ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 5 (1), 2025

https://doi.org/10.23939/ictee2025.01.034

SHAP-BASED EVALUATION OF FEATURE IMPORTANCE IN BGP ANOMALY DETECTION MODELS

M. Kyryk[1][ORCID: 0000-0001-9156-9347], S. Maruniak[1][ORCID: 0009-0006-0635-512X], T. Andrukhiv[2]

[1] Lviv Polytechnic National University, 12, S. Bandery str., Lviv, 79013, Ukraine
[2] Lviv branch of JSC "Ukrtelecom", 3, Doroshenko str., Lviv, 79000, Ukraine

Corresponding author: S. Maruniak (e-mail: stanislav.t.maruniak@lpnu.ua)

(Received 8 January 2025)

The classification of Border Gateway Protocol (BGP) anomalies is essential for maintaining Internet stability and security, as such anomalies can impair network functionality and reliability. Previous studies has examined the impact of key features on anomaly detection; however, current methodologies frequently demonstrate high computational costs, complexity, and usage challenges. The article presents a novel approach for evaluating feature importance based on SHAP (SHapley Additive Explanations), which provides a simplified, interpretable and efficient alternative specifically designed for LSTM-based classification models. A dedicated tool was developed to effectively evaluate feature impact, combining statistical analysis with visualizations to improve comprehension. This tool enables the assessment of global feature influence across datasets, emphasizing features that consistently increase classification performance. Furthermore, it offers insights into the impact of features on a per-class basis, demonstrating the varying contributions of individual features to the detection of different types of anomalies. Various datasets representing distinct anomaly types, such as direct, indirect, and outage anomalies, were utilized to validate the approach's applicability across a range of scenarios. This level of detail enables researchers to enhance LSTM models for particular anomaly categories while preserving overall efficacy. We suggested a structured algorithm to facilitate these developments, showing how feature impact evaluation can directly improve model optimization and detection tactics. Stability tests performed on various datasets demonstrate the reliability of feature rankings, thereby reinforcing the validity of the proposed methodology. The SHAP-based framework described in this paper makes complex analyses easier to understand while also providing useful insights. This approach enhances the efficiency of anomaly detection systems by allowing researchers to identify critical features, integrate new metrics, and refine existing LSTM models. The advancements enhance the security and resilience of infocommunication networks, effectively addressing emerging challenges in network security through a scalable and interpretable solution.

Keywords: *BGP, SHAP, anomaly detection, datasets, machine learning.*
UDC: 004.7

Introduction

The Border Gateway Protocol (BGP) is a critical component of the Internet's infrastructure that allows communication between global networks. Although its open and decentralized architecture is necessary for scalability, it also makes it vulnerable to a variety of abnormalities [1]. BGP operations can be disrupted by misconfigurations, deliberate assaults, and infrastructure failures, therefore creating major connectivity and security concerns. Addressing these difficulties necessitates excellent methods for not just detecting but also classifying anomalies in order to determine their root causes and inform mitigation efforts.

Classification is critical in this circumstance because it lets network operators identify the type of abnormality. It could be a routing error, a deliberate hijack, or a disruption caused by external factors such as earthquakes or floods [2]. Recent advances in machine learning have increased the accuracy of anomaly classification by extracting a range of features from BGP update signals. These features can be constructed using traditional volume measures as well as more complicated graph-based and geographical variables. Nonetheless, the individual contributions of each feature to classification outcomes have not been properly studied.

This paper seeks to fill this gap by proposing a novel approach to measuring feature influence in BGP anomaly classification. This work uses SHAP (SHapley Additive Explanations) to measure feature impact, resulting in a lightweight and interpretable technique for both academics and practitioners. The findings demonstrate how specific features influence classification performance in both positive and negative ways, which can be beneficial for developing novel feature extraction techniques or improving models.

2. Related work

The classification of BGP anomalies has been extensively researched due to its importance in preserving the stability and security of the Internet. Anomalies were detected through the use of simple statistical metrics and volume-based indicators, such as variations in AS path length or update frequency using early techniques. While these techniques were excellent at detecting unusual behavior, they lacked the ability to classify anomalies, leaving operators without actionable information.

Machine Learning Models for Classification of BGP Anomalies [3], published in 2012, was a big step forward. This study used multiple machine learning algorithms to classify anomalies based on attributes extracted from BGP update packets. The study examined the classification of various anomalies, such as route hijacks, misconfigurations, and connection issues, using Support Vector Machines (SVMs), Decision Trees, and Hidden Markov Models (HMMs). SVMs shown remarkable efficacy in handling feature spaces that were linearly separable, whereas HMMs performed rather well in gathering and predicting temporal trends within BGP updates.

The study also proposed a feature selection procedure to boost classification accuracy. The most informative features were identified using metrics such as Fisher's score and Minimum Redundancy Maximum Relevance (mRMR). These included variations in AS path length, announcement volumes, and the frequency of unique prefixes. By refining feature sets, the models improved significantly in accuracy, laying the groundwork for future research into feature engineering for BGP anomaly classification.

Researchers built on these results in later years using more sophisticated criteria and classification systems. Graph-based characteristics comprising AS connection metrics and path topology proved better than conventional volume-based measurements, according to Fonseca et al [4]. Paiva et al. [5] then included geographic features into their models, including average AS lengths, therefore enabling more complex diagnosis of abnormalities including route leaks and connection failures.

Despite these advancements, one major challenge remains unmet: the systematic evaluation of feature value in anomaly categorization. Most studies concentrate on overall model accuracy rather than how specific factors influence performance. Traditional feature selection methods, such as ablation studies, are typically computationally expensive and inappropriate for real-time or large-scale applications. Furthermore, researchers need lightweight, interpretable tools that allow them to quickly test the usability of new features.

This paper fills these gaps by presenting a novel approach to determining feature importance called SHAP (SHapley Additive Explanations) [6]. Building on previous research approaches, this paper presents a realistic methodology for feature evaluation, allowing for a more in-depth knowledge of feature contributions to BGP anomaly categorization.

3. Data processing

BGP anomalies are classified using well-structured datasets that contain a wide range of events and routing characteristics. This study evaluates six well-documented anomalies, divided into three categories: direct anomalies, indirect anomalies, and lilnk failures, to assess the influence of different classification standards. Direct anomalies are typically caused by faulty routing arrangements [7]. For example, the AS9121 Routing Table Leak in 2004 resulted in the unexpected announcement of a large number of prefixes, upsetting global routing tables, whilst the AWS Route Leak in 2016 caused traffic to be routed along undesired paths due to incorrectly configured routing settings.

In contrast, indirect anomalies are caused by external interruptions like the spread of Internet worms. Notable examples include the Nimda Worm in 2001 and the Slammer Worm in 2003, both of which swamped routing infrastructures and caused widespread disruption. Whereas link failures are caused by physical network disturbances. Examples include the 2005 Moscow Blackout, in which a large power outage affected routing operations, and the 2011 Japanese earthquake, which caused serious connection interruptions.

Routing data from known BGP monitoring sites was used to examine these instances. RIPE RIS [8] and Route Views [9] provided historical data compiled from a network of monitoring locations spread across multiple Autonomous Systems. For large-scale abnormalities, additional data was obtained from the Center for Applied Internet Data Analysis (CAIDA), a repository for network research. The open-source tool BGPStream was used to quickly download data for the specified time periods of interest, allowing for targeted examination of the six occurrences.

The analysis depends on BGP update messages, which provide the primary data source for categorization models. These communications contain essential routing information, including AS pathways, IP prefixes, next-hop data, and routing policies. Features derived from these messages cover both conventional metrics, including announcement volumes and AS path lengths, as well as sophisticated attributes, such as AS relationship structures, degree variance, and geographical distances between Autonomous Systems. The latter were computed utilizing location databases to capture spatial dynamics relevant to routing behavior.

This study uses current datasets and extracted features taken from previous research [10] to evaluate the relevance and influence of individual aspects in the classification of BGP anomalies. The processing pipeline ensures that the datasets and features are well-structured for feature significance analysis.

4. SHAP-based feature evaluation for BGP classification

SHAP (SHapley Additive Explanations) [6] is a powerful framework for analyzing machine learning model results. It gives a consistent and fair measure of feature relevance by comparing marginal contributions across all possible feature subsets. Unlike classic methods like permutation importance or ablation studies, SHAP uses cooperative game theory ideas to assure interpretability and consistency. These properties make it especially appropriate for evaluating the complex LSTM-based (Long short-term memory) classifier used in this work for BGP anomaly identification.

The SHAP value for a feature is its contribution to the output of the model, computed as [6]:

$$\varphi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}\left[f(S \cup \{i\}) - f(S)\right], \tag{1}$$

where φ_i is the SHAP value for feature x_i, N is the set of all features, S is a subset of features excluding x_i, and $f(S)$ is the model’s prediction based on the subset S.

SHAP decomposes a model’s prediction into a sum of contributions from individual features:

$$f(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i, \tag{2}$$

where φ_0 is the base value of the model (the mean prediction across the dataset).

Moreover, by averaging the magnitude of SHAP values for every feature over all data points, the Mean Absolute SHAP Value (MASV) presents a global viewpoint on feature importance. MASV estimates the overall influence of a feature, allowing for a clear comparison of its value:

$$MASV_i = \frac{1}{n}\sum_{j=1}^{n} \left|\varphi_{i,j}\right|, \tag{3}$$

where $\varphi_{i,j}$ is the SHAP value of feature i for instance j, and N is the total number of instances.

Using SHAP, we calculated the global influence of characteristics on anomaly classification. The findings show that traditional indicators like the number of announcements remain consistently important across all anomaly classes (Fig. 1). This is expected, as anomalies frequently emerge as major changes in the number of announcements, making this a feature that is generally applicable. Metrics like maximum AS path length are also important worldwide, representing the impact of anomalies on routing routes, such as anomalous path extensions induced by misconfigurations or route leaks.

In addition to standard characteristics, unexpected factors such as geographical distances and AS path degree variance emerged as important contributors. These parameters give a contextual layer to classification by capturing structural and spatial elements that standard metrics may miss. Geographic distances, for example, can provide information about localized disruptions, whereas AS path degree variance reveals network-wide connectivity fluctuations. These additional variables serve to improve the model's predicting skills across all anomaly categories.

These findings are consistent with prior studies by Fonseca et al. [4] and Paiva et al. [5], which highlighted the usefulness of advanced measures for anomaly detection. The use of SHAP in this study expands previous research by giving a quantitative, interpretable quantification of feature impact, confirming the need of integrating traditional and innovative metrics for robust anomaly categorization.

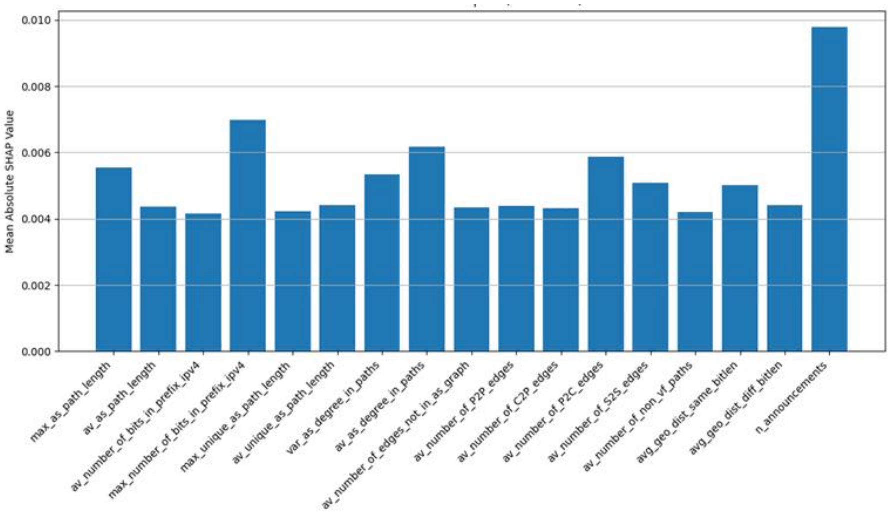

Fig. 1. Global feature impact across all anomaly classes

The influence of features varied dramatically among anomaly classes, as evidenced by SHAP values. Traditional measures like the maximum AS path length and the number of unique prefixes had the greatest impact on direct anomalies, such as misconfigurations and hijacks (Fig. 2). The maximum AS path length frequently includes significant variances caused by routing failures or malicious operations. For example, during a direct anomaly, AS pathways may become abnormally long as a result of unexpected loops or inaccurate route ads. Similarly, the number of unique prefixes reveals changes in advertising routes, which are frequently impacted by such abnormalities. These features frequently produced greater SHAP values than others, demonstrating their utility for detecting routing discrepancies that identify direct anomalies.

For indirect anomalies induced by Internet worms like Nimda and Slammer, graph-based features, specifically the number of peer-to-peer edges, were found to be more relevant (Fig. 3). These anomalies frequently disrupt the structural equilibrium of the AS network, causing anomalous changes in peer relationships as a result of rapid traffic propagation or malicious activity. The high SHAP values of peer-to-peer edges indicate their capacity to detect structural abnormalities. For example, during the Slammer worm attack, the anomaly caused widespread peer-to-peer relationship fluctuations, resulting in a greater impact for this statistic. These findings highlight the value of utilizing network structure-based features to detect disruptions caused by indirect anomalies.

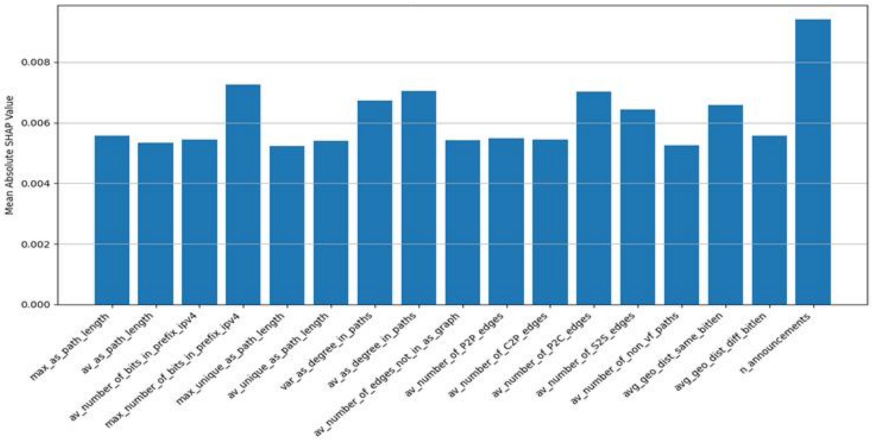

Fig. 2. Feature impact for direct anomalies

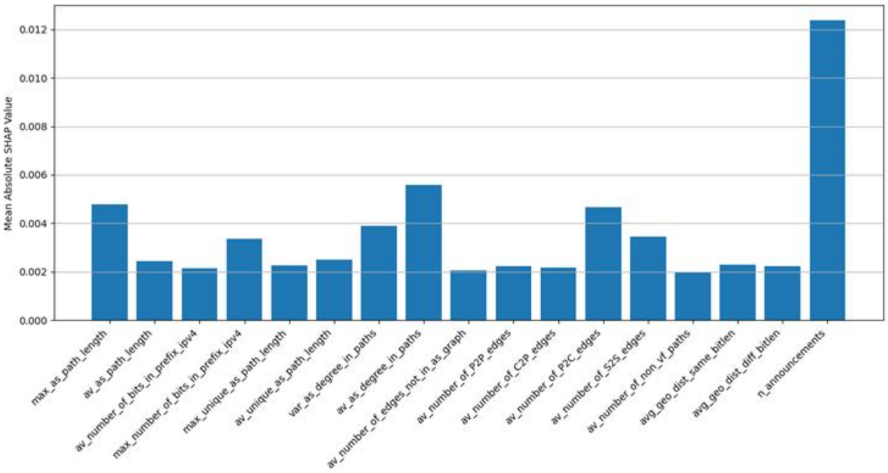

Fig. 3. Feature impact for indirect anomalies

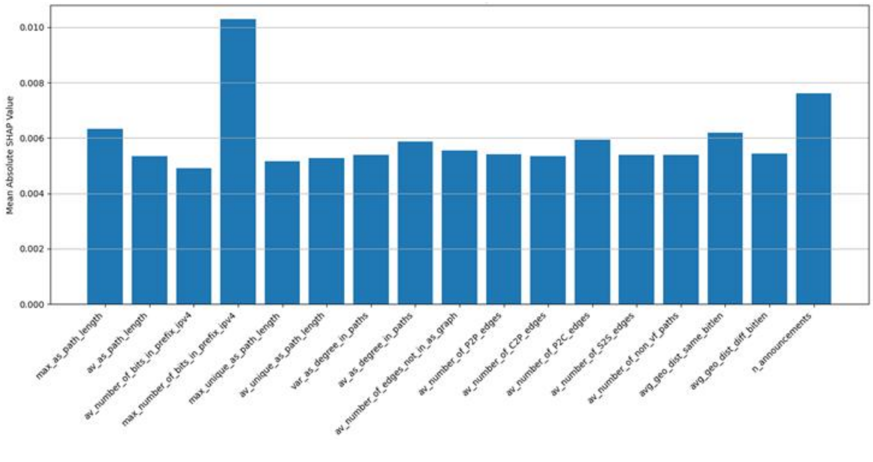

Fig. 4. Feature impact for outage anomalies

Outage anomalies followed a clear pattern, with geographical characteristics such as distance between ASes having the biggest influence (Fig. 4). Outages, such as those produced by the Japanese earthquake, frequently result in regionally limited interruptions. The SHAP study found that geographical distance indicators regularly rated higher in relevance because they accurately reflected the impact of outages on ASes in afflicted locations. Interestingly, the average AS path length has little impact on outages. This is most probable because disruptions tend to affect broader geographic areas, making individual path statistics less useful. In contrast, measures like the maximum amount of bits in the prefix had a much greater influence. Outages often disrupt larger prefix ranges, and SHAP values shown the growing relevance of this metric in classification of such events.

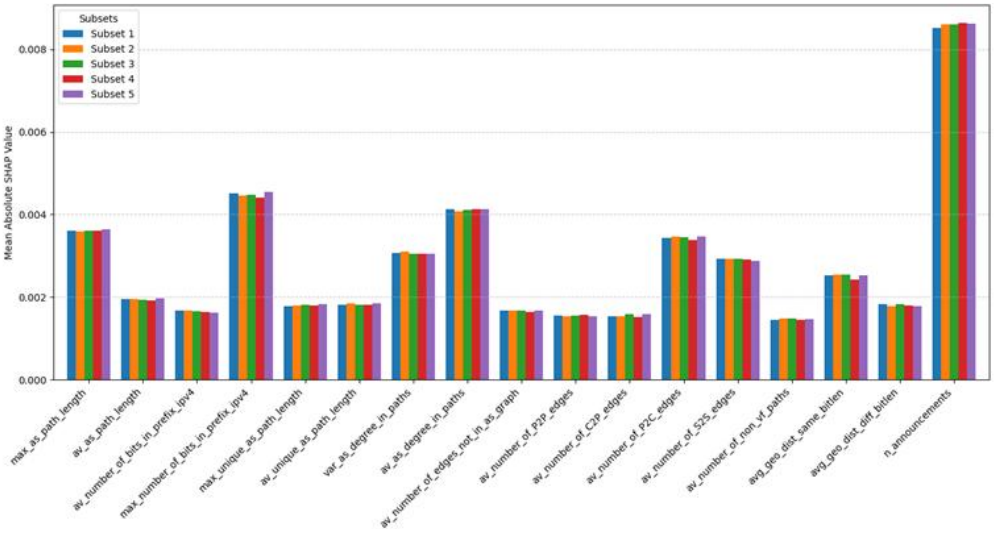

Fig. 5. Stability analysis of feature impact across random subsets

The per-class SHAP study shows overall how feature importance varies depending on type of anomaly. While graph-based features shine at identifying indirect abnormalities, traditional measures beat in cases involving direct changes. Capture of the spatial elements of outages depends on new geographical metrics. With significant features routinely surpassing others in their anomalous classes, the projected SHAP values provide a quantitative basis for these results. Stability analysis, which was performed on various randomized subsets of the dataset, validated the resilience of feature importance rankings (Fig. 5). The findings revealed that all features yielded fairly consistent rankings across subsets, with little variation in relative importance. This consistency demonstrates the trustworthiness of SHAP-based evaluations for feature impact analysis.

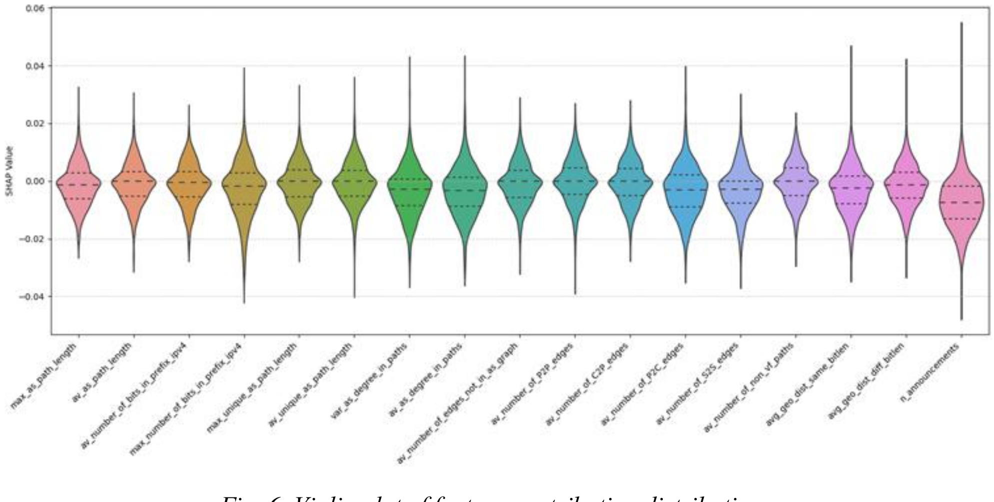

Fig. 6. Violin plot of feature contribution distributions

The variety of feature contributions was examined using violin plots, which depict the distribution of SHAP values (Fig. 6). Features such as the maximum AS path length displayed consistently high importance across instances, as evidenced by their narrow distributions around central values. This shows their general applicability over a range of anomalies, which makes them consistent indicators of disturbances. Conversely, geographic distances showed more spread and indicated that their influence changed depending on the background of the anomaly. For outages that affect certain areas, for instance, geographic distances could be more crucial, but their significance reduces in anomalies like route hijacking or less spatially dependent designs.

This diversity emphasizes the subtle relevance of such factors in classification. The larger distributions indicate that, while these features may not always be among the best generally, they do provide unique insights into certain anomaly cases. This conditional relevance emphasizes the need of adding a varied set of variables to represent the multifaceted nature of BGP anomalies, allowing models to effectively adapt to various categorization issues.

The heatmap (Fig. 7) illustrates the correlations between features and SHAP values across various anomaly categories. The maximum amount of bits in a prefix had a substantial positive connection with outages, demonstrating its efficacy in detecting disruptions to prefix-level routing. In contrast, average AS path length was found to be negatively related to outages, indicating that it is becoming less relevant in such instances.

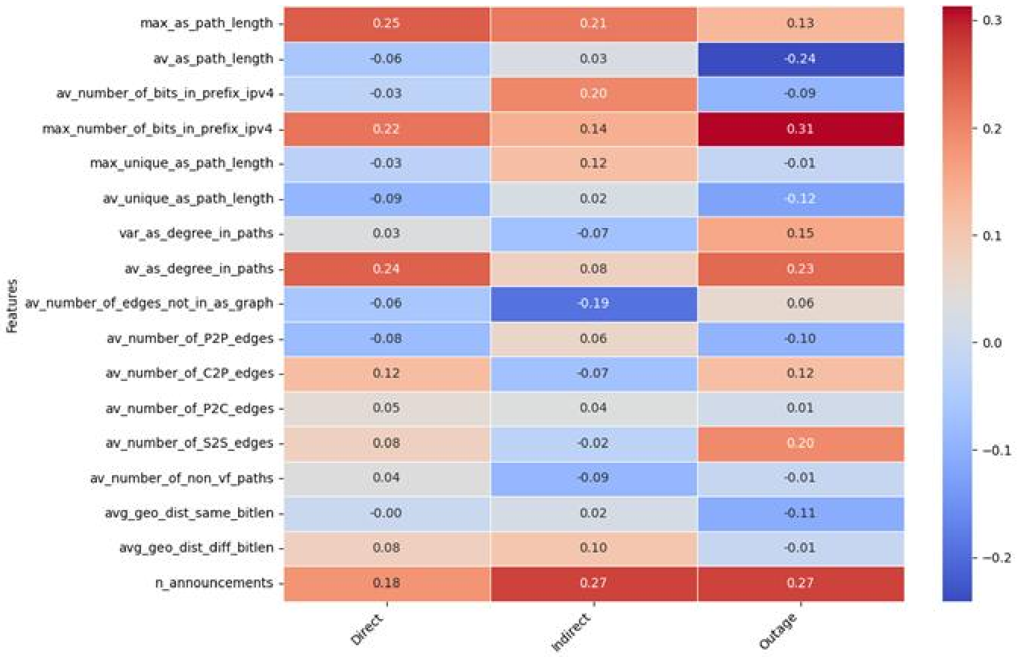

Fig. 7. Correlation heatmap of features with SHAP values across anomaly classes

These findings are consistent with the global feature rankings, but they contrast from the violin plot's conclusions for average AS path length, where the broader contribution was clearer. Such differences suggest that while average AS path length has a general utility, it may lack specific relevance in particular anomaly types like outages. The heatmap's correlations further emphasize the nuanced contributions of graph-based and geographical features, which align well with the classification of indirect anomalies and outages.

These results offer fresh understanding of how features contribute to BGP anomaly classification. While new elements including geographical distances and graph-based qualities are very essential for addressing the special characteristics of outages and structural disruptions, traditional metrics remain vital for direct and indirect anomalies. Using SHAP, this study provides a more thorough and interpretable knowledge of feature importance than other studies, therefore opening the path for improvement of anomaly classification models and direction of future feature engineering efforts.

A block scheme (Fig. 8) is provided to demonstrate the possibility of SHAP-based analysis for refining BGP anomaly classification models. This method describes a systematic strategy for incorporating SHAP evaluations into the feature selection and model improvement process. Starting with a set of BGP update messages, features are extracted and utilized to train an LSTM model. The SHAP analysis then determines the relevance of each characteristic, which guides further model improvements.

Features determined to have a negative effect are excluded, and the model training procedure is repeated. Features significantly affecting particular anomaly classes are further optimized by assigning weights to increase their impact on the classification process. The final output presents feature impact measurements, guaranteeing that only features with positive contributions are preserved in the model. This method improves model accuracy and interpretability by repeated optimization of the feature set, all while preserving computing efficiency. This method aligns with the previously reported results and offers a clear structure for incorporating SHAP into BGP anomaly detection operations, facilitating systematic feature engineering and model improvement initiatives.

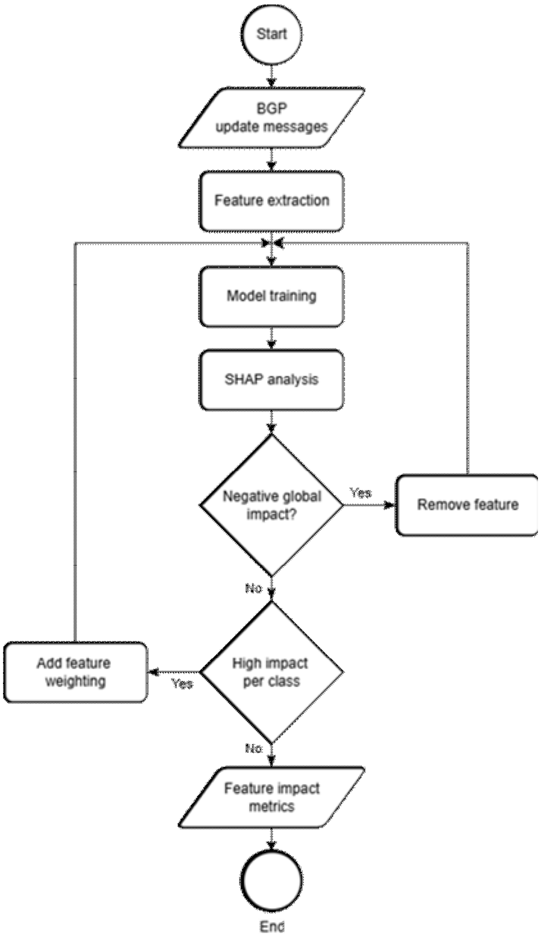

Fig. 8. SHAP-based model refinement process

Conclusions

This paper presents a new, systematic method for assessing feature significance in BGP anomaly categorization with SHAP. The research examines the contributions of different features, emphasizing their collective influence on classification performance and illustrating the consistency of feature ranks across datasets. This framework [11] provides a consistent and understandable way to clarify the relevance of particular metrics in complex models.

The outcomes offer significant advantages for next studies since they help to efficiently evaluate new features and improve low computational burden classification algorithms. This work integrates feature evaluation with model building to help to develop more strong and interpretable solutions for BGP anomaly detection and network security.

References

[1] Rekhter, Y., Li, T. and Hares, S. (2006), “A border gateway protocol 4 (BGP-4)”, Internet Requests for Comments, RFC Editor, RFC 4271, January, available at: http://www.rfc-editor.org/rfc/rfc4271.txt (Accessed 25 November 2024). DOI: 10.17487/RFC4271

[2] Hammood, N.H., Al-Musawi, B. and Alhilali, A.H. (2022), “A survey of BGP anomaly detection using machine learning techniques”, in Pokhrel, S.R., Yu, M. and Li, G. (eds) Applications and Techniques in Information Security. ATIS 2021. Communications in Computer and Information Science, vol. 1554. Springer, Singapore. DOI: 10.1007/978-981-19-1166-8_9

[3] Al-Rousan, N.M. and Trajković, L. (2012), “Machine learning models for classification of BGP anomalies”, 2012 IEEE 13th International Conference on High Performance Switching and Routing, Belgrade, Serbia, pp. 103–108. DOI: 10.1109/HPSR.2012.6260835

[4] Fonseca, P., Mota, E. S., Bennesby, R. and Passito, A. (2019), “BGP dataset generation and feature extraction for anomaly detection”, 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, pp. 1–6. DOI: 10.1109/ISCC47284.2019.8969619

[5] Paiva, T. B., Siqueira, Y., Batista, D. M., Hirata, R. and Terada, R. (2021), “BGP anomalies classification using features based on AS relationship graphs”, 2021 IEEE Latin-American Conference on Communications (LATINCOM), Santo Domingo, Dominican Republic, pp. 1–6. DOI: 10.1109/LATINCOM53176.2021.9647824

[6] Lundberg, S.M. and Lee, S.-I. (2017) “A unified approach to interpreting model predictions”, in Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774. DOI: 10.48550/arXiv.1705.07874

[7] Al-Musawi, B., Branch, P. and Armitage, G. (2017), “BGP anomaly detection techniques: A survey”, IEEE Communications Surveys & Tutorials, 19(1), pp. 377–396. DOI: 10.1109/COMST.2016.2622240

[8] RIPE (1999), “RIPE Network Coordination Centre”, available at: https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris (Accessed: 30 November 2024).

[9] RouteViews (2013), “University of Oregon RouteViews Project”, Eugene, OR., available at: http://www.routeviews.org (Accessed: 30 November 2024).

[10] Paiva, T., “BGP anomaly classification dataset”, available at: https://github.com/thalespaiva/bgp-anomaly-classification/blob/main/data.zip (accessed: 30 November 2024).

[11] Maruniak, S., “BFRank”, available at: https://github.com/MaruniakS/BFRank (Accessed: 15 December 2024).

ОЦІНКА ВПЛИВУ ОЗНАК У МОДЕЛЯХ ВИЯВЛЕННЯ АНОМАЛІЙ BGP НА ОСНОВІ SHAP

Мар’ян Кирик[1], Станіслав Маруняк[1], Тарас Андрухів[2]

[1]Національний університет “Львівська політехніка”, вул. С. Бандери, 12, 79013, Львів, Україна
[2] Львівська філія ВАТ “Укртелеком”, вул. Дорошенка, 3, Львів, 79000, Україна

Класифікація аномалій з використанням Протоколу Граничного Шлюзу (BGP) важлива для забезпечення стабільності та безпеки інтернету, оскільки такі аномалії можуть порушувати роботу та знижувати надійність мережі. У попередніх дослідженнях цієї предметної області було проаналізовано вплив базових характеристик повідомлень оновлення BGP на моделі виявлення аномалій, проте описані підходи часто використовують методи із високою обчислювальною складністю, важкі для розуміння та можуть спричиняти труднощі під час заміни наборів даних чи тренувальних моделей. У статті викладено новий підхід до оцінювання важливості характеристик на основі методів SHAP (SHapley Additive Explanations), який пропонує спрощену, зрозумілу та ефективну альтернативу, спеціально розроблену для моделей класифікації на основі LSTM. Розроблено спеціалізований інструмент для ефективної оцінки впливу характеристик, який поєднує статистичний аналіз із візуалізаціями для кращого розуміння результатів. Цей інструмент дає змогу оцінювати глобальний вплив характеристик для різних наборів даних, як позитивний, так і негативний. Крім того, він надає інформацію про вплив характеристик для кожного класу, демонструючи, як окремі характеристики по-різному впливають на виявлення різних типів аномалій. Для перевірки стабільності отриманих результатів використано набори даних, що представляють декілька типів аномалій, такі як прямі, непрямі та збої. Такий рівень деталізації дає змогу дослідникам покращувати моделі LSTM для окремих категорій аномалій, зберігаючи загальну ефективність. Запропоновано структурований алгоритм поліпшення моделей класифікації аномалій BGP із урахуванням оцінки впливу характеристик. Виконано тести стабільності на різних наборах даних для підтвердження надійності ранжування характеристик, що додатково збільшує достовірність запропонованої методології. Описаний у статті підхід підвищує ефективність систем виявлення аномалій, даючи дослідникам змогу ідентифікувати критичні характеристики, впроваджувати нові метрики та вдосконалювати наявні моделі LSTM, що, своєю чергою, дає можливість підвищити безпеку та стійкість інформаційно-комунікаційних мереж, ефективно відповідаючи на нові виклики у сфері мережевої безпеки.

Ключові слова: *BGP, SHAP, виявлення аномалій, набори даних, машинне навчання.*