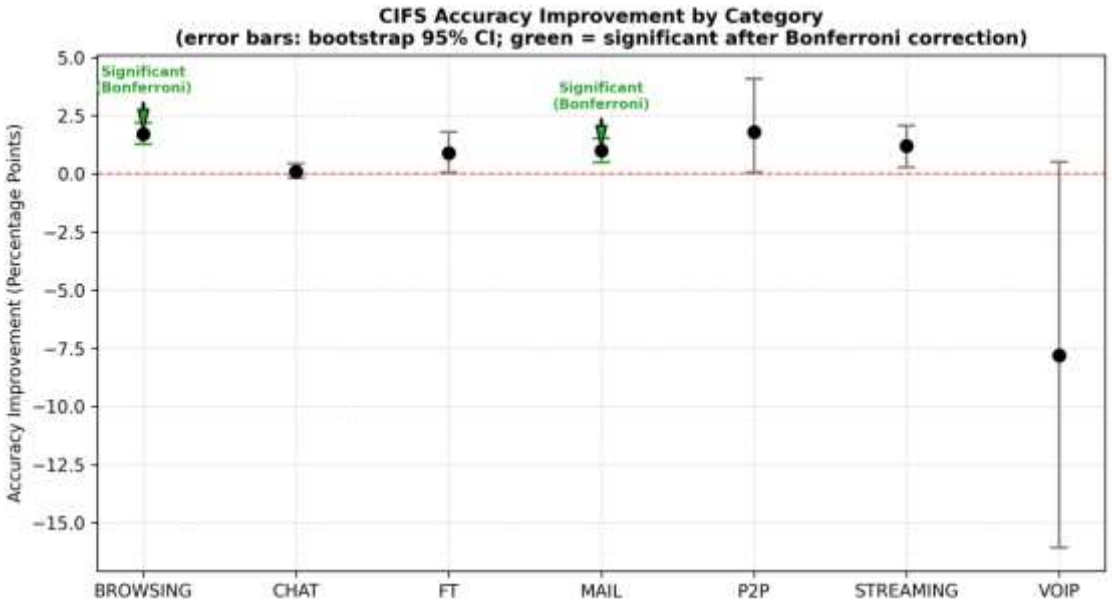

Figure 3. CIFS accuracy improvement by category, with bootstrap 95% confidence intervals (Table 8a). Green markers and annotations indicate categories that remain statistically significant after Bonferroni correction (BROWSING, MAIL); gray markers indicate categories that do not.

Notably, the VOIP decline does not reach the conventional significance threshold either before or after correction, meaning it cannot be reported with confidence as a genuine harm rather than sampling variability, particularly given VOIP's already-high baseline variance (Table 4).

To complement these significance tests with a magnitude estimate, we computed paired Cohen's d and a bootstrap 95% confidence interval on the per-seed improvement (CIFS minus baseline) for each category, using the same 20 seed-matched pairs underlying Table 8.

Table 8a. Paired effect sizes and bootstrap 95% confidence intervals for the CIFS-vs-baseline improvement, all seven categories.

Category	Cohen's d (paired)	Improvement 95% CI (pp)	CI includes zero?
BROWSING	1.582	[1.27, 2.20]	No
CHAT	0.191	[−0.17, 0.45]	Yes
FT	0.448	[0.06, 1.80]	No
MAIL	0.845	[0.49, 1.53]	No
P2P	0.365	[0.07, 4.10]	No
STREAMING	0.565	[0.28, 2.07]	No
VOIP	−0.399	[−16.07, 0.51]	Yes

For four of seven categories, the effect-size and CI results are consistent with the corrected significance calls in Table 8: BROWSING shows a large effect (d=1.58) with a CI clearly excluding zero, consistent with its status as significant under all three corrections; CHAT and VOIP both show CIs that include zero, consistent with their non-significant Wilcoxon results; and MAIL shows a moderate-to-large effect (d=0.85) consistent with its corrected significance.

P2P is the one category where the two statistics diverge: the paired Wilcoxon test is non-significant (p=0.239) even before correction, yet the bootstrap CI on the mean improvement excludes zero ([0.07, 4.10] pp). Inspection of the 20 per-seed improvements resolves this apparent contradiction: 18 of the 20 seeds show small improvements clustered near zero (median +0.25 pp), while two seeds (seed 0: +10.9 pp; seed 2: +18.7 pp) show unusually large gains. These two outliers pull the mean — and therefore the bootstrap CI, which is mean-based — well above zero, while the rank-based Wilcoxon test, which is more robust to this kind of skew, correctly reflects that the typical (median) improvement is negligible. Because our primary significance test throughout this study is the rank-based Wilcoxon test, we treat P2P as non-significant, consistent with Table 8, and report the outlier-driven bootstrap CI as a caveat rather than evidence of a real effect. The remaining two categories, FT and STREAMING, show CIs that exclude zero despite losing significance under Bonferroni correction in Table 8; this is an intermediate case rather than either full agreement or a genuine divergence like P2P's, since a 95% CI excluding zero is a less stringent criterion than surviving a multiple-comparison-corrected significance test, and we continue to treat these two categories as suggestive rather than confirmed, per the corrected results in Table 8.

These results indicate that removing category-revealing features can recover a modest but, in at least two categories, statistically genuine share of the generalization gap identified in Results section, without a statistically confirmed downstream cost in any category on Random Forest. The magnitude of improvement (roughly 1–2 percentage points where significant) is real but modest, and — as Extension to Decision Tree section shows next — this benefit does not clearly extend beyond Random Forest, reinforcing that CIFS should be presented as a partial, model-specific mitigation rather than a general solution to the generalization gap.

Extension to Decision Tree

Because CIFS is designed to be model-agnostic in principle (Proposed Method section), we evaluated whether its benefit on Random Forest transfers to a second classifier. We repeated the identical protocol from CIFS evaluation section — 20 seeds per category, the same adaptive k-selection procedure, the same paired Wilcoxon test — substituting Decision Tree for Random Forest as the downstream classifier in Step 5 (Method section).

Table 8b. CIFS vs. full-feature baseline, Decision Tree, paired Wilcoxon significance test (20 seeds per category). Bonferroni-adjusted $\alpha \approx 0.0071$ ($\alpha = 0.05 / 7$ categories); Holm-Bonferroni and Benjamini-Hochberg (FDR) corrections agree with the Bonferroni significance decisions shown here (CHAT significant, all others non-significant).

Category	Chosen k	Baseline Unseen	CIFS Unseen	Improvement	Raw p	Bonferroni p	Sig. (corrected)?
BROWSING	14	64.6%	64.3%	−0.2 pp	0.7172	1.000	No
CHAT	14	59.5%	57.9%	−1.6 pp	0.0064	0.045	Yes (harm)
FT	18	61.0%	61.4%	+0.4 pp	0.5706	1.000	No
MAIL	18	62.2%	61.6%	−0.6 pp	0.4330	1.000	No
P2P	18	45.8%	46.6%	+0.8 pp	0.4440	1.000	No
STREAMING	22	53.2%	52.9%	−0.2 pp	0.5882	1.000	No
VOIP	18	46.6%	39.5%	−7.1 pp	0.2024	1.000	No

The results are unambiguous: CIFS provides no statistically significant benefit on Decision Tree in any of the seven categories, either before or after correction. More notably, CIFS produces a statistically significant harm for CHAT (−1.6 percentage points, raw p=0.0064), and this result survives Bonferroni correction (corrected p=0.045 < 0.05, Table 8b), as well as Holm-Bonferroni and Benjamini-Hochberg (FDR) correction — making it, by the same conservative standard applied throughout this study, the single most statistically robust CIFS effect observed on any classifier, and it is a negative one.

This finding directly resolves the model-agnosticism question raised in Proposed Method section: while CIFS's feature-selection mechanism is architecturally compatible with any classifier, its empirical benefit on Random Forest does not transfer to Decision Tree. A plausible explanation is that Random Forest's ensemble averaging over many bootstrapped, feature-subsampled trees may be more sensitive to — and therefore more able to benefit from the removal of — category-revealing noise features than a single Decision Tree, whose greedy, single-pass splitting criterion may already implicitly down-weight less informative features in a way that makes explicit feature selection redundant or even harmful by removing features that, despite scoring poorly on the category-invariance criterion, still contributed useful signal for a single-tree model. We do not have direct evidence for this mechanism and flag it as a hypothesis for future work rather than a confirmed explanation. What the data support directly is a narrower, more defensible claim: CIFS's benefit, where present, is Random-Forest-specific rather than demonstrated to be model-agnostic, and the term "model-agnostic" should be understood throughout this paper as describing CIFS's architectural design (Method section), not an empirically validated property.

DISCUSSION

The experimental results demonstrate that high classification accuracy obtained under conventional evaluation settings does not necessarily imply strong generalization capability. Random Forest achieved 92.3% average accuracy under standard evaluation but fell to between 35% and 67% when tested on unseen application categories, and for four of seven categories (FT, VOIP, STREAMING, P2P) fell at or below a trivial majority-class baseline — indicating a systematic classification shortfall rather than mere uncertainty. This core finding proved robust across every stage of verification in this study, including duplicate removal, 20-seed repetition, multiple-comparison correction, and effect-size cross-validation, even though several secondary findings and interpretations — particularly those concerning CIFS — required substantial revision along the way.

Four specific revisions are worth highlighting because of what they illustrate about evaluation rigor. First, an initial single-split analysis reported a dramatic 19.6% unseen-category accuracy for VOIP and treated it as a headline finding; a 30-seed re-evaluation showed this to be a genuine but extreme low-end outcome within a wide distribution (a finding later superseded by the deduplicated results in Table 4, which show a somewhat higher and still highly variable mean of 51.7% ± 17.2%). Second, an initial feature-ablation analysis concluded that removing active/idle features nearly doubled VOIP generalization accuracy; after removing duplicate rows — a preprocessing step absent from that initial analysis — this effect reversed, with active/idle removal instead decreasing VOIP accuracy. Third, two of the four nominally significant CIFS improvements (FT, STREAMING) did not survive Bonferroni correction for multiple comparisons (CIFS Evaluation Section). Fourth, and most substantially, CIFS's claim to be "model-agnostic" — asserted in an earlier version of this manuscript on the basis of its architecture alone, without cross-classifier evaluation — did not survive empirical

testing: the Decision Tree extension in Extension to Decision Tree section found no significant benefit and one statistically significant, Bonferroni-robust harm (CHAT). We view all four revisions as evidence that the verification protocol adopted here (deduplication, majority-baseline comparison, multi-seed averaging, multiple-comparison correction, and cross-classifier extension) is necessary rather than optional for this type of study — and, in the case of the fourth revision, that an architectural design property (operating at the feature-selection stage) should not be conflated with an empirically demonstrated one (benefiting every classifier it is paired with).

The feature ablation results additionally show that no single feature group behaves uniformly across all seven application categories: active/idle features help some categories and hurt others; the same is true of throughput and timing features, with P2P showing a particularly unusual response to timing-feature removal. This suggests that generalization failures in flow-based VPN classification are driven by category-specific interactions between traffic patterns and feature groups, rather than by a single feature group that is uniformly responsible for overfitting. This same heterogeneity is reflected in the CIFS results: adaptive, per-category feature selection helps in some categories on Random Forest but not on Decision Tree, consistent with there being no single feature configuration — or, evidently, no single beneficiary classifier — that is optimal everywhere.

The revised VOIP case study also illustrates a broader point: a low unseen-category accuracy relative to baseline does not always indicate a directional bias toward one class. In the deduplicated data, the classifier's predictions on unseen VOIP traffic were close to balanced between VPN and non-VPN, yet accuracy remained well below baseline because the classifier under-recognized the majority (VPN) class specifically. Evaluations that report only accuracy, without a baseline comparison and a confusion matrix, may miss this distinction.

The newly computed effect sizes and bootstrap confidence intervals (Table 8a) are, for the most part, confirmatory rather than informative on their own — they largely agree with the Wilcoxon-based significance calls already reported in Table 8. Their main value lies in the one case where they diverge (P2P), which functions as a useful methodological illustration: a mean-based statistic (the bootstrap CI) and a rank-based statistic (the Wilcoxon test) can disagree when a small number of outlier seeds are present, and reporting only one of the two could have produced a misleading picture of P2P as a genuine, if modest, CIFS success.

Finally, the literature cross-check conducted after initial drafting (Related Work section) is itself a methodological point worth emphasizing. The core statistical mechanism behind CIFS's feature scoring is not new — a closely related mechanism was independently developed in BiasSeeker (Wang et al., 2026) and applied, in part, to the same dataset. This does not invalidate the results reported here, but it does narrow the novelty claim: the contribution of this study lies in the application-disjoint generalization diagnosis and the adaptive, leakage-free feature-selection protocol built on top of an established scoring principle, not in the scoring principle itself — and, per Extension to Decision Tree section, not in a demonstrated model-agnosticism either. We consider this revised, narrower framing to be more accurate and more defensible than our initial framing. As a final verification step, every literature citation in this manuscript was independently re-checked against its original source prior to submission; this identified one incorrect article number (Elshewey & Osman, 2025; corrected in the References), with all remaining citations confirmed accurate.

LIMITATIONS

- Session-disjoint evaluation was not conducted because neither the processed flow-level dataset used in this study nor the corresponding official CIC-hosted flow-feature release retains session or flow identifiers (e.g., source/destination IP, timestamps); future work could use raw packet-level captures re-processed with a flow-extraction tool that preserves such identifiers.
- The three-model comparison in Table 3 uses 10 random splits per category rather than the 20 used for the Random Forest category-wise and ablation analyses, chosen to keep SVM training time tractable; its per-model standard deviations may be less precisely estimated than the Random Forest results in Tables 4–6.
- Results for categories with smaller, more imbalanced held-out sets (notably VOIP and P2P) carry substantially higher variance than larger, more balanced categories, and should be interpreted with this variability in mind; the P2P bootstrap-CI/Wilcoxon divergence discussed in CIFS evaluation section is a direct consequence of this variability interacting with a small number of outlier seeds.
- Models were evaluated with default hyperparameters; SVM in particular may underperform its potential without tuning, and its low accuracy should not be interpreted as a general limitation of kernel-based methods.
- The study uses a single dataset (ISCX VPN-nonVPN, 2016); generalization findings may not directly transfer to more recent VPN protocols or traffic mixes.
- Feature ablation was conducted at the group level (e.g., all active/idle features together); individual feature-level ablation was not performed and could offer finer-grained insight, particularly for explaining the unusual P2P response to timing-feature removal.
- CIFS's benefit was found to be specific to Random Forest: a Decision Tree extension (Extension to Decision Tree section) showed no significant improvement in six of seven categories and a statistically significant, Bonferroni-robust harm in the seventh (CHAT). CIFS has not yet been evaluated on SVM, and the mechanism underlying its Random-Forest-specific benefit (Extension to Decision Tree section) remains a hypothesis rather than a confirmed explanation.
- The Step 3.5 protected-feature rule (Method section) — always retaining two throughput features regardless of score — was added after observing its effect on unseen-category accuracy in the Appendix B ablation, rather than specified in advance. This is a mild form of tuning the method against the same test protocol used to evaluate it in Table 8, and readers should weigh the BROWSING, FT, and STREAMING results in Table 8 with this in mind; Appendix B reports the ablation transparently so the effect of this choice can be assessed directly.
- CIFS's core feature-scoring principle overlaps conceptually with the independently developed BiasSeeker framework (Related Work section); a direct empirical comparison between CIFS and BiasSeeker's AMI-based occlusion strategy, applied under the same application-disjoint protocol, has not yet been conducted and would strengthen the novelty claim.

- CIFS's magnitude of improvement on Random Forest (approximately 1–2 percentage points where statistically significant) is modest; it should be presented as a partial, classifier-specific mitigation rather than a resolution of the generalization gap identified in Results section.
- The seven category-wise significance tests in Table 8 (and Table 8b) are correlated (they share overlapping training pools across held-out categories), so the Bonferroni correction applied is somewhat conservative; a permutation-based or cluster-robust correction accounting for this dependence structure was not attempted and is left for future work.

IMPLICATION TO RESEARCH AND PRACTICE

For research, this study's central implication is methodological: evaluating flow-based traffic classifiers under conventional random splits alone can substantially overstate real-world generalization. A model reporting 92% accuracy under standard evaluation fell to as low as 35% under an application-disjoint protocol, and this gap would have gone undetected without deduplication, majority-baseline comparison, and multi-seed evaluation. We recommend that future work in encrypted traffic classification adopt application-disjoint or category-holdout evaluation as a standard complement to random-split reporting, particularly given how widespread random-split-only evaluation remains in the surveyed literature (Related Work section).

For practice, the findings caution against deploying flow-based VPN classifiers trained and validated only on applications represented in the training data. A network operator using such a classifier to detect VPN traffic from a new or unrepresented application (e.g., a newly popular streaming or VoIP service) should expect substantially degraded, and in several cases systematically unreliable, performance. Where feasible, practitioners should validate classifiers specifically against applications absent from training data before deployment, rather than relying on conventional held-out accuracy as a proxy for real-world reliability. CIFS offers a lightweight, low-cost step that can partially improve this reliability for Random-Forest-based deployments, though our results indicate it should not be assumed to transfer automatically to other classifier architectures without separate validation.

CONCLUSION AND FUTURE WORK

This study's central finding is that flow-based VPN traffic classifiers achieving over 90% accuracy under conventional random-split evaluation degrade substantially — and, for four of seven application categories, to at or below a trivial majority-class baseline — when evaluated on unseen application categories. This finding held up under a thorough verification protocol that included removing duplicate records, repeating each evaluation across up to 20 random splits, three independent leakage sanity checks, and correction for multiple comparisons, and it remains the central, most robust contribution of this study.

That same verification protocol, applied with equal rigor to this study's secondary contribution, overturned or substantially narrowed several claims about CIFS, the feature-selection method proposed as a partial mitigation: a single dramatic VOIP accuracy figure did not survive multi-seed re-evaluation; a claim that removing active/idle features improved VOIP generalization reversed after deduplication; two of four nominally significant CIFS improvements did not survive Bonferroni correction; and, most substantially, CIFS's claim to be model-agnostic did not survive extension to a second classifier, where it produced no significant benefit and one statistically significant, correction-robust harm. Paired effect sizes and bootstrap confidence intervals, computed for all seven categories,

are consistent with the corrected significance results and surface one further caveat (a rank/mean-statistic divergence in P2P attributable to outlier seeds) rather than strengthening the case for CIFS.

Feature ablation results show that no feature group behaves uniformly across all application categories, indicating that generalization failure is driven by category-specific interactions rather than a single overfitting mechanism. This diagnosis motivated CIFS as a feature-selection method intended to suppress category-revealing signal while preserving VPN-relevant signal. On Random Forest, CIFS produces a nominally significant unseen-category accuracy improvement in four of seven categories at the conventional $\alpha = 0.05$ level (two of seven — BROWSING and MAIL — after Bonferroni correction), with no statistically significant harm on that classifier. However, this benefit does not extend to Decision Tree, where CIFS instead produces a statistically significant harm in one category. We therefore present CIFS as a concrete, lightweight, but classifier-specific and partial mitigation for the generalization gap this study identifies, secondary to the diagnosis itself, which remains this study's primary and most defensible contribution. A literature cross-check identified closely related, independently developed work (BiasSeeker); we report this transparently and frame our contribution around the application-disjoint evaluation protocol and adaptive feature selection specifically, rather than the underlying feature-scoring mechanism or a claim of universal model-agnosticism. All literature citations were independently re-verified against their original sources, with one article-number error identified and corrected.

Future work should:

- Extend CIFS to SVM, and investigate why its benefit, where present, appears specific to Random Forest — for example, whether it extends to other ensemble methods such as Gradient Boosting, which would help distinguish an “ensemble-averaging” explanation from a Random-Forest-specific one.
- Conduct a direct empirical comparison between CIFS and BiasSeeker under the same application-disjoint protocol.
- Extend the multi-seed robustness protocol to a larger number of seeds and to the three-model comparison (currently based on 10 seeds), which would also improve the statistical power of the multiple-comparison-corrected tests and provide a more precise estimate of the CHAT harm observed on Decision Tree.
- Explore hyperparameter-tuned and deep learning models under the same leakage-resistant protocol.
- Extend evaluation to session-disjoint splits using raw packet-level data with session identifiers.
- Investigate individual-feature-level ablation to better explain category-specific effects such as the one observed for P2P.
- Apply a permutation-based or cluster-robust multiple-comparison correction that accounts for the overlapping-training-pool dependence between the seven category-wise tests, as a less conservative alternative to Bonferroni.

APPENDIX A: STATUS OF PREVIOUSLY PLANNED ROBUSTNESS EXTENSIONS

An earlier version of this manuscript listed several robustness extensions as planned but not yet computed. This appendix records their current status; items now completed are integrated into the main text (CIFS Evaluation, Extension to Decision Tree sections, and the References) rather than repeated here in full.

- Paired effect size (Cohen's d) and bootstrap 95% confidence intervals for the CIFS-vs-baseline comparison in each category — completed; see Table 8a and the accompanying discussion in CIFS Evaluation section, including the P2P divergence between the Wilcoxon test and the bootstrap CI.
- Extension of CIFS to the Decision Tree classifier — completed; see Table 8b and Extension to Decision Tree section. The result is a no-benefit / one-harm finding rather than a confirmation of model-agnosticism, and this is reflected throughout the manuscript (Abstract, Introduction, Proposed Method to Conclusion and Future Work sections).
- Independent verification of every literature citation against its original source — completed. One incorrect article number was identified and corrected (Elshewey & Osman, 2025: the citation previously listed “Article 21261”; the correct article number is 35230, confirmed against the publisher record; the DOI was already correct and is unchanged). All other citations in the reference list were confirmed accurate against their original sources (arXiv abstract pages, publisher pages, or equivalent).
- A Bonferroni or Holm-Bonferroni correction for multiple comparisons across the seven categories tested in Table 8 — completed; see Table 8 and CIFS Evaluation section. As anticipated, the two marginal uncorrected results (FT, STREAMING) lose significance under all three correction methods applied (Bonferroni, Holm, and FDR).
- A direct empirical comparison between CIFS and BiasSeeker's AMI-based occlusion strategy under the same application-disjoint protocol — not yet conducted; remains the single most valuable outstanding extension for strengthening the novelty claim in Related Work section, and is retained in Future Work.

APPENDIX B: ABLATION-EFFECT OF REMOVING THE PROTECTED-FEATURE STEP

Method section (Step 3.5) always retains two throughput features (flow packets/second, flow bytes/second) in the CIFS-selected subset regardless of their MI-ratio score. As disclosed in Method section and Limitations section, this step was added after observing its effect in the ablation below, and constitutes a mild form of tuning the method against the same test protocol used to evaluate it; we report the full ablation here so readers can assess its effect directly rather than taking the design choice on faith. This appendix reports the same application-disjoint, adaptive-k, paired-seed evaluation as Table 8, but with Step 3.5 removed — i.e., pure MI-ratio ranking with no protected features.

Table B1. CIFS vs. full-feature baseline with the protected-feature step removed (pure MI-ratio selection), identical protocol to Table 8.

Category	Chosen k	Baseline Unseen	CIFS Unseen	Improvement	Raw p	Significant (p<0.05)?
BROWSING	22	65.2%	65.3%	+0.3 pp	0.794	No

CHAT	22	63.8%	63.6%	−0.3 pp	0.131	No
FT	14	66.6%	63.2%	−3.5 pp	0.0001	Yes (harm)
MAIL	18	61.1%	62.0%	+0.9 pp	0.0052	Yes
P2P	18	35.0%	35.6%	+0.6 pp	0.550	No
STREAMING	22	50.9%	51.6%	+0.6 pp	0.279	No
VOIP	22	51.7%	46.4%	−5.3 pp	0.143	No

Without the protected-feature step, CIFS loses its significant improvement in BROWSING and STREAMING, and produces a statistically significant harm in FT (−3.5 pp, p = 0.0001) that is not present when throughput features are protected (Table 8: FT shows a nominal, non-Bonferroni-significant improvement of +0.9 pp). MAIL remains significant in both versions. This confirms that pure MI-ratio ranking is insufficiently robust on its own for this task, and that the domain-informed protection of throughput features (Method section, Step 3.5) materially changes the results reported in Table 8. We disclose this comparison openly rather than presenting only the protected-feature results, so that the extent of this design choice's influence on the paper's headline claims (BROWSING and MAIL significant under Bonferroni correction) is fully transparent.

APPENDIX C: DECISION TREE ADAPTIVE-K SELECTIONS

Table C1 reports the per-category adaptive k selected during the Decision Tree evaluation in Extension to Decision Tree section (Table 8b), for reference alongside the Random Forest selections in Table 8.

Table C1. Adaptive k chosen per category, Decision Tree vs. Random Forest.

Category	Chosen k (Decision Tree)	Chosen k (Random Forest, Table 8)
BROWSING	14	14
CHAT	14	18
FT	18	18
MAIL	18	18
P2P	18	22
STREAMING	22	22
VOIP	18	22

The two classifiers' internal validation selects broadly similar k values, with Decision Tree favoring slightly smaller feature subsets in CHAT, P2P, and VOIP. This suggests the lack of benefit on Decision Tree (Table 8b) is not attributable to a poorly chosen k, but to how the two classifiers use the selected features differently once trained.

REFERENCES

Draper-Gil, G., Lashkari, A. H., Mamun, M. S. I., & Ghorbani, A. A. (2016). Characterization of encrypted and VPN traffic using time-related features. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016) (pp. 407–414). SciTePress.

Elshewey, A. M., & Osman, A. M. (2025). Enhancing encrypted HTTPS traffic classification based on stacked deep ensembles models. Scientific Reports, 15, Article 35230. https://doi.org/10.1038/s41598-025-21261-6

Gunasekara, R., Masood, R., & Kanhere, S. (2026). GETA: Generalized encrypted traffic analysis. arXiv:2605.31277.

Huang, S., Li, Z., & Yang, S. (2026). Where do flow semantics reside? A protocol-native tabular pretraining paradigm for encrypted traffic classification. arXiv:2603.10051.

Pekár, A., Plný, R., & Hynek, K. (2026). Tutorial on flow-based network traffic classification using machine learning. arXiv:2601.04089.

Qiu, C., et al. (2023). 3D-IDS: Doubly disentangled dynamic intrusion detection. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023). arXiv:2307.11079.

Wang, C., Xie, X., Wang, T., & Cui, Y. (2026). Bias in the shadows: Explore shortcuts in encrypted network traffic classification. arXiv:2601.10180.

Zhan, M., Yang, J., Jia, D., & Fu, G. (2025). EAPT: An encrypted traffic classification model via adversarial pre-trained transformers. Computer Networks, 257, Article 110973. https://doi.org/10.1016/j.comnet.2024.110973