Linking Class-Wise SHAP Attribution Drift To Detection Degradation Under Cross-Dataset Shift In Network Intrusion Detection

Aditya Rameshwar Karmase[1], Prof. Smt. K.N. Gadakh[2]
[1]*Department of Computer Science and Applications, K.R.T. Arts, B.H. Commerce and A.M. Science (KTHM) College, Nashik, Maharashtra, India*
[2]*Asst. Prof., Department of Computer Science and Applications, K.R.T. Arts, B.H. Commerce and A.M. Science (KTHM) College, Nashik, Maharashtra, India*
doi.org/10.64643/IJIRT211653

***Abstract*—Machine-learning-based network intrusion detection systems (NIDS) can achieve high classification performance on benchmark datasets but may fail to generalize to different network environments. Explainable artificial intelligence techniques such as SHapley Additive exPlanations (SHAP) provide insight into model feature importance, but the stability of these explanations does not necessarily indicate successful detection transfer. This study experimentally investigates the relationship between class-wise SHAP attribution drift and detection degradation under cross-dataset distribution shift. Random Forest and XGBoost classifiers were trained and evaluated bidirectionally using selected captures from CIC-IDS2017 and CSE-CIC-IDS2018. Seven harmonized traffic classes and 23 common features were retained, with 5,000 samples per class in each prepared dataset. Class-wise SHAP attribution rankings were compared using Spearman correlation and Top-10 Jaccard similarity. Detection degradation was measured through changes in class-wise recall, while exact permutation tests and Holm correction were used to examine associations between attribution drift and recall degradation. Cross-dataset Macro F1 declined to 0.2500–0.2780, compared with source-holdout scores of 0.8864–0.9695. Despite this decline, mean class-wise Spearman correlations ranged from 0.8226 to 0.9366 for Random Forest and from 0.8602 to 0.9286 for XGBoost, indicating relatively stable feature-attribution rankings despite substantial cross-dataset detection performance degradation. Several classes exhibited high attribution-ranking similarity despite near-zero target recall. The association between attribution drift and recall degradation varied across models and transfer directions, with no association remaining statistically significant after multiple-testing correction. These findings demonstrate that global SHAP attribution-ranking stability and cross-dataset detection performance should be evaluated as distinct properties of intrusion detection models.**

***Index Terms*—Network Intrusion Detection, Explainable Artificial Intelligence, SHAP, Cross-Dataset Generalization, Distribution Shift, Random Forest, XGBoost.**

I. INTRODUCTION

Machine-learning-based network intrusion detection systems are increasingly used to identify malicious network activity by learning patterns from network traffic. However, strong performance on a particular dataset does not guarantee that a trained model will perform effectively in another network environment. Differences in traffic characteristics, attack implementations, feature distributions and data collection procedures can cause substantial performance degradation.

Cross-dataset generalization is therefore an important challenge in NIDS research. Cantone et al. [1] demonstrated that classifiers achieving high performance within a dataset can experience substantial degradation when evaluated across datasets. This raises concerns about the practical reliability of intrusion detection models trained and tested using a single benchmark.

Explainable artificial intelligence (XAI) provides additional insight into model behaviour. SHAP, introduced by Lundberg and Lee [2], assigns feature attribution values to model predictions using a unified additive explanation framework. In intrusion detection, SHAP can help identify which network

traffic characteristics contribute to particular class predictions.

However, an important distinction remains: stable feature-attribution rankings do not necessarily mean that a model can successfully detect the corresponding attack in another dataset. A model may preserve a similar ordering of influential features while its predictive performance deteriorates.

This study investigates that distinction using two publicly available network intrusion detection datasets, CIC-IDS2017 and CSE-CIC-IDS2018 [3, 4].

The research addresses three questions:

1. How substantially do Random Forest and XGBoost detection performances deteriorate under bidirectional cross-dataset evaluation?
2. How stable are class-wise SHAP feature-attribution rankings across the two datasets?
3. Is class-wise SHAP attribution drift associated with class-wise recall degradation?

The study contributes an experimental comparison of detection performance and SHAP attribution-ranking stability across seven harmonized traffic classes. It also examines the association between attribution drift and recall degradation using exact permutation testing with multiple-testing correction. The contribution is deliberately limited to these experimental conditions rather than claiming that attribution drift universally predicts detection failure.

II. LITERATURE REVIEW

Machine-learning-based Network Intrusion Detection Systems (NIDS) have been widely studied for identifying malicious network traffic. The CIC-IDS2017 dataset, introduced by Sharafaldin et al. [3], provides labelled benign and attack traffic, including brute-force, denial-of-service, web, and other attack categories. It has become a commonly used benchmark for evaluating intrusion detection models. However, performance measured on a single benchmark does not necessarily demonstrate that a model will generalize to traffic collected under different network conditions.

Cross-dataset generalization is therefore an important concern. Cantone et al. [1] investigated machine-learning-based NIDS using four datasets, including CIC-IDS2017 and CSE-CIC-IDS2018. Their findings showed that models achieving strong results when trained and tested on the same dataset could experience substantial performance deterioration when evaluated across different datasets. This highlights the influence of dataset heterogeneity and the limitations of relying exclusively on within-dataset evaluation.

Alongside predictive performance, explainability has gained importance in cybersecurity because analysts need to understand the factors influencing model predictions. Lundberg and Lee [2] introduced SHAP (SHapley Additive exPlanations), a unified framework based on Shapley values that assigns feature contributions to model predictions. SHAP can be used to examine which network-flow features influence intrusion detection models. However, observing similar feature-importance rankings across datasets does not automatically establish that a model's explanations are faithful or that its predictions remain accurate [5].

These research areas motivate evaluating predictive performance and explanation behaviour together. Cross-dataset studies demonstrate the risk of detection degradation, while SHAP provides a method for examining model attribution patterns.

Investigating whether changes in class-wise SHAP feature rankings are associated with class-wise detection degradation can provide an additional perspective on model behaviour under dataset shift.

III. RESEARCH GAP AND OBJECTIVES

Existing research has examined cross-dataset generalization in network intrusion detection and the use of SHAP for interpreting machine-learning predictions. However, the relationship between class-wise SHAP attribution-ranking drift and class-wise detection degradation across different intrusion detection datasets warrants further empirical investigation.

This study aims to:

1. Evaluate Random Forest and XGBoost models using within-dataset and bidirectional cross-dataset experiments on CIC-IDS2017 and CSE-CIC-IDS2018.
2. Measure class-wise SHAP attribution-ranking similarity using Spearman correlation and Top-10 Jaccard similarity.

3. Examine the association between attribution drift and class-wise recall degradation.
4. Assess the sampling stability of the observed SHAP attribution rankings through repeated sampling.

The study investigates whether stable feature-attribution rankings can coexist with substantial detection degradation under cross-dataset shift.

IV. EXPERIMENTAL METHODOLOGY

4.1. Datasets and class selection

The experiments used CIC-IDS2017 and CSE-CIC-IDS2018. Rather than processing every capture file, three selected captures from each dataset were used.

For CIC-IDS2017, the selected files were:

- Monday-WorkingHours.pcap_ISCX.csv
- Tuesday-WorkingHours.pcap_ISCX.csv
- Wednesday-workingHours.pcap_ISCX.csv

For CSE-CIC-IDS2018, the selected files were:

- Wednesday-14-02-2018_TrafficForML_CICFlowMeter.csv
- Thursday-15-02-2018_TrafficForML_CICFlowMeter.csv
- Friday-16-02-2018_TrafficForML_CICFlowMeter.csv

Seven shared traffic classes were selected: Benign, FTP-BruteForce, SSH-Bruteforce, DoS-Hulk, DoS-GoldenEye, DoS-Slowloris and DoS-SlowHTTPTest. Dataset-specific label variants were mapped to common class names. Sparse attack categories without suitable representation in both selected datasets were excluded. The overall experimental workflow is summarized in Figure 1.

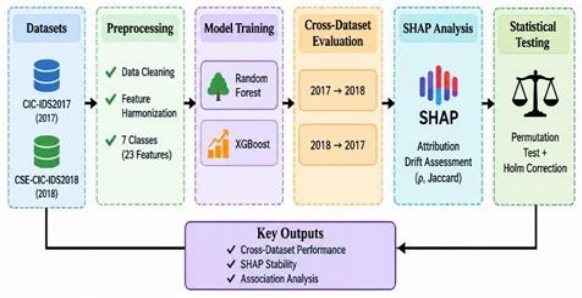

Figure 1. Experimental Workflow for Cross-Dataset XAI Analysis.

4.2. Feature harmonization and pre-processing

An exact-name comparison of the available CSV schemas identified 27 common candidate predictive features. Four features that contained only zero values in both prepared datasets were removed, leaving 23 retained features.

The retained features represented flow duration, forward and backward inter-arrival times, active and idle flow statistics, directional flags and the down/up ratio.

Pre-processing included numeric conversion, replacement of infinite values, and treatment of negative Flow IAT Min and Fwd IAT Min values as missing. Rows with all retained features missing were discarded.

A class-wise random-key reservoir sampling procedure selected 5,000 eligible records per class from each dataset, producing 35,000 records per prepared dataset.

This produced balanced experimental samples rather than samples representative of the original traffic prevalence.

The CIC-IDS2017 prepared sample contained 19 missing feature cells, while the CSE-CIC-IDS2018 sample contained none. Missing values were handled using median imputation, with the imputer fitted exclusively on the source training partition.

4.3. Model training and cross-dataset evaluation

Two classifiers were evaluated: Random Forest and XGBoost.

Random Forest used 300 estimators, square-root feature selection and random state 42. XGBoost used 300 estimators, maximum tree depth 6, learning rate 0.08, subsample 0.8, column subsampling 0.8 and the multiclass soft-probability objective.

For each source dataset, an 80:20 stratified split was used for training and source-holdout evaluation. The models were trained on the source training partition and evaluated on both the source holdout and the complete prepared target dataset.

The experiments were conducted in both directions:

- CIC-IDS2017 to CSE-CIC-IDS2018.
- CSE-CIC-IDS2018 to CIC-IDS2017.

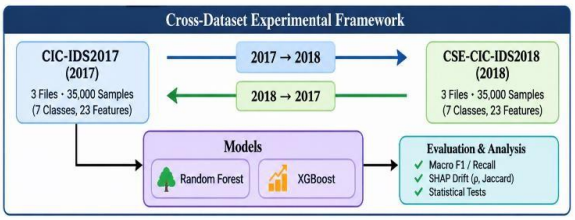

Figure 2. Cross-Dataset Experimental Framework.

Accuracy, macro precision, macro recall and Macro F1 were calculated. Macro F1 was emphasized because it gives equal weight to each of the seven classes.

4.4. SHAP attribution analysis

TreeExplainer was used to calculate SHAP values for both classifiers. For each class, mean absolute SHAP values were aggregated using the true-class output for records belonging to that class. This approach evaluates attribution rankings associated with the actual class, including records that the model may have misclassified.

For every class, source and target feature-importance rankings were compared using Spearman rank correlation.

Attribution drift was defined as:

Attribution drift = 1 − Spearman correlation

Top-10 Jaccard similarity was also calculated to measure overlap between the ten highest-ranked features in the source and target evaluations.

4.5. Repeated-sampling stability analysis

To examine sensitivity to sample selection, the SHAP comparison was repeated ten times. Each repetition selected 500 records per class from the source holdout and target evaluation pools using reproducible random seeds.

The trained models and explainers were held fixed. Mean class-wise Spearman correlations, Top-10 Jaccard similarities and within-class sampling standard deviations were recorded. These repetitions measure sampling sensitivity within the available evaluation pools, not independent model replications.

4.6. Association and statistical testing

Class-wise recall degradation was calculated as:

Recall degradation = Source recall − Target recall

Spearman correlation was then calculated between attribution drift and recall degradation. Eight planned tests were conducted: all-class and attack-only analyses for both models and transfer directions.

Exact permutation tests were used, followed by Holm correction across the eight tests. Given the small number of class-level observations, the association analysis was treated as exploratory.

V. RESULTS

5.1. Cross-dataset detection performance

Both classifiers achieved high source-holdout performance but experienced substantial degradation during cross-dataset evaluation.

Training→ Evaluation	Model	Source Macro F1	Cross-dataset Macro F1
2017 → 2018	Random Forest	0.9690	0.2780
2017 → 2018	XGBoost	0.9695	0.2717
2018 → 2017	Random Forest	0.8864	0.2739
2018 → 2017	XGBoost	0.8874	0.2500

The absolute Macro F1 reductions ranged from 61.25 to 69.78 percentage points. Neither classifier maintained its source-domain performance when evaluated on the other dataset.

These results demonstrate substantial cross-dataset generalization degradation under the selected experimental conditions.

5.2. Class-wise SHAP attribution drift

Despite the decline in detection performance, class-wise SHAP attribution rankings remained relatively similar across datasets.

Transfer direction	Model	Mean Spearman correlation
2017 → 2018	Random Forest	0.8226
2017 → 2018	XGBoost	0.8602
2018 → 2017	Random Forest	0.9366
2018 → 2017	XGBoost	0.9286

Attribution-ranking agreement was higher in the CSE-CIC-IDS2018 to CIC-IDS2017 direction for both classifiers.

Class-level differences were also observed. In the 2017 to 2018 direction, DoS-SlowHTTPTest showed the lowest Spearman correlation for both Random Forest (0.6117) and XGBoost (0.6522). In the reverse

direction, DoS-Hulk showed the lowest correlations for Random Forest (0.8468) and XGBoost (0.8151).

5.3. Repeated-sampling stability

Across ten repeated sampling experiments, mean class-wise Spearman correlations ranged from 0.8211 to 0.9333 across the four model and transfer-direction combinations. Mean Top-10 Jaccard similarity ranged from 0.6459 to 0.8531.

The mean within-class sampling standard deviation of Spearman correlation ranged from 0.0045 to 0.0117. This indicates relatively low sensitivity to sample selection within the evaluated pools.

However, stability varied by class and transfer direction. DoS-SlowHTTPTest showed the lowest average ranking agreement in the 2017 to 2018 direction, while DoS-Hulk showed the lowest agreement in the reverse direction.

These results indicate that the observed attribution rankings were reasonably consistent under repeated sampling, although this does not establish explanation faithfulness or correctness.

5.4. Association between attribution drift and recall degradation

The association between attribution drift and recall degradation varied across the four experimental conditions.

Transfer direction	Model	All-class ρ	Attack-only ρ
2017 → 2018	Random Forest	0.6071	0.3714
2017 → 2018	XGBoost	0.3571	-0.0286
2018 → 2017	Random Forest	0.9286	0.8857
2018 → 2017	XGBoost	0.4286	0.0857

The strongest observed association occurred for Random Forest in the 2018 to 2017 direction. Its all-class Spearman correlation was 0.9286, with an unadjusted exact permutation p-value of 0.0070. However, the Holm-adjusted p-value was 0.0560, which did not meet the conventional 0.05 significance threshold.

The corresponding attack-only analysis produced a correlation of 0.8857, but its adjusted p-value was 0.2310. None of the eight tests remained statistically significant after Holm correction.

5.5. Attribution stability despite detection failure

An important observation was that high SHAP ranking similarity could coexist with extremely poor target-class detection.

For example, when CSE-CIC-IDS2018-trained XGBoost was evaluated on CIC-IDS2017, FTP-BruteForce exhibited a Spearman correlation of 0.9656, while its target recall was only 0.0002.

Similarly, CSE-CIC-IDS2018-trained Random Forest achieved a correlation of 0.9348 for DoS-SlowHTTPTest, despite obtaining zero target recall.

These examples demonstrate that global class-wise feature-attribution ranking similarity does not guarantee successful detection transfer.

VI. DISCUSSION

The experiments reveal a distinction between predictive generalization and global explanation-ranking stability. Both classifiers experienced severe cross-dataset performance degradation, yet their class-wise SHAP rankings often remained similar.

This suggests that a model may preserve the relative ordering of influential features while failing to recognize the target dataset's attack instances. Similar attribution rankings should therefore not be interpreted as evidence that a model has learned transferable attack behaviour.

The association analysis further indicates that attribution drift is not a consistent predictor of recall degradation across the evaluated models and transfer directions. Although Random Forest showed a strong observed association in one direction, it did not remain statistically significant after multiple-testing correction. XGBoost showed weaker and less consistent relationships.

The findings support a more careful approach to explainable intrusion detection evaluation. Detection metrics and explanation-ranking comparisons should be reported separately because they measure different aspects of model behaviour.

VII. LIMITATIONS

This study has several limitations.

First, only three selected capture files from each dataset were used. The findings therefore cannot be generalized to every capture or attack scenario in CIC-IDS2017 and CSE-CIC-IDS2018.

Second, the prepared datasets were balanced through sampling at 5,000 records per class. Consequently, the reported metrics do not represent performance under the original class prevalence or naturally occurring network traffic distributions.

Third, source evaluation used a random flow-level split rather than a capture-level or time-based split. Related flows may therefore occur across training and source-holdout partitions, potentially inflating source-domain performance.

Fourth, feature harmonization used exact-name matching. Although this provides a conservative shared feature set, it does not prove complete semantic or extraction equivalence. The dataset documentation also identifies different CICFlowMeter implementations, making extraction differences a possible source of domain variation.

Fifth, the association tests used only six or seven class-level observations. The statistical power is limited, and the exchangeability assumption underlying permutation testing across different traffic classes may not fully hold.

Finally, SHAP rank correlations measure similarity in feature-importance ordering. They do not establish explanation faithfulness, causal validity or local explanation correctness. Repeated sampling also used fixed trained models and overlapping evaluation pools, so the resulting variability should not be interpreted as independent experimental replication.

VIII. CONCLUSION

This study experimentally investigated the relationship between class-wise SHAP attribution drift and detection degradation under bidirectional cross-dataset shift in network intrusion detection.

Random Forest and XGBoost achieved high source-holdout Macro F1 scores but experienced substantial performance degradation when transferred between CIC-IDS2017 and CSE-CIC-IDS2018. Despite this decline, class-wise SHAP attribution rankings remained relatively stable across many classes.

The association between attribution drift and recall degradation varied by classifier and transfer direction. No tested association remained statistically significant after Holm correction. Furthermore, examples of near-identical attribution rankings alongside almost zero target recall demonstrate that explanation-ranking stability cannot be treated as a substitute for predictive evaluation.

The main conclusion is that cross-dataset detection performance and global SHAP attribution-ranking stability should be assessed as separate properties of explainable intrusion detection systems. Future research should extend this investigation to additional datasets, capture-level evaluation, temporal validation, larger class samples and further explanation-faithfulness tests.

REFERENCES

[1] M. Cantone, C. Marrocco, and A. Bria, “On the Cross-Dataset Generalization of Machine Learning for Network Intrusion Detection,” *arXiv preprint arXiv:2402.10974*, 2024.

[2] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[3] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” in *Proc. 4th Int. Conf. Inf. Syst. Security and Privacy (ICISSP)*, 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[4] Canadian Institute for Cybersecurity, “CSE-CIC-IDS2018: A Realistic Cyber Defense Dataset,” University of New Brunswick, n.d. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html

[5] E. M. Maseno, Y. Sun, and Z. Wang, “Reliability Auditing of Explanations for Machine-Learning-Based Intrusion Detection Systems,” *Journal of Computer Virology and Hacking Techniques*, vol. 22, Art. no. 80, 2026, doi: 10.1007/s11416-026-00664-7.