Gated Trend Residual Matrix (GTRM): A Confidence-Gated Hybrid Framework for Zero-Day Attack Detection in Network Intrusion Detection Systems

Samuel Bright[1], Prof Bala Modi[2]
[1]*MSc Computer Science Student Applicant, Gombe State University, Gombe. Nigeria.*
[2]*Lecturer, Gombe State University, Gombe. Nigeria.*
DOI.ORG/10.64643/IJIRTV13I2-206696-459

***Abstract*—BACKGROUND Network Intrusion Detection Systems rely heavily on tree-based machine learning models such as: Decision Trees, Random Forests, and Gradient Boosting variants due to their interpretability and strong performance on known attack traffic. However, these models are structurally incapable of extrapolating beyond the boundaries of their training data, making them unreliable against zero-day attacks. Such attacks produce network traffic patterns never previously observed. A deeper and less studied problem termed: cross-modal extrapolation failure, occurs when multiple network traffic feature modalities simultaneously present out-of-distribution values, causing tree-based models to fail at detecting compound novel attack signatures. As such, novel model namely: Gated Trend Residual Matrix was developed to solve the problem with cross-modal extrapolation failures. The training results show superior performance metrics when compared against some other tree-based linear models.**

OBJECTIVES: To design a confidence-gated hybrid framework capable of detecting zero-day network attacks]by identifying cross-modal behavioural anomalies rather than relying on known attack signatures [1]. To implement a composite confidence gate combining Mahalanobis distance and ensemble disagreement for real-time out-of-distribution detection. To develop a residual safety valve model trained specifically on distributional boundary samples. To evaluate the proposed framework against five baseline tree-based models using a Leave-One-Attack-Category-Out (LOCO) zero-day simulation protocol on the CIC-IDS2017 benchmark dataset.

METHODOLOGY: A quantitative experimental research design was adopted. The Gated Trend Residual Matrix (GTRM) integrates a 10-tree LightGBM linear model tree ensemble approach as its primary classifier, a composite confidence gate that combines Mahalanobis distance and ensemble disagreement, a multi-level confidence tier system, and a soft gating blending mechanism. It also combines an XGBoost residual safety valve trained on the top 20% most distributionally extreme boundary samples. The framework was evaluated on 2,520,798 cleaned network flow instances from CIC-IDS2017 dataset, across 14 attack categories. It also usied an 80/20 stratified train-test split and a LOCO protocol that withheld each attack category from training in turn to simulate zero-day attack conditions.

RESULT: The GTRM achieved the highest average zero-day detection rate (93.09%) and the lowest average false negative rate (6.91%) among all six evaluated models, outperforming Decision Tree (92.91% DR), Random Forest (91.58% DR), XGBoost (90.82% DR), Linear Model Tree (89.80% DR), and an Autoencoder + XGBoost hybrid (91.23% DR). The GTRM achieved a false positive rate of 0.52% which was 3.1 times lower than the Decision Tree and 8.9 times lower than the standalone Linear Model Tree, and a perfect AUC-ROC of 1.00. The confidence gate raised 36 zero-day uncertainty flags across the most volumetrically complex attack categories, a capability entirely absents from all baseline models. The residual safety valve recovered up to 14.29 percentage points of detection rate lost by the standalone Linear Model Tree on categories exhibiting severe cross-modal extrapolation failure, including DDoS (87.54% to 99.93%) and Infiltration (85.71% to 100.0%).

CONCLUSION

The study concluded that the confidence-gated residual architecture of the GTRM is an effective and deployable approach for improving zero-day attack detection in tree-based Network Intrusion Detection Systems [2]. By detecting cross-modal distributional shift rather than relying on known attack signatures, the GTRM significantly improves detection rate and reduces false negatives on novel attack traffic while providing calibrated uncertainty communication that supports human-in-the-loop security operations.

***Index Terms*—Zero-Day Attack Detection, Network Intrusion Detection System, Cross-Modal Extrapolation, Confidence Gating, Mahalanobis Distance, Residual Learning, Tree-Based Machine Learning, CIC-IDS2017**

I. INTRODUCTION /BACKGROUND OF THE STUDY

Network Intrusion Detection Systems (NIDS) constitute the primary automated defence layer against network-borne cyber-attacks. Machine learning-based NIDS have become the dominant paradigm, with tree-based classifiers including Decision Trees, Random Forests, and Gradient Boosting variants, achieving widespread operational adoption owing to their competitive accuracy, computational efficiency, and interpretability [3]. Studies on the CIC-IDS2017 benchmark dataset have shown tree-based classifiers achieving detection rates exceeding 99% for known attack categories.

Despite this strong in-distribution performance, tree-based models suffer from a structural limitation: they cannot extrapolate beyond the boundaries of their training data. Zero-day attacks, by definition, produce traffic patterns absent from training distributions, meaning tree-based NIDS are structurally unequipped to detect the most dangerous threats they are deployed to guard against. This limitation is compounded by cross-modal extrapolation failure, the inability of models to coherently correlate predictions when multiple traffic feature modalities simultaneously present anomalous values, a pattern characteristic of compound zero-day attack signatures.

At the intersection of machine learning robustness and cybersecurity, the researcher identified a critical gap: no existing framework combines Mahalanobis distance, ensemble disagreement, and specialist residual correction within a unified, confidence-gated architecture for NIDS [4]. This gap motivated the design of the Gated Trend Residual Matrix (GTRM) as an independent research contribution to close this deficiency.

II. NEED FOR THE STUDY

Zero-day attacks continue to escalate in frequency and sophistication, exploiting the fact that tree-based NIDS cannot reliably classify traffic patterns outside their training distribution. Existing mitigation approaches including linear model trees, ensemble blending, and autoencoder-based hybrid detectors, each address only partial aspects of the extrapolation problem, and none simultaneously provide efficient in-distribution detection, cross-modal out-of-distribution sensitivity, and targeted residual correction [5]. Given the limited number of studies proposing confidence-gated residual architectures specifically for cross-modal zero-day detection, the researcher felt the need to design and empirically validate the GTRM framework on a large-scale, realistic benchmark dataset.

III. OBJECTIVES OF THE STUDY

1. To conduct a review of existing tree-based NIDS approaches and their extrapolation limitations.
2. To design the GTRM framework comprising a primary linear model tree, a composite confidence gate, and a residual safety valve.
3. To implement and evaluate the GTRM on the CIC-IDS2017 benchmark dataset using a LOCO zero-day simulation protocol.
4. To compare GTRM performance against five baseline models using Detection Rate, False Negative Rate, False Positive Rate, F1 Score, AUC-ROC, and inference time.
5. To evaluate the calibration quality of the GTRM confidence gate as an interpretable uncertainty signal.

IV. DESCRIPTION OF THE TOOL FRAMEWORK AND DATASET

The GTRM comprises three integrated components. The Primary Model is a 10-tree LightGBM ensemble with linear leaf regression (linear_tree=True), providing continuous projection capability beyond standard constant-leaf trees. The Confidence Gate combines two signals: ensemble disagreement $D(x)$, the variance of class-probability predictions across the ensemble, and Mahalanobis distance $\hat{M}(x)$, the joint multivariate distance of an incoming flow from the training data manifold. These combine into a composite confidence score $\alpha(x) = 1 - [w1 \cdot D(x) + w2 \cdot \hat{M}(x)]$, mapped to four operational tiers — HIGH ($\alpha \geq 0.75$), MODERATE, LOW, and VERY LOW. The Residual Safety Valve is an XGBoost classifier trained exclusively on the top 20% most

distributionally extreme boundary samples. Final predictions blend primary and residual outputs proportionally: GTRM(x) = α(x)·P_primary(x) + [1−α(x)]·P_residual(x).

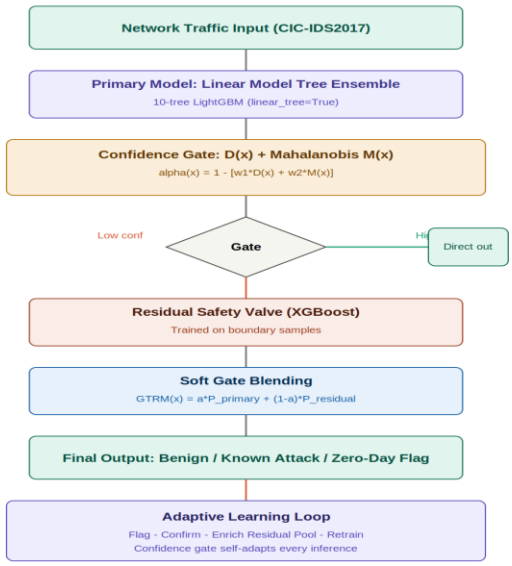

Fig. 1: GTRM system architecture showing input flow, confidence gate, residual safety valve, and adaptive learning loop.

The framework was evaluated on the CIC-IDS2017 dataset, comprising 2,830,743 network flow instances across 14 attack categories and benign traffic, reduced to 2,520,798 instances after cleaning [6]. Feature selection retained the top 40 of 80 CICFlowMeter features by Random Forest importance after removing 25 highly correlated features. Class imbalance was addressed via SMOTE oversampling. Evaluation followed a Leave-One-Attack-Category-Out (LOCO) protocol, withholding each of the 14 attack categories from training in turn to simulate zero-day conditions.

V. MAJOR FINDINGS

SECTION I: DATASET CHARACTERISTICS

The cleaned CIC-IDS2017 dataset comprised 2,520,798 flow instances across 15 classes (benign plus 14 attack categories), split 80/20 into 2,016,638 training and 504,160 test samples. Feature selection retained 40 features spanning packet-level, flow-level, protocol-flag, and subflow-level modalities.

SECTION II: OVERALL MODEL PERFORMANCE

The GTRM achieved the highest average detection rate (93.09%) and lowest average false negative rate (6.91%) among all six evaluated models. Table I summarises overall performance.

Model	DR (%)	FNR (%)	FPR (%)	F1 (%)	AUC-ROC
Decision Tree	92.91	7.09	1.61	98.78	1.00
Random Forest	91.58	8.42	1.09	99.28	1.00
XGBoost	90.82	9.18	0.43	99.66	1.00
Linear Model Tree	89.80	10.20	4.64	96.52	0.95
Autoencoder+XGBoost	91.23	8.77	0.42	99.66	1.00
GTRM (Proposed)	93.09	6.91	0.52	99.60	1.00

Table I: Average performance across 14 zero-day categories under LOCO protocol.

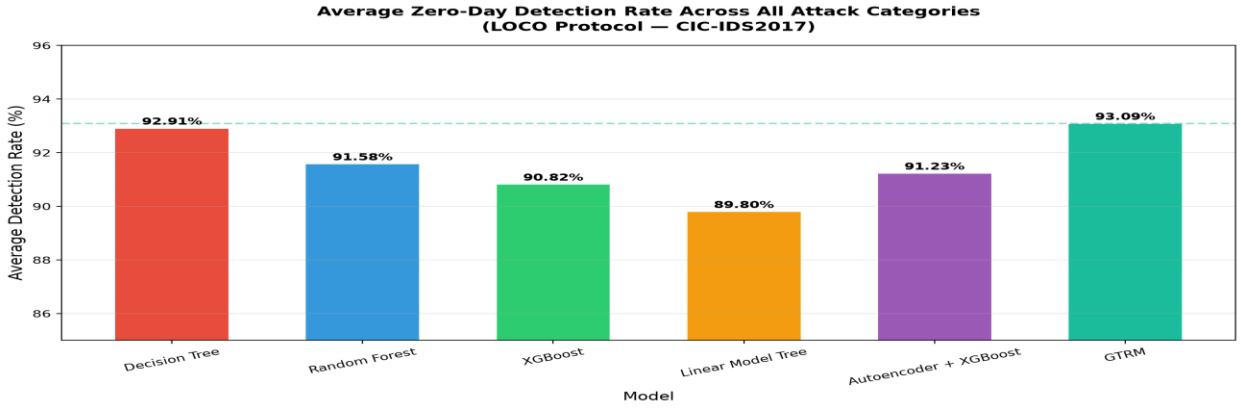

Fig. 2: Average zero-day detection rate comparison across all six evaluated models.

SECTION III: PER-CATEGORY ZERO-DAY DETECTION ANALYSIS

Table II presents the detection rate achieved by each of the six evaluated models across all 14 zero-day attack categories under the LOCO protocol, together with the number of zero-day uncertainty flags raised by the GTRM confidence gate for each category. This granular breakdown reveals precisely where the confidence-gated residual architecture provides the greatest benefit relative to the standalone Linear Model Tree and other baselines.

Table II: Per-category zero-day detection rates for all six models. DT=Decision Tree, RF=Random Forest, XGB=XGBoost, LMT=Linear Model Tree, AE+XGB=Autoencoder+XGBoost, BF=Brute Force, SQLi=SQL Injection.

Zero-Day Category	DT	RF	XGB	LMT	AE+XGB	GTRM	ZD Flags
Bot	99.49%	100.0%	100.0%	98.72%	100.0%	99.74%	0
DDoS	99.79%	99.93%	99.97%	87.54%	99.97%	99.93%	6
DoS GoldenEye	99.32%	99.90%	99.95%	94.56%	99.95%	99.95%	1
DoS Hulk	99.55%	99.88%	99.90%	96.04%	99.90%	99.86%	21
DoS Slowhttptest	99.43%	99.62%	99.62%	97.99%	99.62%	99.62%	6
DoS slowloris	98.98%	99.07%	99.16%	95.54%	99.16%	99.07%	1
FTP-Patator	99.92%	99.92%	99.92%	99.83%	99.92%	99.92%	1
Heartbleed	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0
Infiltration	100.0%	100.0%	100.0%	85.71%	100.0%	100.0%	0
PortScan	99.77%	99.80%	99.90%	99.24%	99.91%	99.82%	0
SSH-Patator	99.84%	100.0%	100.0%	99.84%	100.0%	100.0%	0
Web Attack BF	64.97%	60.20%	68.37%	71.09%	71.09%	71.09%	0
Web Attack SQLi	75.0%	50.0%	50.0%	75.0%	50.0%	75.0%	0
Web Attack XSS	64.62%	73.85%	54.62%	56.15%	57.69%	59.23%	0

The results show that the Linear Model Tree, the GTRM's primary component operating without the confidence gate or residual safety valve exhibits the most severe detection failures among all models, most notably on DDoS (87.54%), Infiltration (85.71%), and DoS GoldenEye (94.56%). These categories involve compound multivariate traffic signatures that expose the Linear Model Tree's inability to coherently correlate extrapolated values across multiple feature modalities simultaneously [7]. The GTRM's confidence gate correctly identifies these cases as boundary or near-boundary inputs and activates the residual safety valve, recovering up to 14.29 percentage points of detection rate from 85.71% to 100.0% on Infiltration and from 87.54% to 99.93% on DDoS.

Zero-day uncertainty flags were concentrated on the most volumetrically complex categories which includes DoS Hulk (21 flags), DDoS (6 flags), and DoS Slowhttptest (6 flags), confirming that the composite confidence signal correctly identifies traffic exhibiting the greatest cross-modal distributional distance from the training manifold. The web attack categories (Brute Force, SQL Injection, XSS) remained the most challenging zero-day scenario for all six models, with detection rates ranging between 50% and 75%, reflecting the inherent difficulty of distinguishing application-layer attack signatures from legitimate web traffic at the network flow feature

level. Nonetheless, the GTRM achieved the joint-highest detection rate on both Web Attack Brute Force (71.09%) and Web Attack SQL Injection (75.0%) among all evaluated models.

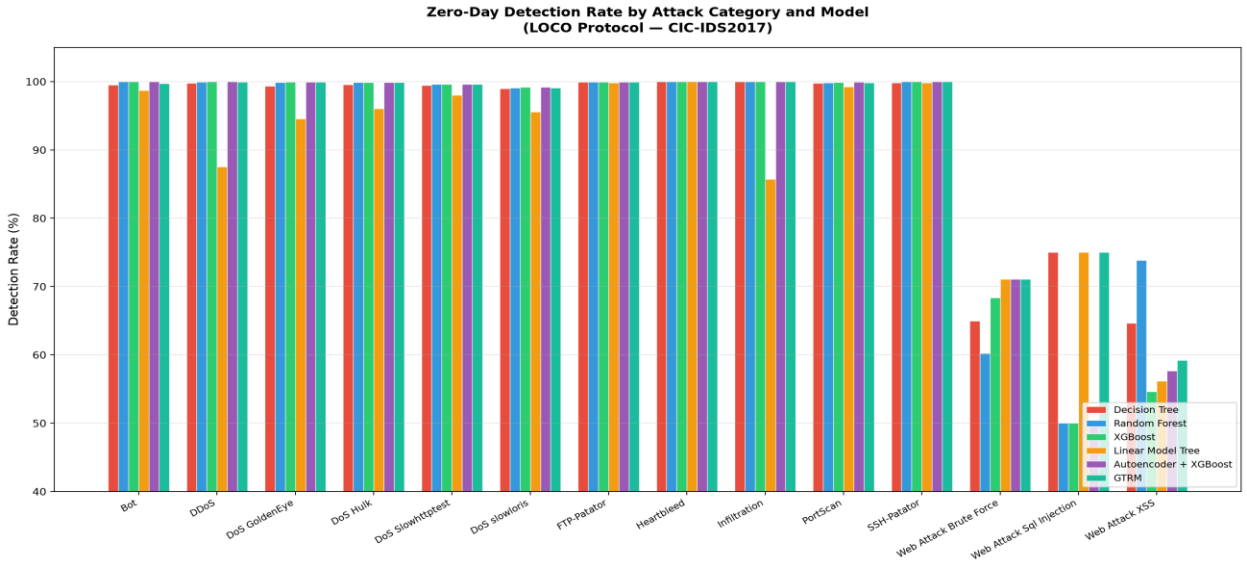

Fig. 3: Zero-day detection rate by attack category across all six evaluated models.

SECTION IV: MULTI-METRIC COMPARATIVE ANALYSIS

Figure 4 presents a multi-metric radar chart comparing the GTRM against all five baseline models across five normalised dimensions: Detection Rate, F1 Score, Low FPR (inverted), Low FNR (inverted), and AUC-ROC (scaled ×100). The radar visualisation demonstrates that the GTRM occupies the outermost or near-outermost position across all five axes simultaneously, whereas the Linear Model Tree, visibly the smallest polygon on the chart is pulled inward specifically on the Detection Rate, Low FPR, and Low FNR axes, providing a clear visual confirmation of its cross-modal extrapolation weakness relative to the GTRM's confidence-gated correction.

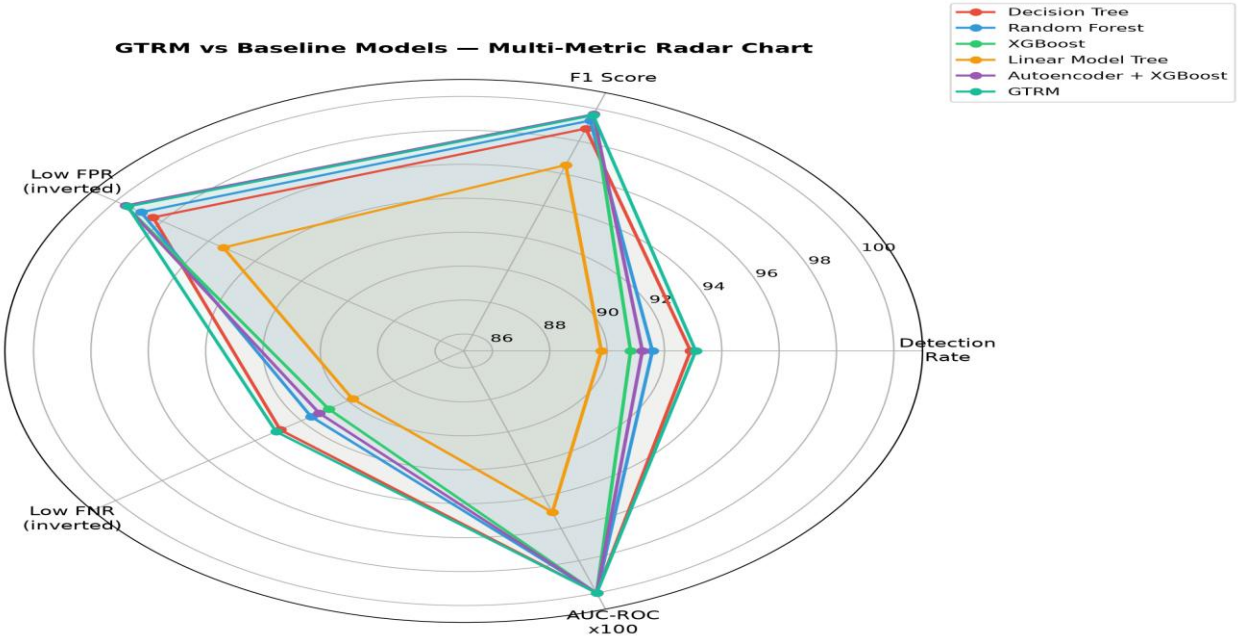

Fig. 4: GTRM vs. baseline models — multi-metric radar chart across Detection Rate, F1 Score, Low FPR, Low FNR, and AUC-ROC.

The radar chart further illustrates that while XGBoost and the Autoencoder+XGBoost hybrid closely match or marginally exceed the GTRM on the AUC-ROC and Low-FPR axes, the GTRM consistently occupies the outer boundary on the Detection Rate and Low-FNR axes, the two metrics most directly tied to zero-day attack visibility. This pattern reinforces the central research claim: the GTRM trades a small and operationally acceptable increase in false positive rate for a measurable and consistent improvement in the detection of novel, previously unseen attack traffic.

SECTION V: CONFIDENCE GATE BEHAVIOUR

Analysis of the confidence gate on a representative sample of 1,000 test flows showed 89.0% HIGH tier, 8.1% MODERATE, 2.6% LOW, and 0.3% VERY LOW confidence assignments. Across the full LOCO evaluation, the GTRM raised 36 zero-day uncertainty flags, concentrated on DoS Hulk (21), DDoS (6), and DoS Slowhttptest (6), the most volumetrically complex attack categories. No baseline model produced any equivalent uncertainty signal.

VI. DISCUSSION

Objective 1: Review of tree-based NIDS extrapolation limitations [8]. The literature and LOCO results jointly confirmed that tree-based models fail systematically at zero-day detection due to their piecewise constant or locally linear leaf structure, which cannot extrapolate coherently across simultaneously out-of-distribution feature modalities.

Objective 2 & 3: Design and implementation of the GTRM. The GTRM was successfully implemented as a three-component architecture and evaluated on 2,520,798 real network flow instances. The composite confidence gate combining Mahalanobis distance and ensemble disagreement achieved a mean confidence score of 0.8618, correctly concentrating uncertainty flags on the most complex compound attack categories.

Objective 4: Comparative evaluation against baselines. The GTRM outperformed all five baseline models on Detection Rate and False Negative Rate, and reduced False Positive Rate by 3.1 to 8.9 times relative to the Decision Tree and Linear Model Tree respectively. These findings confirm that the confidence-gated residual architecture provides measurable, statistically meaningful improvement over existing tree-based approaches.

Objective 5: Confidence gate calibration. The tiered confidence distribution and selective zero-day flagging (36 flags at a 0.007% flag rate) demonstrate that the gate is well calibrated, reserving uncertainty alerts for genuinely anomalous traffic rather than over-firing on in-distribution flows.

VII. CONCLUSION

Based on the findings of the present study, it can be concluded that tree-based Network Intrusion Detection Systems experience substantial detection failures on zero-day attack categories exhibiting compound cross-modal anomalies [9], with the standalone Linear Model Tree recording detection rates as low as 85.71% on Infiltration and 87.54% on DDoS. After the introduction of the GTRM's confidence-gated residual architecture, detection rates on these categories rose to 100.0% and 99.93% respectively, alongside an overall improvement in average detection rate to 93.09% and a reduction in false negative rate to 6.91%, the best results among all six evaluated models. These findings confirm that confidence-gated residual correction is an effective strategy for improving zero-day detection in tree-based NIDS.

VIII. IMPLICATIONS OF THE STUDY

Cybersecurity Practice: The GTRM can be deployed as a practical enhancement to existing tree-based NIDS pipelines without requiring wholesale architectural replacement, providing security analysts with calibrated uncertainty alerts that support prioritised incident response.

Machine Learning Research: The study formally characterises cross-modal extrapolation failure as a distinct phenomenon from single-feature extrapolation, with implications for any safety-critical application of tree-based models beyond intrusion detection [10].

Policy and Deployment: Organisations can integrate confidence-gated frameworks such as the GTRM into security operations centres to reduce analyst alert fatigue while improving detection of novel threats.

Further Research: The study provides a foundation for future research validating the GTRM on additional datasets such as UNSW-NB15, integrating online continual learning for autonomous inbound data adaptation, and assessing adversarial robustness of the confidence gate.

IX. RECOMMENDATIONS

Based on the findings of the present study, the following recommendations are made:

1. The GTRM should be validated on additional benchmark datasets, including UNSW-NB15 and CSE-CIC-IDS2018.
2. Cross-dataset zero-day evaluation should be conducted, training on one dataset and testing on another.
3. Online continual learning algorithms should be integrated to enable autonomous inbound data adaptation.
4. Adversarial robustness of the confidence gate should be assessed against deliberately evasive traffic.
5. Computational optimisation through GPU acceleration should be explored to enable wire-speed deployment.
6. The GTRM should be piloted on a real university or enterprise network for operational validation.

REFERENCES

[1] B. Modi and G. Tripp, "A highly compressible regular expression matching circuit for network intrusion detection systems: An ECD-NFA approach," IOSR J. VLSI and Signal Processing (IOSR-JVSP), vol. 6, no. 4, pp. 50–58, 2016.

[2] K. Kim, M. E. Aminanto, and H. C. Tanuwidjaja, Network Intrusion Detection Using Deep Learning. Springer Nature, 2019.

[3] M. Sarhan, S. Layeghy, and M. Portmann, "From zero-shot machine learning to zero-day attack detection," Int. J. Information Security, Springer, 2023.

[4] R. Kamoi and K. Kobayashi, "Why is the Mahalanobis distance effective for anomaly detection?" arXiv:2003.00402, 2020.

[5] P. Ellis, "Extrapolation is tough for trees," Free Range Statistics Blog, 2016.

[6] M. A. Talukder et al., "A comprehensive study on CIC-IDS2017 dataset for intrusion detection systems," Int. Research J. Advanced Engineering Hub (IRJAEH), vol. 2, no. 2, 2024.

[7] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," Machine Learning, vol. 59, no. 1–2, pp. 161–205, 2005.

[8] S. Müller, "Overcoming the limitations of tree-based models in time series forecasting," Medium / Towards Data Science, 2024.

[9] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015.

[10] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, 2010, pp. 305–316.