Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee4156

MACHINE LEARNING-BASED DETECTION OF BGP ROUTING ANOMALIES FOR ENHANCED NETWORK SECURITY AND RESILIENCE: A HYBRID MULTI-SCALE TEMPORAL LEARNING APPROACH

Zhou Lu[1*], XuanPing Zou[2], KaiLang Wei[3]
[1]*Northeastern University, San Jose 95113, USA.*
[2]*Zoom, Inc., San Jose 95113, USA.*
[3]*Emory University, Atlanta 30322, USA.*
[*]*Corresponding Author: Zhou Lu*

Abstract: The Border Gateway Protocol (BGP) lacks built-in validation, so worm outbreaks, power outages, and other disruptive events propagate as storms of routing updates that threaten the stability of the Internet. Machine learning (ML) detectors trained on historical incidents report high accuracy on the incidents they were trained on, whereas their ability to recognize previously unseen events, which is what matters in operation, has received little attention. This paper studies BGP anomaly detection under a leave-one-event-out protocol on real RIPE RIS traces of the Code Red I, Nimda, and Slammer worms and the Moscow blackout. We show that per-minute classifiers generalize poorly because anomalies are sustained multi-hour deviations hidden in bursty traffic, and we propose PyraGuard, a hybrid multi-scale temporal learning framework. PyraGuard fuses, at the decision level, a learned sequence branch with MS-GBDT, a gradient-boosted model on causal multi-scale descriptors spanning 1 to 240 minutes; for the learned branch, we also design MSTAN, an interpretable temporal-pyramid attention network. Multi-scale descriptors alone raise the mean F1 score of LightGBM from 60.2% to 79.1%, and decision-level fusion further improves every competent learner. PyraGuard with a GRU branch achieves a mean F1 score of 83.7% and an average precision of 87.1% on unseen events, outperforming seven baselines, and raises no false alarm on unseen-era 2017 traffic, at about 0.20 ms per decision on a single CPU core.
Keywords: Border Gateway Protocol; Anomaly detection; Routing security; Network resilience; Multi-scale temporal learning; Gradient boosting; Deep learning

1 INTRODUCTION

The Border Gateway Protocol (BGP) is the inter-domain routing protocol that glues together the tens of thousands of autonomous systems (ASes) that form the Internet [1]. BGP was designed under the assumption that the information exchanged between neighbors is trustworthy, and it provides no built-in mechanism to validate the origin or the path of a route [2]. Consequently, events that are not routing attacks in themselves, such as fast-spreading worms that saturate links and routers, large power outages that disconnect exchange points, misconfigurations, and prefix hijacks, manifest as storms of BGP update messages that destabilize reachability far beyond the networks directly affected [3-4]. Detecting such anomalies quickly and reliably is therefore a cornerstone of both the security and the resilience of the Internet routing system.

Security extensions to BGP have been studied for more than a decade, yet none of them has struck an adequate balance between comprehensive protection and deployment cost [2]. Operators therefore rely heavily on monitoring: public route collectors stream update messages from many vantage points, and detection systems analyze these streams to raise alarms [4-5]. Machine learning (ML) has become the dominant tool for this task. Classical classifiers such as support vector machines (SVMs), naive Bayes, and hidden Markov models [6-8], recurrent neural networks (RNNs) [9-10], multi-scale long short-term memory (LSTM) networks [11-12], weakly supervised attention models [13], and graph attention networks [14] have all been applied to minute-level features extracted from BGP update messages.

Two aspects of the problem remain insufficiently addressed. First, most studies train and test on partitions of the same anomalous events, so the reported accuracy reflects the recognition of incidents that have already been observed. In operation, the next incident is by definition unseen: it may stem from a different cause, occur in a different era of Internet growth, and exhibit a different magnitude and temporal profile. Second, routing anomalies are rarely instantaneous spikes. They are sustained deviations that build up and decay over minutes to hours and are superimposed on bursty background churn. Detectors that decide from a single minute or a short window are therefore exposed to noise and to the absolute traffic level of the training collectors, both of which change from one event to the next.

This paper revisits ML-based BGP anomaly detection from the perspective of generalization to unseen events. Using real RIPE RIS traces of the Code Red I and Nimda worms (2001), the Slammer worm (2003), and the Moscow power blackout (2005), together with anomaly-free RIPE and BCNET traffic [15], we adopt a leave-one-event-out (LOEO) protocol in which the test event never contributes training data. A preliminary analysis reveals that the dominant factor

© By the Author(s) 2022, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

for transferable detection is multi-scale temporal context: aggregating evidence over horizons from one minute to four hours markedly improves cross-event performance, whereas per-minute decisions lag far behind. Building on this observation, we propose PyraGuard, a hybrid multi-scale temporal learning framework for BGP anomaly detection. The main contributions are as follows.

• We formulate BGP anomaly detection as an online, per-minute decision problem evaluated under a strict LOEO protocol, and we quantify the collector- and era-induced distribution shift that makes this setting challenging.

• We design the Multi-Scale Temporal Attention Network (MSTAN), a lightweight causal model that combines a squeeze-and-excitation feature gate, an equal-budget temporal pyramid at 1-, 4-, and 16-minute resolutions, dilated convolutions with attention pooling, and a scale-attention fusion layer, and that exposes the features and time scales that drive each decision.

• We introduce MS-GBDT, a gradient-boosted model on causal multi-scale descriptors, and fuse it at the decision level with a learned sequence branch. The fusion is agnostic to the learner; we instantiate it with GRU, TCN, and MSTAN branches and show that it consistently improves every competent learner.

• Experiments with seven baselines show that PyraGuard attains a mean F1 score of 83.7% and an AP of 87.1% on unseen events, 2.8 points above the strongest stand-alone model, and raises no false alarm on unseen-era 2017 traffic, while MSTAN reveals the features and time horizons that drive each alarm.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data and formulates the problem. Section 4 presents PyraGuard. Section 5 details the experimental setup, Section 6 reports and discusses the results, and Section 7 concludes the paper.

2 RELATED WORK

2.1 BGP Security and Anomaly Detection

BGP-4 propagates reachability information hop by hop [1], and its lack of authentication enables prefix hijacks, route leaks, and the rapid amplification of failures [2]. Al-Musawi *et al.* surveyed BGP anomaly detection techniques and classified anomalies into direct intended anomalies such as hijacks [3], direct unintended anomalies such as misconfigurations, indirect anomalies such as worm outbreaks, and link failures such as power outages. Hijack-oriented systems exploit control-plane semantics: ARTEMIS enables an AS to detect hijacks of its own prefixes from public monitoring streams and to neutralize them within a minute [4], and AP2Vec embeds IP prefixes to flag functional changes of the ASes along new routes in an unsupervised manner [5]. Such systems excel at prefix-level attacks, but the large-scale instabilities caused by worms and blackouts perturb the aggregate dynamics of the update stream rather than individual origin-prefix pairs. PyraGuard targets this class of indirect anomalies and link failures and complements prefix-level defenses.

2.2 Machine Learning for BGP Anomaly Detection

A long line of work summarizes update messages into volume and AS-path features at a one-minute granularity and trains supervised classifiers. Al-Rousan and Trajković employed SVM, hidden Markov, and naive Bayes models to classify the Code Red I, Nimda, and Slammer worms, and subsequent studies compared further classical learners and SVM kernels [6-8, 16]. Li *et al.* applied LSTM and gated recurrent unit (GRU) networks to detect worms, ransomware, and blackouts in RIPE and Route Views data and extended the comparison to broad learning systems [9-10]. Cheng *et al.* showed that the time scale used to preprocess the traffic strongly affects accuracy and proposed the multi-scale LSTM [11], which was later combined with wavelet decomposition and hierarchical attention over time scales [12]. Dong *et al.* distilled knowledge from existing detection systems into a self-attention LSTM under weak supervision for ISP self-operated detection [13], and Peng *et al.* combined seasonal-trend decomposition with graph attention networks to model relations among features and time steps [14]. These studies establish the value of temporal modeling; PyraGuard differs in that it is evaluated strictly on unseen events and combines learned and descriptor-based multi-scale representations within a single detector.

2.3 Multi-Scale Temporal Modeling

Temporal convolutional networks (TCNs) cover long histories efficiently with dilated causal convolutions [17], and Transformers model long-range interactions through self-attention [18]. Multi-resolution representations have recently proved effective for long sequences: Pyraformer builds a pyramidal graph whose coarse scales summarize long-range dependencies at linear complexity [19], and the Anomaly Transformer exploits association discrepancies for unsupervised time-series anomaly detection [20]. Channel recalibration with squeeze-and-excitation (SE) blocks and the focal loss for extreme class imbalance are widely used building blocks [21-22]. MSTAN adapts these ideas to BGP monitoring through an explicit equal-budget temporal pyramid, strictly causal processing for online detection, and interpretable feature and scale attention.

3 DATA AND PROBLEM FORMULATION

3.1 BGP Datasets

We use the public BGP datasets of the Communication Networks Laboratory at Simon Fraser University [15], which have been employed in several prior studies [9-10]. Update messages collected by the RIPE RIS route collector rrc04 were parsed and summarized into 37 features per minute. Volume features include the numbers of announcements, withdrawals, announced and withdrawn network layer reachability information (NLRI) prefixes, duplicate announcements, implicit and duplicate withdrawals, the update arrival rate, the numbers of updates with IGP, EGP, and INCOMPLETE origin attributes, and the average packet size. AS-path features include the average and maximum AS-path lengths, the average unique AS-path length, the maximum and average edit distances between AS paths, and binned indicators of the maximum AS-path length and the maximum edit distance. Each minute is labeled as anomalous or regular according to the documented period of the event.

Table 1 Statistics of the BGP Traces (RIPE rrc04 and BCNET)

Trace	Period	Minutes	Anomalous (%)	Ann./min (reg.)	Ann./min (anom.)
Code Red I	Jul. 2001	7,200	600 (8.3)	64	190.5
Nimda	Sep. 2001	8,609	1,301 (15.1)	77	194
Slammer	Jan. 2003	7,200	869 (12.1)	48	216
Moscow blackout	May 2005	7,200	240 (3.3)	76	1,978
WannaCrypt†	May 2017	11,520	5,760 (50.0)	821	762
RIPE regular	5 days	7,196	0	163	—
BCNET regular	1 day	1,440	0	57	—

†Excluded from the anomaly evaluation because its labeled period has no routing signature; its pre-event regular traffic serves as the unseen-era false-alarm test.

Table 1 summarizes the traces. Code Red I (July 2001) and Nimda (September 2001) exploited vulnerabilities of Microsoft IIS web servers, Slammer (January 2003) infected SQL servers by probing randomly generated IP addresses, and the Moscow blackout (May 2005) followed a substation failure that disconnected the MSK-IX exchange point [15]. Each event trace covers the event together with the two days before and after it, and anomaly-free traces from RIPE (five days) and BCNET (one day) provide additional regular behavior. A WannaCrypt ransomware trace (May 2017) is also available. However, its labeled four-day period shows no measurable routing signature: the median number of announcements per minute is 762 during the labeled period and 821 outside it, and the two-sample Kolmogorov–Smirnov statistic between the two periods is only 0.095. Since a detector cannot be expected to recognize an event that leaves the control plane untouched, we exclude WannaCrypt from the anomaly evaluation and use its pre-event regular traffic as an unseen-era false-alarm test.

3.2 Distribution Shift Across Events and Eras

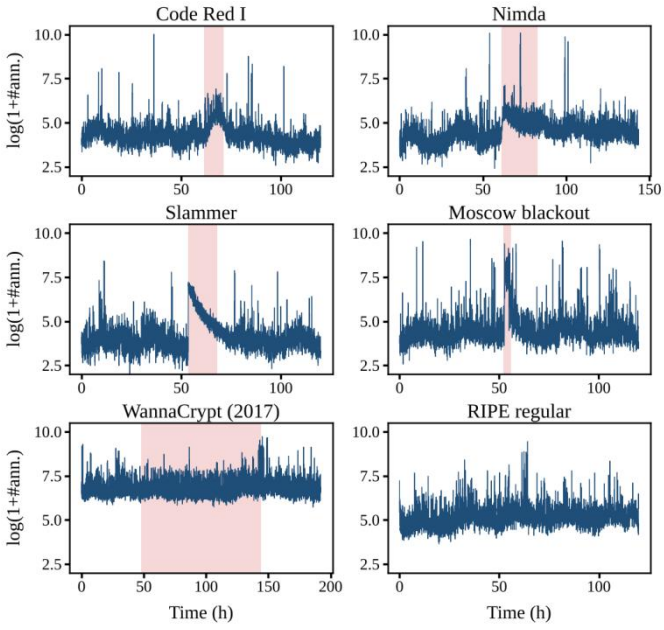


Figure 1 Announcements per Minute (Logarithmic Scale) in the RIPE rrc04 Traces
Note: Shaded areas mark the labeled anomalies. The WannaCrypt period shows no routing signature, and the baseline level differs across eras.

Figure 1 plots the number of announcements per minute for each trace, and three properties stand out. First, the anomalies are sustained: Code Red I produces a multi-hour hump, Slammer and Nimda produce a sharp surge followed by a slow decay, and the Moscow blackout produces a short but intense burst. Second, minute-level fluctuations are large relative to the anomalous elevation, so that single-minute observations are ambiguous. Third, the baseline level drifts across collectors and eras: the median announcement count of the regular RIPE trace (163 per minute) exceeds

those of the regular periods of the 2001–2005 traces (48–77 per minute), and the 2017 trace runs at 821 per minute. A detector that learns absolute thresholds from a few historical events thus risks both missing weaker anomalies and raising alarms on benign traffic growth.

3.3 Problem Formulation and Protocol

Let $\mathbf{x}_t \in \mathbb{R}^F$ (F = 37) denote the feature vector of minute t and $y_t \in \{0, 1\}$ its label. An online detector observes the causal history $\mathbf{x}_{t-W+1:t}$ and outputs a score $p_t \in [0, 1]$, and an alarm is raised when $p_t \geq 0.5$. Given the set of anomalous events $\mathcal{E}$ = {Code Red I, Nimda, Slammer, Moscow blackout}, the LOEO protocol trains a detector on $\mathcal{E} \setminus \{e\}$ together with the regular RIPE and BCNET traces and evaluates it on the unseen event e, for every $e \in \mathcal{E}$. The first day of each test trace serves only as history, and scoring starts at minute 1,440. We report the precision (P), recall (R), and F1 score of the anomalous class at the fixed threshold, together with the threshold-free average precision (AP). In addition, every LOEO model is applied to the second day of the WannaCrypt trace, which contains 1,440 regular minutes from 2017, and we report the false-positive rate (FPR) averaged over the four LOEO models as a measure of robustness to an unseen-era shift.

4 PROPOSED METHOD: PYRAGUARD

4.1 Overview

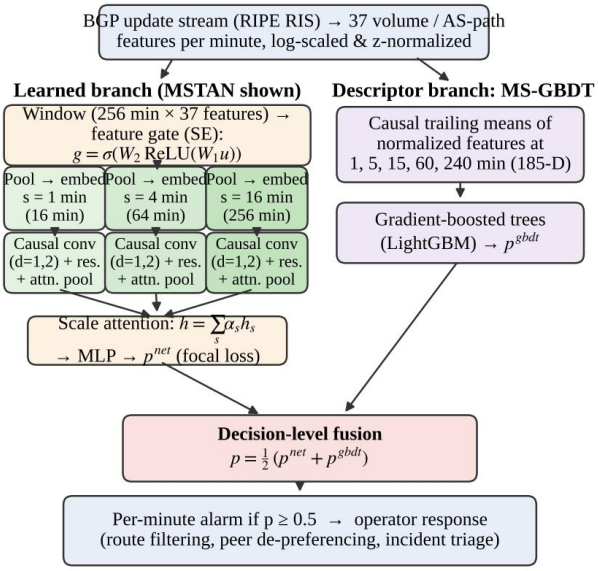


Figure 2 Architecture of PyraGuard, Shown with the Proposed MSTAN Learned Branch
Note: Any causal sequence learner, such as a GRU or a TCN, can replace MSTAN, while the descriptor branch MS-GBDT and the decision-level fusion remain unchanged.

Figure 2 depicts the architecture of PyraGuard. At every minute, the 37 features are log-transformed as $\tilde{x} = \log(1 + x)$ and standardized with the mean and standard deviation of the training traces, yielding $\hat{x}_t$. Two branches then score the current minute. The learned branch processes a causal window of recent minutes with a sequence model; we design MSTAN for this role (Section 4.2), and the framework equally accepts standard learners such as the GRU and the TCN. The descriptor branch, MS-GBDT, applies gradient-boosted trees to causal multi-scale summaries, and the two scores are fused at the decision level.

4.2 MSTAN: Multi-Scale Temporal Attention Network

1) Feature gate: BGP features are heterogeneous; some are dominated by announcement surges, whereas others reflect path exploration. Following the squeeze-and-excitation principle [21], MSTAN summarizes the window $\mathbf{X} \in \mathbb{R}^{W \times F}$ by its temporal mean and maximum and computes a gate

$$g = \sigma(W_2\, \mathrm{ReLU}(W_1\, [\mathrm{mean}_t(X)\,;\, \max_t(X)])), \qquad X' = X \odot (1\, g^T) \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, $W_1 \in \mathbb{R}^{(F/2)\times 2F}$, and $W_2 \in \mathbb{R}^{F\times(F/2)}$. The gate rescales each feature over the whole window and exposes which features drive a decision.

2) Equal-budget temporal pyramid: For the scale set $\mathcal{S}$ = {1, 4, 16} and a token budget L = 16, the branch of scale s average-pools the most recent sL minutes into L tokens,

$$z_{s,\ell} = (1/s) \sum_{j=1}^{s} x'_{t-(L-\ell+1)s+j}, \qquad \ell = 1, \ldots, L \tag{2}$$

so that the three branches observe horizons of 16, 64, and 256 minutes at resolutions of 1, 4, and 16 minutes, respectively. Assigning the same number of tokens to every scale keeps the cost of each branch constant and prevents

the finest resolution from dominating the representation, which occurs when a single full-resolution sequence is processed. Because pooling is linear, it is applied before a shared token embedding $E(z) = LN(W_e\, z + b_e)$, which reduces the computation of a coarse branch by a factor of s.

3) Intra-scale encoder: Each scale applies two causal convolutions with kernel size 3 and dilations 1 and 2, a residual connection, and layer normalization, $H_s = LN(E_s + \varphi(C_2\, \varphi(C_1\, E_s)))$, where φ is the GELU activation. An additive attention layer then pools the L tokens,

$$a_{s,\ell} = \mathrm{softmax}_\ell(v_s^T \tanh(U_s\, h_{s,\ell})), \qquad h_s = \varphi(W_s\, [\Sigma_\ell\, a_{s,\ell}\, h_{s,\ell}\, ;\, h_{s,L}]) \tag{3}$$

where the concatenation with the most recent token $h_{s,L}$ preserves the current state of each scale.

4) Scale attention: The scale representations are fused by a second attention layer with learnable scale embeddings e_s,

$$\alpha_s = \mathrm{softmax}_s(v^T \tanh(U\, (h_s + e_s))), \qquad h = \Sigma_s\, \alpha_s\, h_s \tag{4}$$

and a two-layer perceptron maps h to the logit of the score p_t^{net}. The weights α_s indicate which horizon supports a decision (Section 6.5).

5) Training objective: Anomalous minutes account for only 5.7%–8.8% of the training minutes in the LOEO folds. MSTAN is therefore trained with the focal loss [22],

$$\mathscr{L} = -(1/N)\, \Sigma_i\, \alpha_i\, (1 - q_i)^\gamma\, \log\, q_i \tag{5}$$

where q_i is the predicted probability of the true class of window i, and α_i equals α for anomalous and $1 - \alpha$ for regular windows. We use $\gamma = 2$ and $\alpha = 0.75$, which down-weights the abundant, easily classified regular minutes.

4.3 MS-GBDT: Multi-Scale Descriptor Boosting

The second branch encodes multi-scale context explicitly. For every horizon $k \in \mathcal{K} = \{1, 5, 15, 60, 240\}$ minutes, it computes the causal trailing mean of the standardized features,

$$m_t^{(k)} = (1/k)\, \Sigma_{j=0}^{k-1}\, \hat{x}_{t-j}, \qquad d_t = [m_t^{(1)}\, ;\, \ldots\, ;\, m_t^{(240)}] \in \mathbb{R}^{185} \tag{6}$$

and a LightGBM classifier maps d_t to the score p_t^{gbdt}. Gradient-boosted trees learn axis-aligned thresholds on individual multi-scale levels [23], such as the hourly mean of announced prefixes, which transfer well to events whose anomalous elevation is moderate but persistent.

4.4 Decision-Level Fusion

$$p_t = \tfrac{1}{2}\, (p_t^{net} + p_t^{gbdt}), \qquad \hat{y}_t = \mathbb{1}[p_t \geq 0.5] \tag{7}$$

The two branches err in different ways. A learned sequence branch captures the shapes of surges and decays and responds sharply to abrupt events, whereas MS-GBDT relies on level thresholds and tolerates slow ramps. Averaging the two scores allows a confident branch to override an uncertain one and suppresses isolated false alarms that only one branch produces. Fusion requires no additional training and adds negligible latency.

4.5 Complexity

MSTAN has 72,138 trainable parameters. Its cost per decision is dominated by the embedding and convolution of $3L = 48$ pooled tokens and is independent of the window length, so that longer horizons come almost for free. The descriptor branch updates its trailing means in O(1) time per minute with running sums. Both branches are causal and process one minute at a time, which matches the granularity at which BGP features are produced.

5 EXPERIMENTAL SETUP

5.1 Baselines

We compare PyraGuard and its branches with seven baselines that represent the main model families used for BGP anomaly detection. The GRU and TCN baselines also serve as alternative learned branches, which yields the variants PyraGuard-GRU and PyraGuard-TCN, whereas PyraGuard-MSTAN uses the proposed MSTAN. The per-minute baselines are an SVM with a radial basis function kernel [8, 24], a random forest (RF) [25], and LightGBM [23], following the classical feature-based approach [6-7]. The deep sequence baselines are trained on the same windows as MSTAN: a GRU with 64 hidden units, a three-layer one-dimensional convolutional neural network (CNN) with global average and max pooling [9, 26], a TCN with five residual blocks and dilations from 1 to 16 [17], and a two-layer Transformer encoder with 8-minute patch tokens [18]. The GRU uses a 64-minute window, the TCN a 128-minute window that matches its receptive field, and the CNN and Transformer the same 256-minute window as MSTAN.

5.2 Implementation Details

All deep models are implemented in JAX and trained with AdamW at a learning rate of 2×10^{-3} with cosine decay, a weight decay of 10^{-4}, gradient clipping at 1.0, a batch size of 256, and 200 optimization steps on windows sampled with a stride of two minutes. The baselines are trained with the binary cross-entropy loss, and MSTAN uses an embedding size of 48. The SVM is fitted on 8,000 randomly sampled training minutes, the RF uses 100 trees, and both LightGBM models use 300 trees, a learning rate of 0.05, 31 leaves, and row and feature subsampling ratios of 0.8. The decision threshold is fixed at 0.5 for all methods (the sign of the decision function for the SVM). The SVM, RF, LightGBM,

MS-GBDT, GRU, and MSTAN models are run with three random seeds, and the CNN, TCN, and Transformer models and all ablation variants with two seeds; every PyraGuard variant fuses the learned and MS-GBDT models of the same seed. All experiments run on a single CPU core.

6 RESULTS AND DISCUSSION

6.1 Detection of Unseen Events

Table 2 F1 Score (%) on Each Unseen Event under the LOEO Protocol (Mean ± Std over Seeds)

Method	Code Red I	Nimda	Slammer	Moscow	Mean
SVM	55.9±1.9	50.5±1.4	64.8±1.8	1.5±0.6	43.2±0.6
RF	57.0±0.4	58.3±0.4	75.1±1.1	12.9±4.4	50.8±1.1
LightGBM	60.3±0.7	63.5±0.4	82.3±0.3	34.9±0.3	60.2±0.1
CNN	33.4±0.5	38.2±3.2	37.9±26.4	25.7±25.7	33.8±0.9
Transformer	57.3±12.4	25.9±10.5	90.1±5.9	63.1±1.1	59.1±7.5
TCN	74.5±0.7	72.7±3.2	91.2±0.7	78.0±0.8	79.1±0.6
GRU	73.5±0.6	70.9±2.7	95.3±0.8	**84.2±2.7**	80.9±0.3
MSTAN (ours)	46.8±23.2	71.3±7.4	71.9±30.4	71.2±8.8	65.3±8.1
MS-GBDT (ours)	82.9±0.4	69.3±0.2	88.3±4.2	75.7±0.7	79.1±1.0
PyraGuard-MSTAN	**83.4±1.5**	72.2±1.0	94.6±0.9	76.8±0.4	81.8±0.8
PyraGuard-TCN	82.4±0.5	74.1±0.1	**96.8±0.8**	77.3±0.2	82.6±0.0
PyraGuard-GRU	82.6±0.4	**74.8±1.8**	96.1±0.6	81.3±0.6	**83.7±0.4**

Table 2 reports the F1 score on each held-out event, and Table 3 summarizes the precision, recall, AP, false-positive rate on 2017 traffic, model size, and inference time. Three observations stand out. First, per-minute classifiers transfer poorly to unseen events. LightGBM, the best of them, reaches a mean F1 score of only 60.2%, and the SVM, RF, and LightGBM all collapse on the Moscow blackout (at most 34.9%), whose short and intense burst resembles no training event when it is observed one minute at a time. Second, temporal context is essential but not sufficient: the GRU and the TCN reach 80.9% and 79.1%, whereas the CNN and the Transformer remain at 33.8% and 59.1%, which shows that generic sequence models do not automatically extract transferable cues from three training events. Third, multi-scale context transfers remarkably well even without deep learning: feeding the same LightGBM learner with causal descriptors over horizons from 1 to 240 minutes raises the mean F1 score from 60.2% to 79.1% and the AP from 69.8% to 84.6%.

Table 3 Mean Detection Performance (%), False-Positive Rate on Unseen-Era 2017 Traffic (%), Model Size, and Inference Time per Minute

Method	P	R	F1	AP	FPR_{2017}	Params	ms/min
SVM	68.8	31.9	43.2	60.9	0.00	—	0.053
RF	73.7	39.8	50.8	67.8	0.72	100 trees	0.008
LightGBM	77.6	50.8	60.2	69.8	0.83	300 trees	0.012
CNN	69.1	26.9	33.8	61.6	0.00	16.3k	0.111
Transformer	80.0	51.8	59.1	66.6	0.08	53.9k	0.195
TCN	86.4	73.3	79.1	85.0	0.00	19.4k	0.243
GRU	89.2	74.6	80.9	84.2	0.00	19.6k	0.192
MSTAN (ours)	82.8	59.4	65.3	80.3	0.00	72.1k	0.142
MS-GBDT (ours)	90.1	72.4	79.1	84.6	0.00	300 trees	0.007
PyraGuard-MSTAN	92.2	75.0	81.8	85.2	0.00	72.1k+GBDT	0.148
PyraGuard-TCN	91.3	76.7	82.6	**87.7**	0.00	19.4k+GBDT	0.249
PyraGuard-GRU	96.6	74.9	**83.7**	87.1	0.00	19.6k+GBDT	0.198

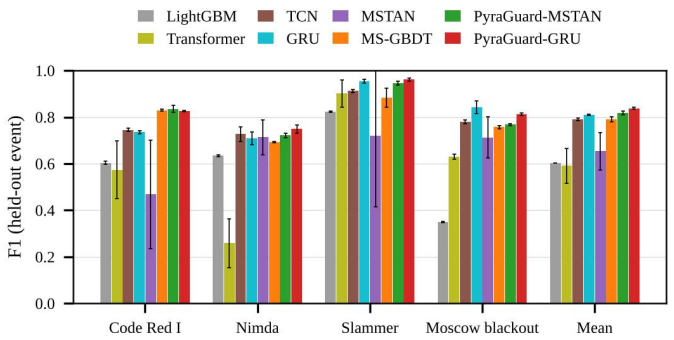


Figure 3 Per-Event F1 Scores under the LOEO Protocol (Mean and Standard Deviation over Seeds)

PyraGuard combines these strengths. With a GRU learned branch, it attains the best mean F1 score (83.7% ± 0.4), which is 2.8 points above the strongest stand-alone model and 4.7 points above MS-GBDT, at a precision of 96.6% and

a recall of 74.9%. PyraGuard-TCN achieves the highest AP (87.7%), and PyraGuard-MSTAN raises its learned branch from 65.3% to 81.8%. Fusion also stabilizes the learned branch: the seed-to-seed standard deviation of the mean F1 score falls from 8.1 points for MSTAN to 0.8 points for PyraGuard-MSTAN. Figure 3 visualizes the per-event results.

6.2 Complementarity of the Two Branches

The per-event results explain why the fusion is effective. MS-GBDT dominates on Code Red I (82.9% versus 73.5% for the GRU), whose anomaly is a slow multi-hour ramp that long-horizon level thresholds capture well. The learned branches dominate on abrupt events: the GRU reaches 95.3% on Slammer and 84.2% on the Moscow blackout, compared with 88.3% and 75.7% for MS-GBDT. Figure 4 shows the corresponding score trajectories. On Slammer, the MS-GBDT score temporarily falls below the threshold during the event while the GRU score remains high, and the fused score keeps the alarm active; after the Nimda period, isolated GRU spikes are attenuated by the low MS-GBDT score. As a result, PyraGuard-GRU attains 82.6% on Code Red I and 96.1% on Slammer. Nimda remains the most difficult event for all methods (the best F1 score is 74.8%), because the slow decay that follows its onset is only weakly visible in the features.

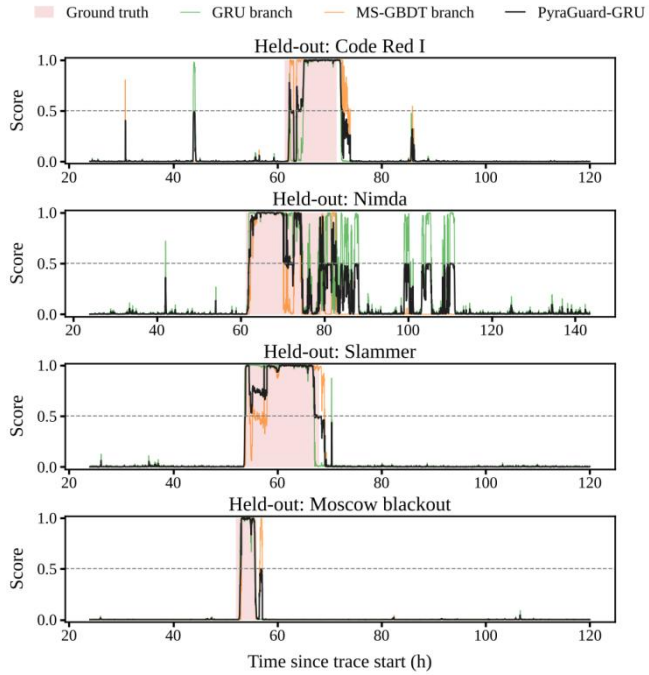


Figure 4 Score Trajectories on Each Held-Out Event (Seed 0) for the GRU Branch, the MS-GBDT Branch, and PyraGuard-GRU

Note: The dashed line marks the alarm threshold, and the shading marks the ground truth.

6.3 Ablation Study

Table 4 Ablation and Fusion Analysis (Mean F1, %, Seeds 0 and 1); ΔF1 Is Relative to the Full PyraGuard-MSTAN

Variant	Branch F1	Fused F1	ΔF1
Learned branch (MSTAN)			
Full PyraGuard-MSTAN	67.5	81.5	0.0
w/o feature gate	77.7	81.7	+0.2
w/o temporal pyramid (single scale)	61.7	80.0	−1.5
w/o focal loss (BCE)	71.8	82.7	+1.2
Descriptor branch (MS-GBDT horizons)			
Horizons up to 60 min	74.6	78.4	−3.2
Horizons up to 15 min	58.4	68.4	−13.2
Fusion rule			
Maximum	—	81.6	+0.1
Geometric mean	—	73.3	−8.2
Choice of learned branch			
GRU	80.8	83.9	+2.4
TCN	79.1	82.6	+1.1
CNN	33.8	71.5	−10.0
Transformer	59.1	75.6	−5.9

Table 4 analyzes the components of PyraGuard on seeds 0 and 1. The temporal pyramid is the key element of MSTAN: replacing it with a single full-resolution sequence over the same 256-minute horizon lowers the stand-alone F1 score

from 67.5% to 61.7% and the fused F1 score from 81.5% to 80.0%, and the single-scale model drops to 29.3% on the Moscow blackout. Removing the feature gate leaves the fused F1 score essentially unchanged (81.7%), so the gate provides interpretability at no cost to the fused detector. Training the learned branch with the binary cross-entropy instead of the focal loss reaches 82.7%, which indicates that re-weighting hard examples is unnecessary once the descriptor branch supplies stable level evidence.

The descriptor horizons matter even more. Restricting MS-GBDT to horizons of up to 60 or 15 minutes reduces its F1 score from 79.2% to 74.6% and 58.4%, and the fused F1 score decreases accordingly. Among the fusion rules, averaging and the maximum perform similarly (81.5% and 81.6%), whereas the geometric mean (73.3%) lets the more conservative branch veto correct alarms. Finally, the choice of the learned branch confirms that the fusion benefits every competent learner: the GRU and TCN branches yield 83.9% and 82.6%, whereas the weak CNN and Transformer branches (33.8% and 59.1% alone) pull the fused score below that of MS-GBDT alone.

6.4 Robustness to Unseen-Era Traffic

The second day of the 2017 WannaCrypt trace contains only regular traffic, but its median of 821 announcements per minute is about 5.0 times that of the regular RIPE trace. All PyraGuard variants and all other models except the RF, LightGBM, and the Transformer raise no false alarm on these 1,440 minutes, whereas these three models flag 0.72%, 0.83%, and 0.08% of them. The long horizons of the descriptor branch are essential for this robustness: MS-GBDT restricted to horizons of up to 15 minutes flags 47.07% of the 2017 minutes, whereas the full descriptor raises no alarm. Short-horizon statistics are dominated by the higher churn of the 2017 Internet, whereas descriptors that include hour-scale horizons separate sustained anomalous elevations from benign high-volume traffic. For operators, this property translates directly into fewer spurious incident tickets as traffic patterns evolve.

6.5 Interpretability

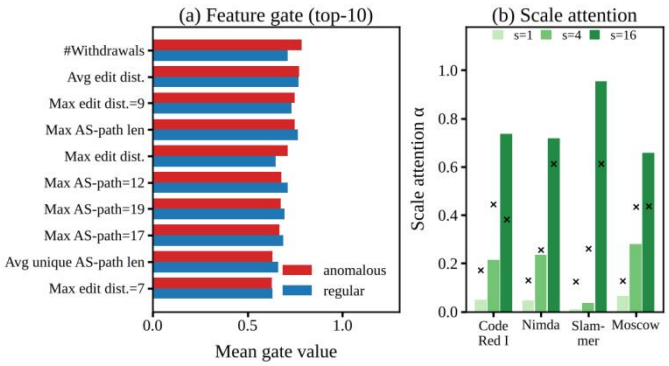


Figure 5 Interpretability of MSTAN on Held-Out Events: (a) Mean Gate Values of the Ten Most Strongly Gated Features on Anomalous and Regular Minutes; (b) Scale-Attention Weights on Anomalous Minutes (Bars) and Regular Minutes (Crosses)

MSTAN exposes the evidence behind its decisions. Figure 5(a) shows that the feature gate emphasizes withdrawals and path-exploration features, namely the edit distances between AS paths and the maximum AS-path length, and that the gates of withdrawals and of the maximum edit distance rise further on anomalous minutes (0.78 vs. 0.71 for withdrawals and 0.71 vs. 0.65 for the maximum edit distance). In contrast, the raw announcement count, the arrival rate, and the number of EGP-origin updates receive the smallest gates (0.23–0.24), which is consistent with the strong dependence of volume levels on the collector and the era observed in Figure 1. Figure 5(b) shows that the scale attention shifts toward the coarsest scale on anomalous minutes for every event; for example, the weight of the 256-minute horizon rises from 0.61 on regular minutes to 0.95 on anomalous minutes of Slammer and from 0.38 to 0.74 on Code Red I. These patterns corroborate the central finding of this study: transferable evidence of routing anomalies resides in sustained, hour-scale deviations of withdrawal and path-exploration activity rather than in minute-level announcement volume.

6.6 Efficiency and Deployment

All models are lightweight. On a single CPU core, MS-GBDT scores a minute in 0.007 ms, the GRU in 0.192 ms, and MSTAN in 0.142 ms (amortized over batches), so that PyraGuard-GRU requires about 0.20 ms per decision, more than five orders of magnitude below the one-minute granularity of the features. The learned branches have between 19,393 and 72,138 parameters, and training one branch for a LOEO fold takes 15–59 s. PyraGuard can therefore run next to a route collector or inside an ISP monitoring pipeline and raise an alarm within the minute in which a sustained anomaly becomes evident, giving operators time to de-preference unstable peers, dampen route flapping, or escalate an incident before the instability propagates further.

7 CONCLUSION

This paper studied machine learning-based detection of BGP routing anomalies under a strict leave-one-event-out protocol that mirrors the operational need to recognize unseen incidents. Real RIPE traces revealed that routing anomalies are sustained deviations hidden in bursty traffic and that the baseline level of the update stream drifts across collectors and eras. Building on these observations, we proposed PyraGuard, a hybrid multi-scale temporal learning framework that fuses a learned sequence branch with MS-GBDT, a gradient-boosted model on causal multi-scale descriptors, and we designed MSTAN, an interpretable temporal-pyramid attention network for the learned branch. Multi-scale descriptors raised the mean F1 score of a gradient-boosted detector by 18.8 points, decision-level fusion improved every competent learner, and PyraGuard-GRU reached a mean F1 score of 83.7% and an AP of 87.1% on unseen events without any false alarm on unseen-era 2017 traffic. Future work will extend the evaluation to prefix hijacks and route leaks, combine multiple route collectors, and adapt the detector online as the Internet evolves.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Rekhter Y, Li T, Hares S. A Border Gateway Protocol 4 (BGP-4). Internet Engineering Task Force (IETF), RFC 4271, 2006.
[2] Butler K, Farley T R, McDaniel P, et al. A survey of BGP security issues and solutions. Proceedings of the IEEE, 2010, 98(1): 100-122.
[3] Al-Musawi B, Branch P, Armitage G. BGP anomaly detection techniques: A survey. IEEE Communications Surveys & Tutorials, 2017, 19(1): 377-396.
[4] Sermpezis P, Kotronis V, Gigis P, et al. ARTEMIS: Neutralizing BGP hijacking within a minute. IEEE/ACM Transactions on Networking, 2018, 26(6): 2471-2486.
[5] Shapira T, Shavitt Y. AP2Vec: An unsupervised approach for BGP hijacking detection. IEEE Transactions on Network and Service Management, 2022, 19(3): 2255-2268.
[6] Al-Rousan N M, Trajković L. Machine learning models for classification of BGP anomalies. In: Proceedings of the IEEE 13th International Conference on High Performance Switching and Routing (HPSR), Belgrade, Serbia, 2012: 103-108.
[7] Ding Q, Li Z, Batta P, et al. Detecting BGP anomalies using machine learning techniques. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 2016.
[8] Batta P, Singh M, Li Z, et al. Evaluation of support vector machine kernels for detecting network anomalies. In: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018.
[9] Li Z, Gonzalez Rios A L, Trajković L. Detecting Internet worms, ransomware, and blackouts using recurrent neural networks. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020: 2165-2172.
[10] Li Z, Gonzalez Rios A L, Trajković L. Machine learning for detecting anomalies and intrusions in communication networks. IEEE Journal on Selected Areas in Communications, 2021, 39(7): 2254-2264.
[11] Cheng M, Xu Q, Lv J, et al. MS-LSTM: A multi-scale LSTM model for BGP anomaly detection. In: Proceedings of the IEEE 24th International Conference on Network Protocols (ICNP), Singapore, 2016: 1-6.
[12] Cheng M, Li Q, Lv J, et al. Multi-scale LSTM model for BGP anomaly classification. IEEE Transactions on Services Computing, 2021, 14(3): 765-778.
[13] Dong Y, Li Q, Sinnott R O, et al. ISP self-operated BGP anomaly detection based on weakly supervised learning. In: Proceedings of the IEEE 29th International Conference on Network Protocols (ICNP), Dallas, TX, USA, 2021: 1-11.
[14] Peng S, Nie J, Shu X, et al. A multi-view framework for BGP anomaly detection via graph attention network. Computer Networks, 2022, 214: 109129.
[15] Simon Fraser University Communication Networks Laboratory. Border Gateway Protocol (BGP) routing records from Reseaux IP Europeens (RIPE) and BCNET. IEEE DataPort, 2021. DOI: 10.21227/98aa-sh66.
[16] Li Z, Gonzalez Rios A L, Xu G, et al. Machine learning techniques for classifying network anomalies and intrusions. In: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 2019.
[17] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271, 2018.
[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS), 2017, 30: 5998-6008.
[19] Liu S, Yu H, Liao C, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In: Proceedings of the International Conference on Learning Representations (ICLR), 2022.
[20] Xu J, Wu H, Wang J, et al. Anomaly Transformer: Time series anomaly detection with association discrepancy. In: Proceedings of the International Conference on Learning Representations (ICLR), 2022.
[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 7132-7141.

[22] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2980-2988.
[23] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems (NIPS), 2017, 30: 3149-3157.
[24] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27:1-27:27.
[25] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.
[26] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014: 1724-1734.