https://doi.org/10.31891/2219-9365-2026-85-33
UDC 004.738

KYRYK Marian
Lviv Polytechnic National University
https://orcid.org/0000-0001-9156-9347
e-mail: marian.i.kyryk@lpnu.ua

MARUNIAK Stanislav
Lviv Polytechnic National University
https://orcid.org/0009-0006-0635-512X
e-mail: stanislav.t.maruniak@lpnu.ua

RIY Andriy
Lviv Polytechnic National University
https://orcid.org/0009-0005-0252-1533
e-mail: andrii.i.rii@lpnu.ua

A PERTURBATION-BASED XAI APPROACH FOR CLASS-SPECIFIC FEATURE SENSITIVITY ANALYSIS IN BGP ANOMALY CLASSIFICATION

Міждоменна маршрутизація відіграє центральну роль у забезпеченні глобальної зв'язності Інтернету, однак аномальна поведінка BGP залишається складною для надійної класифікації в експлуатаційних умовах. Хоча моделі машинного навчання часто демонструють високі узагальнені показники точності, їхня ефективність нерівномірна для різних типів аномалій. Для окремих класів характерні стабільно низькі значення повноти або систематична плутанина з близькими категоріями, тоді як агреговані метрики приховують ці недоліки. З практичної точки зору така класозалежна нестабільність обмежує придатність автоматизованих класифікаторів.

У статті запропоновано підхід до уточнення класифікації, керований аналізом чутливості, який дозволяє покращити деградовані класи без модифікації архітектури моделі. Метод ґрунтується на контрольованому оклюдуванні окремих ознак маршрутизації в часових вхідних послідовностях. Вимірюючи зміни ймовірностей передбаченого та істинного класів після вилучення ознаки, підхід дає змогу виявити входи, що негативно впливають на розпізнавання конкретного класу. На відміну від глобальних рейтингів важливості, аналіз обмежується неправильно класифікованими сегментами найслабшої категорії, що забезпечує цільову діагностику помилок класифікації.

Підхід оцінено на основі класифікатора LSTM із двома різними наборами ознак, сформованими з історичних подій BGP. Поділ за подіями забезпечив тестування на раніше невідомих аномаліях. У першій конфігурації аналіз оклюзії виявив ознаку, яка суттєво пригнічувала розпізнавання відмов; її вилучення підвищило повноту майже з нульових значень до 0,64 та збільшило F1-міру до 0,78 без погіршення результатів для інших класів. У другій конфігурації уточнення призвело до помірного, але стабільного зменшення плутанини між відмовами та непрямими аномаліями.

Отримані результати свідчать, що аналіз чутливості на основі збурень може слугувати не лише інтерпретаційним механізмом, а й практичним інструментом для покласового покращення багатокласових систем класифікації аномалій BGP.

Keywords: anomaly detection, BGP, machine learning, XAI, perturbation-based.

КИРИК Мар'ян, МАРУНЯК Станіслав, РІЙ Андрій
Національний університет «Львівська Політехніка»

ПОЯСНЮВАЛЬНИЙ ПІДХІД НА ОСНОВІ ЗБУРЕНЬ ДЛЯ ПОКЛАСОВОГО АНАЛІЗУ ЧУТЛИВОСТІ ОЗНАК У ЗАДАЧІ КЛАСИФІКАЦІЇ АНОМАЛІЙ BGP

Міждоменна маршрутизація відіграє центральну роль у забезпеченні глобальної зв'язності Інтернету, однак аномальна поведінка BGP залишається складною для надійної класифікації в експлуатаційних умовах. Хоча моделі машинного навчання часто демонструють високі узагальнені показники точності, їхня ефективність нерівномірна для різних типів аномалій. Для окремих класів характерні стабільно низькі значення повноти або систематична плутанина з близькими категоріями, тоді як агреговані метрики приховують ці недоліки. З практичної точки зору така класозалежна нестабільність обмежує придатність автоматизованих класифікаторів.

У статті запропоновано підхід до уточнення класифікації, керований аналізом чутливості, який дозволяє покращити деградовані класи без модифікації архітектури моделі. Метод ґрунтується на контрольованому оклюдуванні окремих ознак маршрутизації в часових вхідних послідовностях. Вимірюючи зміни ймовірностей передбаченого та істинного класів після вилучення ознаки, підхід дає змогу виявити входи, що негативно впливають на розпізнавання конкретного класу. На відміну від глобальних рейтингів важливості, аналіз обмежується неправильно класифікованими сегментами найслабшої категорії, що забезпечує цільову діагностику помилок класифікації.

Підхід оцінено на основі класифікатора LSTM із двома різними наборами ознак, сформованими з історичних подій BGP. Поділ за подіями забезпечив тестування на раніше невідомих аномаліях. У першій конфігурації аналіз оклюзії виявив ознаку, яка суттєво пригнічувала розпізнавання відмов; її вилучення підвищило повноту майже з нульових значень до 0,64 та збільшило F1-міру до 0,78 без погіршення результатів для інших класів. У другій конфігурації уточнення призвело до помірного, але стабільного зменшення плутанини між відмовами та непрямими аномаліями.

Отримані результати свідчать, що аналіз чутливості на основі збурень може слугувати не лише інтерпретаційним механізмом, а й практичним інструментом для покласового покращення багатокласових систем класифікації аномалій BGP.

Ключові слова: виявлення аномалій, BGP, машинне навчання, XAI, теорія збурень.

Стаття надійшла до редакції / Received 02.01.2026
Прийнята до друку / Accepted 17.02.2026
Опубліковано / Published 05.03.2026

INTRODUCTION

Interdomain routing is one of those Internet mechanisms that usually stays invisible – until it fails. The Border Gateway Protocol (BGP) is the de facto standard that connects autonomous systems and enables global reachability [1]. At the same time, BGP operates in an environment where configuration errors, unexpected routing dynamics, and deliberate attacks can all produce abnormal behavior. In practice, such events range from large bursts of instability to incidents that reroute or blackhole traffic, leading to service disruption and security risks. Long-running measurement studies have shown that routing instability is not a rare edge case: it can be widespread, persistent, and costly in terms of convergence and network performance [2].

Because the operational response depends on the type of incident, the task is not only to detect anomalies but also to classify them. A class label is often what determines the next action: filtering, de-preferencing, contacting peers, or activating incident playbooks. Machine learning has therefore become a common choice for building automated classifiers on top of BGP update streams and derived features, including approaches that learn from update dynamics and event patterns [3]. However, strong average metrics do not always translate into reliable behavior in real deployments. It is common to observe that some anomaly categories are handled well while others remain "difficult" (e.g., lower recall, systematic confusion with neighboring classes). When this happens, the model's overall score hides the practical issue: operators still face the same high-risk cases, such class-dependent instability limits the usefulness of automated classifiers.

This makes explainability relevant for more than transparency alone. Explanations are useful when they help diagnose where the model is fragile and why particular classes degrade. Many works provide only aggregate performance numbers, fewer attempt to explain model behavior in a way that supports debugging and targeted improvements. In the broader machine learning domain, eXplainable artificial intelligence (XAI) methods have been developed to address this challenge by enhancing trust and facilitating model refinement through the systematic analysis of prediction errors rather than merely successful outcomes [4]. For BGP anomaly classification, this suggests a clear direction: explanations should support class-specific analysis, so that weak classes can be investigated and improved in a principled way rather than treated as an unavoidable limitation.

RELATED WORK

Research on abnormal behavior in BGP routing started long before machine learning became common in networking. Early measurement studies showed that routing instability was not an occasional phenomenon but a recurring issue affecting convergence and traffic delivery [2]. Since BGP propagates incremental updates between autonomous systems [1], anomalies tend to appear as unusual patterns in update messages: sudden bursts [5], rapid withdrawals, changes in AS-path structure, or unexpected prefix announcements. These observable effects naturally led researchers to focus on measurable routing characteristics as the basis for anomaly detection.

The availability of public collectors such as RIPE RIS [6] and RouteViews [7], together with tools like BGPStream [8], made it possible to process historical and live update streams at scale. As a result, many works began extracting statistical features from BGP updates. These typically include update rates, AS-path length statistics, prefix-level activity, and various graph-based indicators derived from interdomain topology. Surveys of the field confirm that most detection systems rely on such feature representations, even when the downstream model differs [9-10]. Earlier approaches often applied heuristic thresholds or statistical deviation analysis [11], while later studies incorporated supervised learning methods to improve robustness and reduce false alarms.

Machine learning-based classification of routing anomalies gradually became a separate line of research. Instead of simply identifying "abnormal" behavior, these systems aim to distinguish between types of events, such as outages, worms, or configuration errors. Classical models, including support vector machines and decision-tree ensembles, were among the first to demonstrate that engineered routing features can separate anomaly categories with reasonable accuracy [12]. More recently, deep learning architectures have been introduced to capture temporal dependencies in routing data. Recurrent neural networks and Long Short-Term Memory (LSTM) based models, originally proposed for sequence modeling tasks [13], have been adapted to BGP time-series analysis and often report strong aggregate metrics [14]. In most cases, however, the overall workflow remains similar: feature extraction followed by supervised classification.

Operational routing security tools also show why classification quality matters in practice. Systems such as Argus [15] and HEAP [16] were designed to detect and verify hijacking events quickly, often under strict timing constraints. In these scenarios, the type of anomaly affects what operators do next. A misclassified hijack may trigger unnecessary mitigation, while a missed outage may delay corrective actions. From this perspective, classification errors are not just statistical deviations – they directly influence operational decisions. This suggests that global performance metrics alone may not provide a complete picture of a model's reliability.

As models grew more complex, interest in explainability increased across many domains of machine learning [18-22]. Methods such as Local Interpretable Model-agnostic Explanations (LIME) [23] and SHapley Additive exPlanations (SHAP) [24] were introduced to help interpret model predictions by estimating feature contributions. In networking research, these approaches have been used to analyze which routing features influence classification decisions. In our previous work, a SHAP-based evaluation was performed to quantify feature importance in BGP anomaly detection models, both globally and across anomaly classes [26]. The analysis provided useful insights into dominant and marginal features and helped assess the overall structure of the model.

Feature importance analysis helps identify which inputs contribute most strongly to model decisions, but this view is still aggregated. In real experiments, model behavior is rarely uniform across classes. Some anomaly categories remain consistently easier to recognize, whereas others show lower recall or frequent confusion with similar patterns. Knowing that a feature has high overall importance does not automatically explain why one specific class is unstable. The question is less about which feature is globally dominant and more about how model behavior changes when performance drops for particular anomaly types. As a result, there is still limited understanding of the mechanisms behind class-dependent weaknesses in BGP anomaly classification systems. This gap motivates further investigation into methods that allow more targeted analysis of classification reliability across anomaly types.

FORMULATION OF THE ARTICLE'S OBJECTIVES

The aim of this research is to develop an explainable approach for targeted improvement of BGP anomaly classification models in cases where a specific anomaly class demonstrates reduced predictive performance. The proposed approach is grounded in explainable AI principles, using perturbation-based analysis not only for interpretation but also for guided model refinement. The study focuses on post-training analysis of misclassified or weakly classified classes and on identifying routing features that negatively influence their recognition. To achieve this aim, the following objectives are defined:

(1) to detect anomaly classes with consistently lower classification scores;

(2) to apply a class-oriented explainable analysis method to evaluate the sensitivity of model predictions to individual routing features;

(3) to perform selective feature removal based on the obtained explanations and retrain the model;

(4) to assess whether the explanation-guided modification improves the target class without degrading other classes.

PRESENTATION OF MAIN MATERIAL

Most approaches to BGP anomaly classification rely on a similar general idea. Raw routing updates are first transformed into numerical indicators that summarize behavior within short time intervals. In many studies, this transformation is implemented as an explicit feature-extraction stage, which converts BGP update streams into structured vectors suitable for machine learning [27]. Other works construct features derived from routing topology or AS-relationship graphs before applying a supervised classifier [28]. These feature representations serve as the input layer for various modeling approaches.

Formally, the input to the model is not a single value but a short sequence of feature vectors extracted over consecutive time windows. Let T denote the number of time steps in one observation segment and F the number of extracted routing features. For each time step $t \in \{1, \dots, T\}$, a feature vector $x_t = \{f_1^{(t)}, \dots, f_F^{(t)}\}$ is computed. The full model input can therefore be viewed as a sequence $= \{x_1, \dots, x_T\}$, which in practice corresponds to a matrix of size $T \times F$. Each column represents the temporal evolution of one routing feature, while each row corresponds to a particular observation window. Each sequence X is labeled according to the routing condition it represents (e.g., non-anomaly, direct anomaly, indirect anomaly, outage). After training, the classifier produces a probability distribution over these classes for any unseen input segment.

Performance is evaluated using standard classification measures derived from the confusion matrix. These include precision, recall, accuracy and F1-score:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1\,score = 2\frac{Precision * Recall}{Precision + Recall} \tag{4}$$

where TP (true positives) denotes correctly identified anomaly instances, TN (true negatives) corresponds to correctly classified normal instances, FP (false positives) represents normal instances incorrectly classified as anomalies, and FN (false negatives) denotes anomalies that were not detected by the model.

In multi-class settings, these measures are computed separately for each anomaly category and may also be averaged. They quantify how often the model correctly identifies a class and how reliably it avoids misclassification. In the majority of published works, such metrics are reported in order to compare different models or feature extraction strategies under a fixed dataset and experimental configuration. The emphasis is typically placed on demonstrating competitive or improved values relative to other approaches. Still, metric tables alone do not reveal why certain anomaly types are consistently harder to recognize, or what aspects of the input representation contribute to lower scores.

One practical way to investigate this issue is to observe how model predictions change when parts of the input are deliberately modified. Perturbation-based explanation methods follow this principle by systematically altering selected input components and measuring the resulting variation in output confidence. In computer vision, occlusion has been widely used to identify input regions that influence classification decisions by masking parts of an image and tracking changes in predicted probabilities [29]. Similar perturbation-based reasoning has been formalized as a general explanation strategy in machine learning [30]. Inspired by this idea, we apply a controlled occlusion procedure to routing feature sequences in order to analyze class-specific prediction behavior.

From an operational standpoint, the classifier operates not on isolated one-minute observations, but on short segments formed by combining multiple consecutive time windows into a single input sequence. Each segment therefore reflects routing behavior over a continuous interval rather than a single snapshot. Performance metrics are computed at the level of such segments. To analyze class-specific behavior, we modify the input sequence by occluding one routing feature f_i. This is done by setting its values to zero across the entire segment, effectively removing its temporal signal while keeping all other features unchanged. The modified sequence is then passed through the trained classifier, and the output probabilities are recalculated.

Let $p_{\hat{c}}$ denote the probability assigned by the model to the originally predicted class $\hat{c}$, and let p_c denote the probability of the true class c. After occluding feature f_i, the new probabilities are denoted as $p_{\hat{c}}^{(-i)}$ and $p_c^{(-i)}$, respectively. Two quantities are evaluated:

$$\Delta^{(i)}_{pred\_drop} = p_{\hat{c}} - p_{\hat{c}}^{(-i)} \tag{5}$$

$$\Delta^{(i)}_{target\_gain} = p_c^{(-i)} - p_c \tag{6}$$

The first term measures how strongly feature f_i supports the originally predicted decision. A large positive value indicates that removing the feature substantially weakens that prediction. The second term measures whether occluding f_i increases the probability of the true class. A positive value suggests that the feature may suppress correct recognition.

Based on the aggregated sensitivity measures over misclassified segments of the target class, features are selected according to the following criterion:

$$C = \left\{ i\colon \Delta^{(i)}_{target_{gain}} > \frac{1}{M}\sum_{j}^{M} \Delta^{(j)}_{target_{gain}} \cap \Delta^{(i)}_{pred_{drop}} \leq \Delta^{(j)}_{target_{gain}} + \varepsilon \right\} \tag{7}$$

where $\varepsilon > 0$ is a small tolerance constant introduced to allow minor numerical deviations between the prediction drop and the corresponding target gain. This formulation selects features that provide positive and above-average contribution to the target class while excluding those whose removal leads to disproportionately larger changes in model output.

These quantities are computed for samples belonging to the anomaly category with the weakest metric values, and particularly for segments that were misclassified. By aggregating $\Delta^{(i)}_{pred\_drop}$ and $\Delta^{(i)}_{target\_gain}$ over such segments, we obtain a class-oriented view of how individual routing features influence incorrect decisions. Unlike global feature-importance methods, this analysis is restricted to a specific anomaly type and focuses directly on the model behavior responsible for degraded metric values. Taken together, these considerations lead to a sensitivity-driven refinement approach, summarized in Fig. 1. The procedure starts with training a baseline classifier using the complete set of extracted routing features. Once trained, class-wise precision, recall, and F1-scores are computed on the evaluation dataset. The anomaly category exhibiting the weakest performance is selected as the target for further analysis.

The previously described occlusion-based sensitivity measures are then computed for this class, focusing on misclassified segments. Based on the aggregated values of $\Delta^{(i)}_{pred\_drop}$ and $\Delta^{(i)}_{target\_gain}$, features that consistently demonstrate adverse influence on the target class are identified. Candidate features are temporarily removed from the input set, the model is retrained, and the evaluation metrics are recomputed. If the modification improves the degraded class without causing substantial deterioration in other categories, the change is retained; otherwise, the feature is restored. The resulting configuration defines the refined feature set.

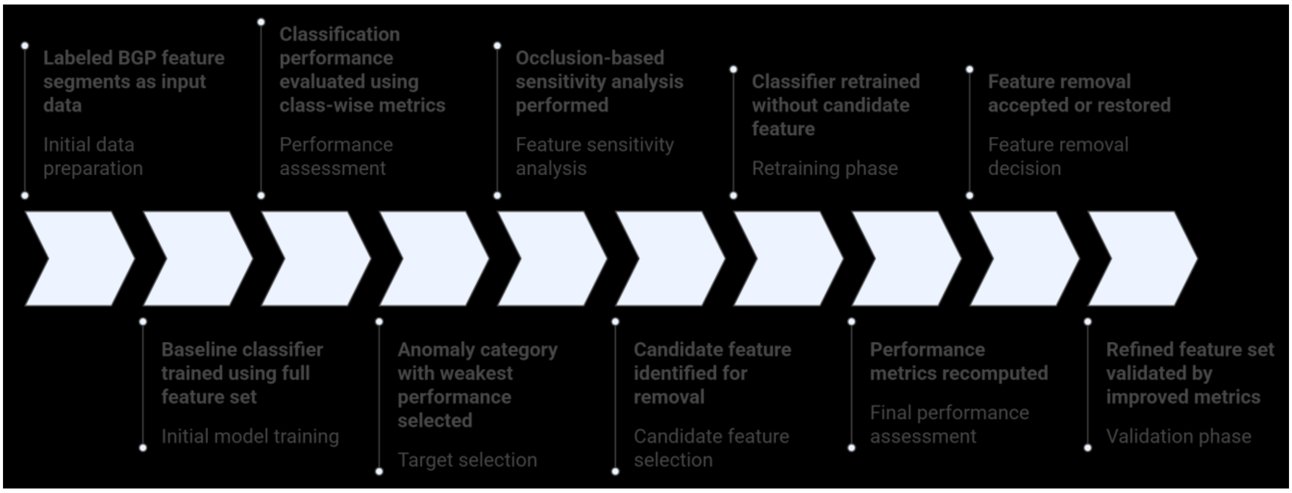

Labeled BGP feature segments as input data
Initial data preparation
Baseline classifier trained using full feature set
Initial model training
Classification performance evaluated using class-wise metrics
Performance assessment
Anomaly category with weakest performance selected
Target selection
Occlusion-based sensitivity analysis performed
Feature sensitivity analysis
Candidate feature identified for removal
Candidate feature selection
Classifier retrained without candidate feature
Retraining phase
Performance metrics recomputed
Final performance assessment
Feature removal accepted or restored
Feature removal decision
Refined feature set validated by improved metrics
Validation phase

Fig. 1. Workflow of the proposed sensitivity-driven refinement approach

To evaluate the proposed approach, a set of well-documented historical BGP anomaly events was selected [28]. The events represent different categories of routing disruptions, including direct anomalies (e.g., route leaks), indirect anomalies caused by large-scale worms, and link failures triggered by infrastructure outages. Using diverse anomaly types ensures that the evaluation does not focus on a single behavioral pattern. The anomalous events considered in this study are listed in Table 1.

Table 1

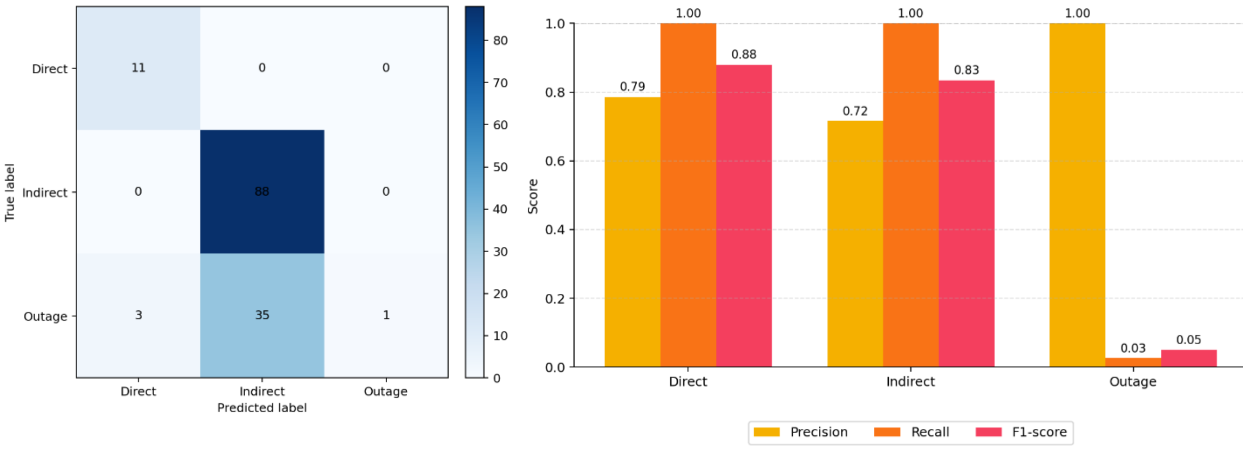
Anomalous events considered in the experimental evaluation

Event	Anomaly Type	Start Time (UTC)	Finish Time (UTC)
AS9121 RTL	Direct	24/12/2004 09:20	24/12/2004 10:03
AWS Route Leak	Direct	22/04/2016 17:10	22/04/2016 20:00
Malaysian Telecom	Direct	12/06/2015 08:42	12/06/2015 10:24
Code Red v2	Indirect	19/07/2001 10:00	19/07/2001 20:00
Nimda	Indirect	18/07/2001 13:00	21/07/2001 12:00
Slammer	Indirect	25/01/2003 05:31	25/01/2003 19:59
Moscow Blackout	Outage	25/05/2005 04:40	25/05/2005 07:40
Japan Earthquake	Outage	11/03/2011 09:13	11/03/2011 15:39

For each event, raw BGP update data were collected for 24 hours before and 24 hours after the anomaly interval. This extended observation period ensures that both normal and anomalous routing behavior are represented in the dataset. All updates were processed using a fixed temporal resolution of 1 minute. For each minute, a routing feature vector was computed, and a corresponding label was assigned according to the routing condition at that time (normal behavior or a specific anomaly type).

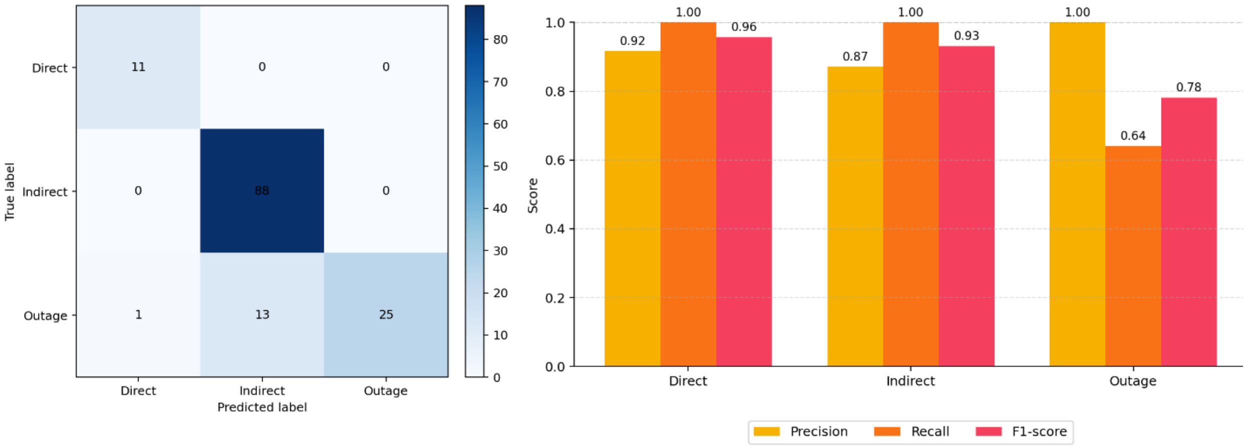
Two independent feature extraction approaches were considered in order to evaluate the robustness of the proposed refinement strategy. First, a statistical feature set consisting of 32 indicators was generated using the BML framework [31]. These features describe temporal and distributional properties of routing updates within each one-minute window. The set includes indicators related to update activity, AS-path statistics, prefix-level characteristics, and variability measures computed over the observation interval. Second, a topology-oriented feature set was constructed using an extractor based on AS-relationship graphs, following the methodology proposed by Paiva et al [28]. These features capture structural properties of routing behavior, including relationship types between ASes and graph-based characteristics derived from AS-path connectivity. The sensitivity-driven refinement procedure described earlier was applied independently to both feature representations using LSTM-driven classifier. An event-based split was adopted to ensure evaluation on previously unseen anomaly events.

Baseline classification results obtained using the full set of 32 BML statistical features are shown in Fig. 2. While direct and indirect anomaly categories are classified with high accuracy, the outage class exhibits significantly degraded performance. In particular, recall for the outage category remains close to zero, indicating that most outage segments are misclassified as other anomaly types. Given this imbalance, the outage class is selected as the target for the subsequent sensitivity analysis and refinement procedure.



a) b)

Fig. 2. Baseline results (BML features): a) confusion matrix; b) metrics



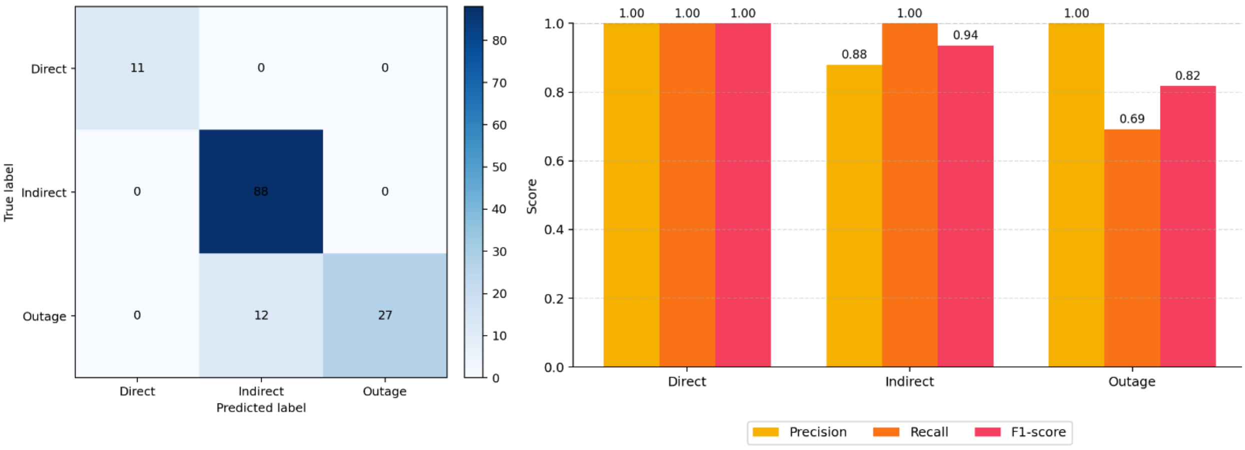
a) b)

Fig. 3. Classification performance after removal of avg_editdist: a) confusion matrix; b) metrics

Among all 32 statistical features, only *avg_editdist* exhibited a substantial positive target gain ($\Delta^{(i)}_{target\_gain}$ ≈ 0.71) when occluded, indicating that this feature strongly contributes to incorrect suppression of the outage class. All remaining features showed marginal effects, with gain values close to zero (within ±0.02). In contrast, the feature *avg_interarrival* demonstrated a high $\Delta^{(i)}_{pred\_drop}$ value (≈ 0.89), meaning that removing it significantly reduces the confidence of the originally predicted class. This behavior suggests that the feature plays an important structural role in the classifier and is not a candidate for removal. After removing the feature *avg_editdist*, the classifier was retrained using the remaining 31 statistical features. The updated performance is shown in Fig. 3.

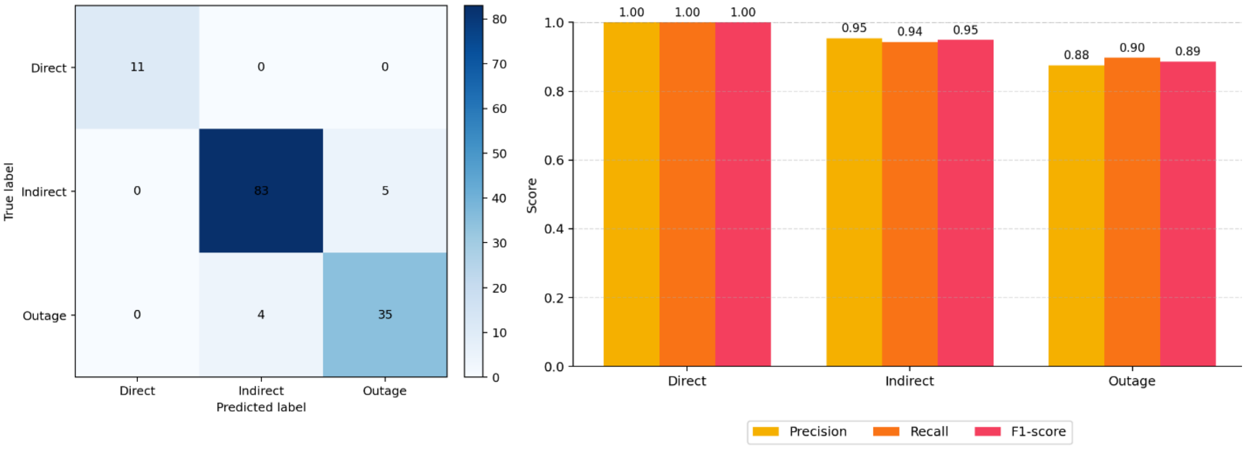
A substantial improvement is observed for the outage class. Recall increases from near-zero in the baseline configuration to 0.64, and the F1-score rises to 0.78. The confusion matrix in Fig. 4(a) shows that the majority of outage segments are now correctly classified (25 instances), with a reduced number of misclassifications as indirect anomalies. In addition to the improvement of the degraded class, performance for other classes also increases. Both categories achieve higher precision and F1-scores compared to the baseline model, indicating that the removed feature negatively influenced overall class separation rather than benefiting other classes. These results demonstrate that the sensitivity-driven refinement not only recovers the degraded class but also enhances global classification consistency.

To further evaluate the generality of the proposed refinement approach, the same experimental procedure was applied to an alternative feature representation based on AS relationship graphs. Unlike the statistical BML features, this topology-oriented set captures structural properties of routing behavior. Baseline results for the topology-based features are shown in Fig. 4. Compared to the statistical configuration, the outage class is recognized more reliably at baseline, with substantially higher recall and F1-score. This suggests that structural routing characteristics are more informative for distinguishing outage-type anomalies. Nevertheless, this category remains the weakest among the three classes, indicating that misclassification patterns are still present and that further sensitivity analysis is warranted.



a) **b)**

Fig. 4. Baseline classification results using topology-based features: (a) confusion matrix; (b) metrics



a) **b)**

Fig. 5. Classification results after removal of av_number_of_bits_in_prefix_ipv4: (a) confusion matrix; (b) metrics

For the topology-based representation, occlusion analysis again focused on misclassified outage segments. The feature *av_number_of_bits_in_prefix_ipv4* demonstrated the highest target gain (≈ 0.98), indicating strong influence on suppression of the outage class. At the same time, it produced a substantial prediction drop(≈ 0.99), suggesting that it plays a significant role in overall model decisions. Given this dual behavior, the feature was temporarily removed to evaluate its structural impact. After removal, a slight redistribution between the indirect and outage classes is observed, reducing their mutual confusion. The improvement is moderate compared to the statistical feature experiment, as the topology-based configuration already provides stronger baseline discrimination of outage anomalies (Fig. 5).

The experimental evaluation across two distinct feature representations demonstrates that the proposed sensitivity-driven refinement approach provides a systematic mechanism for analyzing and correcting class-specific weaknesses in anomaly classification. In the statistical feature configuration, the method revealed a strongly distortive feature whose removal substantially improved outage recognition. In the topology-based representation, the same procedure resulted in a more moderate but still measurable adjustment, confirming that the approach adapts to different feature characteristics. These observations indicate that perturbation-based sensitivity analysis can serve not only as an explanatory tool, but also as a practical instrument for targeted feature configuration refinement.

To ensure result stability, the classification pipeline (including feature extraction, model training, and sensitivity analysis) was repeated 3 times with different random seeds. The identified adverse features (*avg_editdist* and *av_number_of_bits_in_prefix_ipv4*) and performance improvements were consistent across runs, confirming the robustness of the refinement procedure.

CONCLUSIONS

In this study, a perturbation-based sensitivity approach was applied to improve recognition of degraded anomaly categories in BGP classification. The method was evaluated using two different feature sets, statistical features and topology-based features, under the same LSTM model and event-based testing protocol. In both configurations, the analysis identified features that strongly influenced misclassification of the outage category. Their removal followed by retraining resulted in improved recall and F1 score for the problematic class.

The results confirm that occlusion based analysis can be used not only to interpret model behavior but also to guide practical refinement of the feature set. Unlike global feature importance methods (e.g., SHAP) that provide aggregate rankings across all classes, the proposed perturbation-based analysis is inherently class-specific and causal: it directly quantifies how feature removal affects the probability of the true class for misclassified instances of the weakest category, enabling targeted refinement rather than generic interpretation. The approach operates independently of the specific feature representation and does not require architectural changes, which makes it suitable for integration into existing multi-class BGP anomaly detection pipelines.

References

1. Rekhter Y. A Border Gateway Protocol 4 (BGP-4) / Y. Rekhter, T. Li, S. Hares // RFC 4271. – RFC Editor. – January 2006. https://doi.org/10.17487/RFC4271

2. Labovitz C. Internet routing instability / C. Labovitz, G. R. Malan, F. Jahanian // Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. – 1997. – P. 115–126. https://doi.org/10.1145/263105.263151

3. Zhang J. Learning-based anomaly detection in BGP updates / J. Zhang, J. Rexford, N. Feamster // Proceedings of the ACM SIGCOMM Workshop on Mining Network Data (MineNet '05). – 2005. – P. 219–224. https://doi.org/10.1145/1080173.1080178

4. Ribeiro M. T. "Why Should I Trust You?" Explaining the predictions of any classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 1135–1144. https://doi.org/10.1145/2939672.2939778

5. Moriano P. Using bursty announcements for early detection of BGP routing anomalies [Online] / P. Moriano, R. Hill, L. Camp // arXiv preprint. – 2019. https://doi.org/10.48550/arXiv.1905.05835

6. RIPE NCC. Routing Information Service (RIS) [Online]. – 1999. – Available at: https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris – Accessed: January 20, 2026.

7. RouteViews. University of Oregon RouteViews Project [Online]. – 2013. – Available at: http://www.routeviews.org – Accessed: January 20, 2026.

8. Orsini C. BGPStream: A software framework for live and historical BGP data analysis / C. Orsini, A. King, D. Giordano, V. Giotsas, A. Dainotti // Proceedings of the 2016 Internet Measurement Conference (IMC '16). – New York, NY, USA: Association for Computing Machinery, 2016. – P. 429–444. https://doi.org/10.1145/2987443.2987482

9. Al-Musawi B. BGP anomaly detection techniques: a survey / B. Al-Musawi, P. Branch, G. Armitage // IEEE Communications Surveys & Tutorials. – 2016. – Vol. PP. – P. 1–1. https://doi.org/10.1109/COMST.2016.2622240

10. Hammood N. A survey of BGP anomaly detection using machine learning techniques / N. Hammood, B. Al-Musawi, A. Alhilali // In: – 2022. – P. 109–120. – ISBN 978-981-19-1165-1. https://doi.org/10.1007/978-981-19-1166-8_9

11. Prakash B. A. BGP-lens: patterns and anomalies in internet routing updates / B. A. Prakash, N. Valler, D. Andersen, M. Faloutsos, C. Faloutsos // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09). – New York, NY, USA: Association for Computing Machinery, 2009. – P. 1315–1324. https://doi.org/10.1145/1557019.1557160

12. Al-Rousan N. Machine learning models for classification of BGP anomalies / N. Al-Rousan, L. Trajkovic // 2012 IEEE 13th International Conference on High Performance Switching and Routing (HPSR 2012). – 2012. – P. 103–108. – ISBN 978-1-4577-0831-2. https://doi.org/10.1109/HPSR.2012.6260835

13. Hochreiter S. Long short-term memory / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – Vol. 9, No 8. – P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

14. Cheng M. Multi-scale LSTM model for BGP anomaly classification / M. Cheng, Q. Li, J. Lv, W. Liu, J. Wang // IEEE Transactions on Services Computing. – 2021. – Vol. 14, No 3. – P. 765–778. https://doi.org/10.1109/TSC.2018.2824809

15. Xiang Y. Argus: An accurate and agile system to detecting IP prefix hijacking / Y. Xiang, Z. Wang, X. Yin, J. Wu // Proceedings of the International Conference on Network Protocols (ICNP). – 2011. – P. 43–48. https://doi.org/10.1109/ICNP.2011.6089080

16. Schlamp J. HEAP: Reliable assessment of BGP hijacking attacks / J. Schlamp, R. Holz, Q. Jacquemart, G. Carle, E. W. Biersack // IEEE Journal on Selected Areas in Communications. – 2016. – Vol. 34. – P. 1849–1861.

17. Guidotti R. A survey of methods for explaining black box models / R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi // ACM Computing Surveys. – 2018. – Vol. 51, No 5. – Article 93. https://doi.org/10.1145/3236009

18. Ribeiro M. T. "Why Should I Trust You?": Explaining the predictions of any classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). – New York, NY, USA: Association for Computing Machinery, 2016. – P. 1135–1144. https://doi.org/10.1145/2939672.2939778

19. Kyryk M. The potential of the isolation forest method for anomaly detection in network traffic / M. Kyryk, N. Pleskanka, A. Riy // Herald of Khmelnytskyi National University. Technical Sciences. – 2025. – Vol. 349, No 2. – P. 171–177. https://doi.org/10.31891/2307-5732-2025-349-25

20. Shen C. Border Gateway Protocol route leak detection technique based on graph features and machine learning / C. Shen, R. Wang, X. Li, P. Zhang, K. Liu, L. Tan // Electronics. – 2024. – Vol. 13. – Article No 4072. https://doi.org/10.3390/electronics13204072

21. Arreche O. XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems / O. Arreche, T. Guntur, M. Abdallah // Applied Sciences. – 2024. – Vol. 14. – Article No 4170. https://doi.org/10.3390/app14104170

22. Robertson D. Modelling BGP updates for anomaly detection using machine learning [Online] / D. Robertson // Wellington Faculty of Engineering Symposium. – 2023. – Available at: https://ojs.victoria.ac.nz/wfes/article/view/8368 – Accessed: January 21, 2026.

23. Lundberg S. M. A unified approach to interpreting model predictions / S. M. Lundberg, S.-I. Lee // Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17). – Red Hook, NY, USA: Curran Associates Inc., 2017. – P. 4768–4777.

24. Samek W. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models / W. Samek, T. Wiegand, K.-R. Müller // ITU Journal: ICT Discoveries – Special Issue 1: The Impact of Artificial Intelligence (AI) on Communication Networks and Services. – 2017. – Vol. 1. – P. 1–10. https://doi.org/10.48550/arXiv.1708.08296

25. Molnar C. Interpretable machine learning: A guide for making black box models explainable [Elektronnyi resurs] / C. Molnar. – 3rd ed. – 2025. – Available at: https://christophm.github.io/interpretable-ml-book/

26. Kyryk M. SHAP-based evaluation of feature importance in BGP anomaly detection models / M. Kyryk, S. Maruniak, T. Andrukhiv // ICTEE. – 2025. – Vol. 5, No 1. – P. 34–43. https://doi.org/10.23939/ictee2025.01.034

27. Fonseca P. BGP dataset generation and feature extraction for anomaly detection / P. Fonseca, E. Mota, R. Bennesby, A. Passito // 2019 IEEE Symposium on Computers and Communications (ISCC). – 2019. – P. 1–6. https://doi.org/10.1109/ISCC47284.2019.8969619

28. Paiva T. BGP anomalies classification using features based on AS relationship graphs / T. Paiva, Y. Siqueira, D. Batista, R. Hirata Jr., R. Terada // 2021 IEEE Latin-American Conference on Communications (LATINCOM). – 2021. – P. 1–6. https://doi.org/10.1109/LATINCOM53176.2021.9647824

29. Zeiler M. D. Visualizing and understanding convolutional networks / M. D. Zeiler, R. Fergus // In: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds) Computer Vision – ECCV 2014. – Lecture Notes in Computer Science. – Vol. 8689. – Cham: Springer, 2014. https://doi.org/10.1007/978-3-319-10590-1_53

30. Fong R. Interpretable explanations of black boxes by meaningful perturbation / R. Fong, A. Vedaldi // arXiv preprint. – 2017. https://doi.org/10.48550/arXiv.1704.03296

31. Hoarau K. BML: An efficient and versatile tool for BGP dataset collection / K. Hoarau, P.-U. Tournoux, T. Razafindralambo // 2021 IEEE International Conference on Communications Workshops (ICC Workshops). – 2021. – P. 1–6. https://doi.org/10.1109/ICCWorkshops50388.2021.9473737