Hybrid Feature Selection and Explainable Machine Learning for BGP Anomaly Detection

Shadi Motaali, Jorge E. López de Vergara, and Luis de Pedro

Dept. Electronics and Communication Technologies,
Universidad Autónoma de Madrid, Madrid, Spain
{shadi.motaali, jorge.lopez_vergara, luis.depedro}@uam.es

Abstract. Detecting anomalies in the Border Gateway Protocol (BGP) is challenging due to highly imbalanced datasets and feature selection needs. This paper introduces a hybrid feature selection approach that combines six algorithms (ANOVA F-score, Mutual Information, Random Forest Importance, XGBoost Importance, RFE, and Lasso) to identify optimal feature subsets. Unlike previous studies that rely on synthetic oversampling techniques such as SMOTE (which generates unrealistic floating-point values for inherently discrete features), we implement two balanced real-sample selection strategies that preserve data authenticity. Working with the BGP Feature Extractor dataset, our methodology identified a reduced set of 25 high-importance features appearing in at least 75% of the selection algorithm outputs. Experimental results show that our hybrid approach achieves 89.73% accuracy and 89.72% F1-score with XGBoost, outperforming full-feature models by 0.26–1%. The Random Forest classifier showed similar improvements, increasing from 80.86% to 88.00% accuracy when using the hybrid feature subset. SHAP and Gini Index analyses revealed remarkable consistency in top-ranked features across models. PCA analysis confirmed that just 22 components explain 99% of data variance for all classes, while t-SNE visualizations displayed clearer class separation with the hybrid feature subset. This research contributes a robust, explainable framework for BGP anomaly detection that maintains high accuracy while significantly reducing computational complexity.

Keywords: BGP, Anomaly Detection, Feature Selection, Random Forest, XGBoost, SHAP, Gini Importance.

1 Introduction

The Border Gateway Protocol (BGP) is the fundamental routing protocol of the Internet, enabling autonomous systems (ASes) to exchange reachability information. Despite its critical role, BGP lacks robust security mechanisms, making it vulnerable to attacks such as route hijacking, route leaks, and prefix hijacking, which can cause service disruptions and data interception.

Machine learning approaches for BGP anomaly detection face several challenges. First, BGP datasets are highly imbalanced [5], with normal traffic constituting approximately 95% of observations compared to only 5% of anomalous events. Additionally, in our initial experiments with synthetic oversampling techniques such as SMOTE, we

observed the generation of unrealistic feature values—specifically, floating-point values for inherently discrete integer attributes such as announcement counts and origin metrics—potentially compromising model training.

Second, BGP traffic data includes numerous features, many of which are redundant or irrelevant for detecting anomalies. Utilizing all available features increases computational complexity, reducing processing speed, and can introduce noise that degrades model performance. Finally, network security demands not only high detection accuracy but also interpretability, enabling administrators to understand, validate, and respond to flagged anomalies.

This paper addresses these challenges through the following contributions:

1. We propose a hybrid feature selection methodology that combines six distinct algorithms—ANOVA F-score, Mutual Information, Random Forest Importance, XGBoost Importance, Recursive Feature Elimination (RFE), and Lasso—to identify robust and informative feature subsets.
2. We implement a real-sample balancing strategy to construct a uniformly distributed dataset across all classes, avoiding the limitations of synthetic oversampling and preserving data authenticity.
3. Our comprehensive evaluation provides improved classification performance across multiple anomaly types using a compact set of 25 high-importance features that appear in at least 75% of the selected algorithms outputs.
4. We perform explainability analysis using SHAP (SHapley Additive Explanations) for XGBoost and Gini importance for Random Forest, revealing strong consistency in their top-ranked features.
5. Dimensionality reduction techniques (t-SNE and PCA) illustrate improved class separability when using the selected feature subset compared to the full feature space.
6. We performed a Pareto-efficiency analysis to evaluate the trade-off between detection performance and inference time, revealing that hyperparameter tuning improves XGBoost F1 score to 0.897, but at the cost of a 72.7% increase in inference time compared to the non-optimized configuration.

In addition to its classification benefits, the proposed hybrid feature subset significantly simplifies the data preprocessing pipeline. Reducing the feature set from 48 to 25 not only improves computational efficiency but also streamlines the extraction process from raw BGP data sources such as RIPE RIS [11] or RouteViews [14]. This is particularly beneficial for real-time deployment scenarios, where the overhead of computing less relevant features can hinder timely anomaly detection.

The remainder of this paper is organized as follows: Section 2 reviews related work in BGP anomaly detection. Section 3 describes our dataset, feature selection strategy, and classification models. Section 4 presents the experimental results. Section 5 provides an ablation study evaluating the effect of feature count and critical attributes on performance. Finally, Section 6 concludes the study and outlines directions for future research. The complete implementation is available at our GitHub repository.[1]

[1] https://github.com/shadimotaali/Hybrid-Feature-Selection-and-Explainable-Machine-Learning-for-BGP-Anomaly-Detection/tree/main

2 Related Work

To contextualize our contributions, we compare our methodology against key recent works in BGP anomaly detection.

Unlike some studies relying on simulated datasets (e.g., Nassir et al. [7]), our work uses real-world BGP data encompassing normal traffic and three anomaly classes. While their hybrid SGD-RF model achieved high accuracy (99.3%) on simulated data, its applicability to operational networks is limited.

Romo-Chavero et al. [12] applied a hybrid approach combining Median Absolute Deviation (MAD) with multiple machine learning models such as Random Forest and XGBoost, achieving high accuracy (99%) and F1-scores (98%) on major real-world events. However, they employed oversampling techniques for class balancing, which may risk overfitting and depend heavily on chosen thresholds.

Allahdadi et al. [1] proposed a three-phase pipeline with feature extraction, correlation-based feature generation, and anomaly detection via one-class SVM. Their approach yielded high accuracy (up to 98%) and low false alarm rates but was limited by the use of six event-specific training data, with generalizability challenges.

Park et al. [9] used advanced tokenization and deep learning models on real BGP data with SMOTE oversampling. While achieving excellent F1 scores (0.99), the introduction of synthetic samples raised concerns about data realism and model applicability.

In contrast, our hybrid feature selection framework employs an ensemble of six methods and adopts a real-sample balanced selection strategy to avoid unrealistic synthetic data, resulting in a robust and explainable model with balanced performance (Accuracy 89.73%, F1 89.72% using XGBoost). Additionally, we apply explainability techniques such as SHAP, Gini importance, PCA, and t-SNE to interpret the model and features. Some limitations remain, including variance across classes, particularly a partial performance drop in Class 2, related to direct attacks.

3 Methodology

This section details our comprehensive methodology for BGP anomaly detection, encompassing dataset description, preprocessing techniques, our novel balanced sampling approach, feature selection methods, and machine learning modeling strategies.

3.1 Dataset Description and Preprocessing

We utilized the publicly available dataset from the BGP Feature Extractor project [5], which contains both normal BGP traffic and various types of anomalies. The original dataset is highly imbalanced, with approximately 95% of the data representing normal traffic and only 5% representing anomalous activity. The anomalies are grouped into three categories: Class 1 (indirect attacks), Class 2 (direct attacks), and Class 3 (outages). Each data instance is described by a comprehensive set of features, including routing announcements, withdrawals, AS-path characteristics, origin information, and temporal metrics.

Our preprocessing pipeline involved several steps:

- **Initial exploration:** We examined the statistical structure of the dataset and identified Class 3 as the class with the fewest instances (1 217 samples).
- **Balanced sampling strategy:** To address the class imbalance, we avoided synthetic oversampling techniques such as SMOTE. In preliminary tests, SMOTE generated unrealistic floating-point values for inherently discrete features (e.g., announcement counts), which could distort learning. Instead, we used two balanced real-sample selection strategies:
  - **Equal sampling approach:** We randomly sampled exactly 1 217 instances from each class (Class 0 = normal, Class 1–3 = anomalies) to build a uniformly balanced dataset.
  - **Proportional sampling approach:** We sampled 1 217 instances from each anomaly class (Class 1–3) and 3 651 instances (i.e., $3 \times 1\,217$) from the normal class (Class 0) to maintain a more realistic traffic distribution while still mitigating the imbalance.
- **Dataset configurations:** Four binary classification scenarios were constructed to evaluate model performance under varying anomaly types: **Normal vs. Class1**, **Normal vs. Class2**, **Normal vs. Class3**, and **Normal vs. All Anomalies** (aggregated Class1–3).
- **Data partitioning:** Each dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve the class distribution. Although five-fold cross-validation was additionally assessed to ensure robustness, the fixed train-test split yielded more stable and superior performance across all models in our experiments.

These steps ensured the authenticity of the traffic data while addressing the imbalance without introducing synthetic artifacts.

3.2 Feature Selection Methods

Feature selection plays an important role in enhancing model performance, reducing computational complexity, and improving model interpretability. In this study, we deliberately selected six diverse feature selection methods from different algorithmic families to capture complementary perspectives on feature relevance. These are:

- **Statistical methods (ANOVA F-score):** Designed to identify features that exhibit statistically significant differences between class distributions [6].
- **Information-theoretic approaches (Mutual Information):** Used to capture both linear and non-linear dependencies between features and class labels [10].
- **Tree-based methods (Random Forest and XGBoost Importance):** These ensemble techniques evaluate feature relevance by leveraging tree structures to model complex interactions [3, 4].
- **Wrapper methods (Recursive Feature Elimination - RFE):** Iteratively remove the least important features based on model performance [2].
- **Embedded methods (Lasso - L1 Regularization):** Integrate feature selection into model training by shrinking irrelevant coefficients to zero [13].

This comprehensive selection strategy enables more robust and explainable modeling. The specific methods are summarized as follows:

- **ANOVA F-score:** A univariate statistical technique that evaluates each feature by computing the ratio of inter-class variance to intra-class variance. Features with high F-scores are considered more discriminative.
- **Mutual Information:** An information-theoretic metric that quantifies the amount of information a feature contributes about the target variable, enabling the detection of complex, non-linear associations.
- **Random Forest Importance:** Estimates feature importance based on the average decrease in Gini impurity across trees in the forest, offering insight into hierarchical feature usage.
- **XGBoost Importance:** Measures feature contribution based on the gain or frequency of usage during the construction of boosted trees in the XGBoost algorithm.
- **Recursive Feature Elimination (RFE):** A wrapper-based strategy that recursively trains a model (Random Forest in our case), ranks features and eliminates the least important until a desired subset is reached.
- **Lasso (L1 Regularization):** Applies L1 penalty during linear model training to enforce sparsity, effectively reducing the coefficients of less relevant features to zero.

3.3 Modeling Approaches

To assess the effectiveness of our selected features in detecting BGP anomalies, we employed two machine learning classifiers: Random Forest (RF), and XGBoost. These models were selected for their proven performance in prior BGP anomaly detection studies and their complementary characteristics—RF provides interpretable ensemble learning, and XGBoost is known for its accuracy and efficiency on tabular data.

For each model, we trained and evaluated two versions: one using the full feature set and one using the hybrid subset of 25 features selected through our voting-based ensemble approach. All models were trained on stratified 80% training data and evaluated on the remaining 20% test set to ensure balanced representation of all classes.

While hyperparameters were initally kept at default values to isolate feature selection effects and ensure reproducibility, we later applied hyperparameter optimization (including Grid Search, Randomized Search, and Bayesian Optimization), which revealed that model tuning could further enhance detection performance between 1% to 6% in F1 score. Notably, class-specific tuning revealed that different anomaly types benefit from distinct parameter configurations—Class 2 (direct attacks) required deeper trees and stronger regularization, while Class 3 (outages) performed optimally with moderate tree depth and lower regularization. This suggests that adaptive parameter tuning strategies could complement our hybrid feature selection approach in operational deployments.

Explainability techniques were also integrated into our modeling phase. We used SHAP (SHapley Additive exPlanations) [8] to interpret XGBoost predictions and Gini importance scores [3] for feature ranking in Random Forest. This enabled a cross-model

comparison of the semantic consistency of top-ranked features across models and confirmed the relevance of our hybrid-selected subset in practical anomaly detection scenarios.

4 Experimental Results

This section presents the outcomes of our empirical evaluation. We analyze the effectiveness of different feature selection methods, assess classifier performance across multiple BGP anomaly classes, provide insights through explainability and visualization techniques, and conduct a Pareto analysis examining the trade-offs between detection performance and inference time under both default and optimized hyperparameter configurations. The goal is to validate the robustness and interpretability of our hybrid feature selection and modeling pipeline.

4.1 Feature Selection Results

This section evaluates feature importance using four algorithms—ANOVA F-score, Mutual Information, XGBoost, and Random Forest—across three anomaly classes and the full dataset (we excluded Lasso and RFE due to their ranking inconsistencies and tendency to prioritize domain-irrelevant or minor features with limited significance for BGP anomaly detection). Our goal was to identify consistently high-performing features that contribute significantly to BGP anomaly detection while reducing dimensionality.

Figure 1 shows a binary heatmap indicating which of the top 30 features appeared in the top 10 results for a given method and class. 1 denotes presence, and 0 denotes absence in the top-10 list.

These features provide a robust description of routing dynamics, including path structure, message duplication, and prefix announcements—all of which are critical indicators in detecting anomalies in BGP traffic.

In addition to per-class rankings, we applied a hybrid voting strategy to combine the outputs from all four algorithms. From each method, the top 30 features were retained. Features that appeared in at least 3 out of 4 methods were selected as part of the final hybrid subset. This process yielded 25 highly consistent features. Figure 2 shows the frequency with which each selected feature appeared in the top 30 lists.

The consistently high-ranking features across algorithms include:

- **Path-related metrics**: `imp_wd_spath`, `unique_as_path_max`
- **Edit distance metrics**: `edit_distance_dict_1`, `edit_distance_avg`
- **Duplication indicators**: `dups`
- **Network announcement patterns**: `nlri_ann`
- **Origin stability metrics**: `origin_0`, `origin_2`

4.2 Classification Results

To evaluate the impact of feature selection on BGP anomaly classification, we trained and tested both Random Forest and XGBoost classifiers on full and hybrid feature sets.

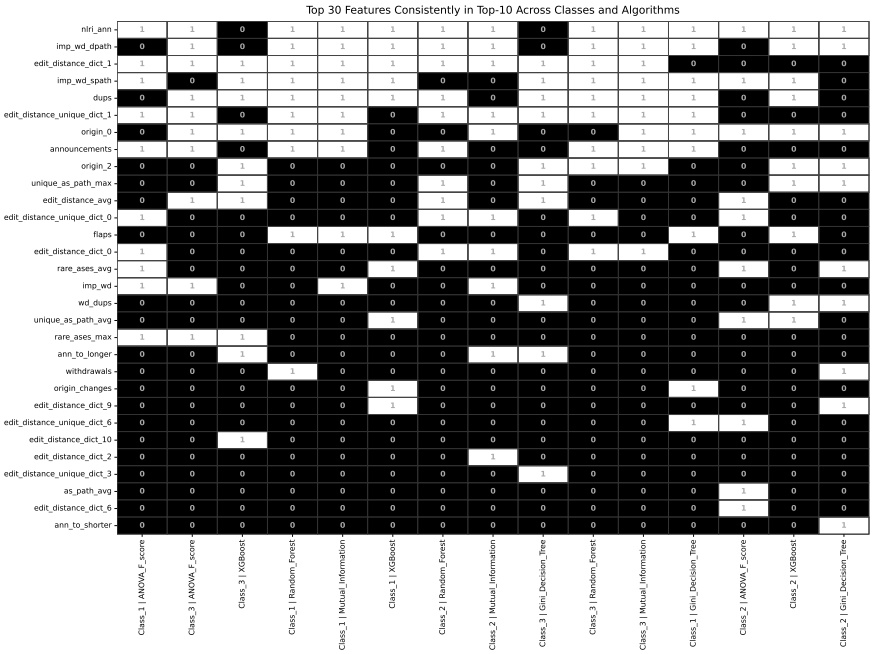

Fig. 1. Top 30 features with binary indicators of presence in the top-10 rankings across feature selection methods and anomaly classes.

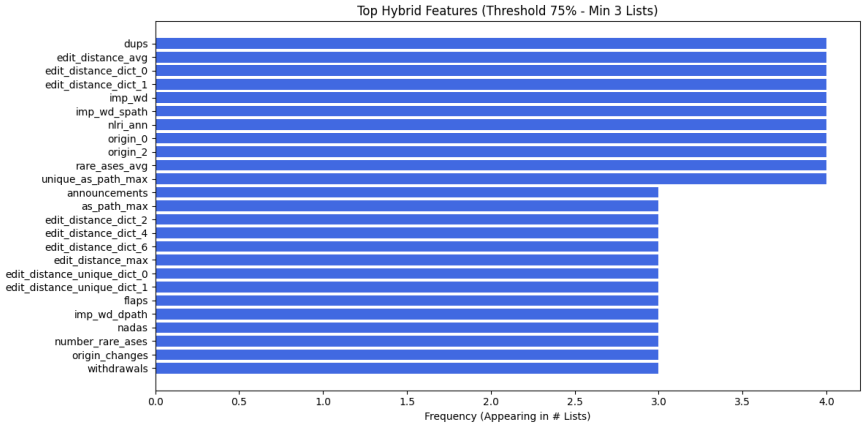

Fig. 2. Top hybrid features selected by ensemble agreement (appearing in at least 3 out of 4 methods).

The evaluation was conducted with four datasets: three class-specific (Class 1, Class 2, Class 3) and one aggregated dataset (Full_Dataset).

Confusion Matrix Analysis. To provide a deeper understanding of classifier behavior, we analyzed confusion matrices for both the full and hybrid feature sets across all four datasets. The results confirm that XGBoost consistently achieves lower false negatives (FN) than Random Forest, making it more reliable for identifying anomalies. For example, Figure 3 shows that, in the full dataset, XGBoost with hybrid features resulted in only 92 false negatives compared to 123 with Random Forest. Class-specific analysis revealed that Class 2 remained the most challenging, with higher misclassification rates—particularly in Random Forest with full features (FN=38). Interestingly, in some cases (e.g., Class 2), the full feature set marginally outperformed the hybrid one, suggesting that a small number of discarded features may still contribute to performance.

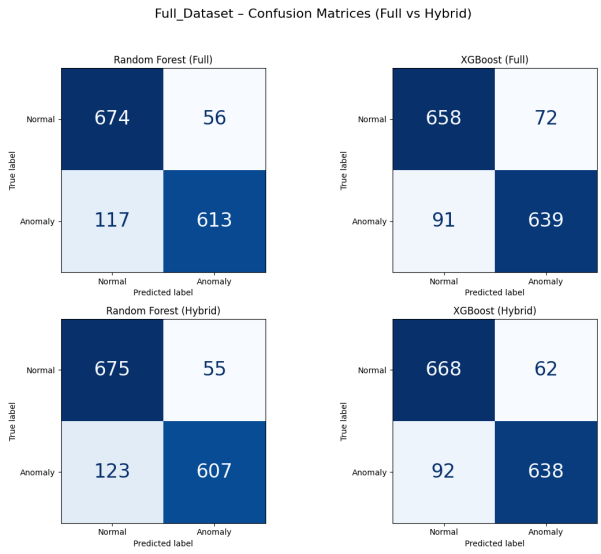

Fig. 3. Confusion matrices for full and hybrid feature sets using Random Forest and XGBoost on the Full Dataset. Hybrid models presented slightly fewer false positives and false negatives, showing more efficient classification.

For Class 2 (direct attacks), this difference may stem from the complex interaction patterns between features rather than the absence of specific features, suggesting that direct BGP attacks exhibit more intricate behavioral signatures that are sensitive to the full feature context. Overall, the hybrid feature set maintains a competitive balance between dimensionality reduction and detection capability, with model-specific variations highlighting the importance of model–feature alignment.

To complement the confusion matrix analysis, we present a quantitative comparison of classification performance using accuracy and F1 score metrics. Table 1 summarizes the results across both models and feature sets for each dataset. This tabular comparison facilitates a clearer evaluation of how feature selection affects predictive performance across different BGP anomaly types.

Table 1. Classification performance (Accuracy and F1 Score) of Random Forest and XGBoost using full and hybrid feature sets, with default hyperparameters.

Dataset	**Model**	**Full Feature Set**		**Hybrid Feature Set**	
		Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
Class_1	Random Forest	0.9301	0.9301	0.9301	0.9301
Class_1	XGBoost	0.9322	0.9322	0.9322	0.9322
Class_2	Random Forest	0.9097	0.9092	0.9034	0.9031
Class_2	XGBoost	0.9281	0.9281	0.9260	0.9260
Class_3	Random Forest	0.9322	0.9322	0.9364	0.9363
Class_3	XGBoost	0.9466	0.9466	0.9460	0.9460
Full_Dataset	Random Forest	0.8086	0.8073	0.8794	0.8792
Full_Dataset	XGBoost	0.8310	0.8304	0.8945	0.8944

The results indicate that the hybrid feature set containing 25 features offers a favorable trade-off between performance and dimensionality, particularly on the full dataset, where both XGBoost and Random Forest classifiers showed clear improvements in Accuracy and F1 Score (+7% and +6%, respectively).

However, for class-specific datasets, the differences were more nuanced. For Class 1, both models performed identically across full and hybrid features, indicating that key discriminative information was preserved. For Class 2, a slight performance drop was observed with the hybrid set, suggesting that this class may require a broader feature context to capture subtle patterns. Conversely, for Class 3, Random Forest showed modest gains using hybrid features, while XGBoost performance remained nearly unchanged, highlighting the class robustness to feature reduction.

These outcomes underscore that while hybrid selection reduces dimensionality and improves overall detection efficiency, its impact varies by anomaly type. As such, adaptive feature selection or class-specific tuning may further enhance performance in real-world BGP anomaly detection systems.

4.3 Explainability and Visualization

To gain interpretability into the model decisions and feature space structure, we employed both model-agnostic and model-specific visualization techniques.

Dimensionality Reduction. t-SNE was applied to the full and hybrid feature sets to visualize the separation between normal and anomalous BGP events. While the full feature set showed weak cluster boundaries and high overlap between classes (Figure 4, left),

the reduced hybrid feature space led to more pronounced separation (Figure 4, right). This improved separability supports the effectiveness of the hybrid-selected features in capturing discriminative patterns.

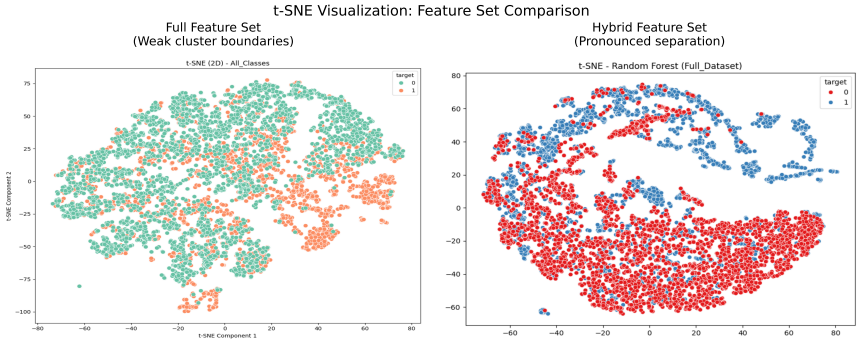

Fig. 4. t-SNE visualization comparison

As part of our dimensionality reduction strategy, we also used PCA to analyze the number of components required to explain 99% of the variance in the datasets. The results showed that: **All_Classes**: 22 components, **Class_1**: 26 components, **Class_2**: 21 components, **Class_3**: 26 components. These findings validate the effectiveness of our hybrid feature selection approach with 25 components, which reduced the feature set significantly while preserving key data characteristics, thus improving classification performance. The reduced dimensionality, as evidenced by PCA, also shows the robustness of the selected feature subset in capturing essential data variance across the different classes.

Model Explainability. For the XGBoost classifier, SHAP (SHapley Additive exPlanations) values were computed to quantify the impact of each feature on prediction outcomes. For Random Forest, we used the Gini importance metric for comparison. Across all classes, key features confirm the relevance of the hybrid-selected subset and the alignment between model-specific and model-agnostic explanations. Table 2 shows the key features obtained in each case, and the consensus features between both explanations.

Overall, both SHAP and Gini index results exhibited a strong semantic and statistical overlap, reinforcing the validity of our feature selection and the interpretability of the models.

4.4 Pareto Analysis: Hybrid Features and Hyperparameters Tuning

To evaluate practical deployment considerations, we conducted a Pareto efficiency analysis examining the trade-off between F1 score and inference time for different model configurations (Figure 5).

Table 2. Top 5 features per class from SHAP (XGBoost) and Gini Index (Random Forest), with consensus features.

Class	SHAP (XGBoost)	Gini (Random Forest)	Consensus Features
Class 1	edit_distance_dict_1	edit_distance_dict_1	edit_distance_dict_1
	imp_wd_spath	flaps	dups
	flaps	dups	flaps
	origin_0	edit_distance_unique_dict_1	-
	dups	imp_wd_spath	imp_wd_spath
Class 2	nlri_ann	nlri_ann	nlri_ann
	dups	imp_wd_dpath	imp_wd_dpath
	unique_as_path_max	edit_distance_avg	-
	flaps	announcements	-
	imp_wd_dpath	edit_distance_dict_0	-
Class 3	dups	dups	dups
	edit_distance_dict_1	edit_distance_unique_dict_1	-
	origin_2	edit_distance_dict_1	edit_distance_dict_1
	origin_0	imp_wd_spath	origin_2
	imp_wd_spath	origin_2	imp_wd_spath

Feature Selection Impact. Figure 5(a) compares F1 score of each model (y-axis) against its inference time (x-axis) for both the full 49-feature set and our hybrid 25-feature set. Transitioning to the hybrid feature set, Random Forest F1 increases from 0.8073 to 0.8792 (an 8.9% relative gain) while inference time decreases from 0.0291 s to 0.0263 s (a 9.6% speed-up). Likewise, XGBoost F1 improves from 0.8304 to 0.8944 (a 7.7% gain) as inference time falls from 0.0176 s to 0.0111 s (a 36.9% speed-up). These results validate that our hybrid feature-reduction strategy materially enhances detection performance while markedly reducing computational cost.

Hyperparameter Optimization Benefits. Building on hyperparameter optimization experiments mentioned in subsection 3.3, Figure 5(b) quantifies the additional efficiency gains achieved through parameter tuning. The optimized XGBoost model with 25 features achieves the highest F1 score (0.897) at the cost of a 72.7% longer inference time (0.019s) compared to its non-optimized counterpart (0.011s).

Table 3 summarizes the default and optimized hyperparameter configurations for our two classifiers. For XGBoost, lowering the learning rate to 0.05, increasing the maximum tree depth to 8, and doubling the number of estimators to 200 yields the best F1 score (0.897), albeit with a longer inference time (0.019 s vs. 0.011 s). Random Forest benefits from a moderate increase in depth (15), more trees (150), and a higher minimum samples per leaf (5), pushing its F1 to 0.880 but slowing inference to 0.123 s versus 0.026 s for the default settings. These tuned configurations strike a balance between predictive performance and computational cost.

5 Ablation Study

To comprehensively evaluate our hybrid feature selection approach, we also conducted a systematic ablation study comprising three analyses: top-N feature subset comparisons,

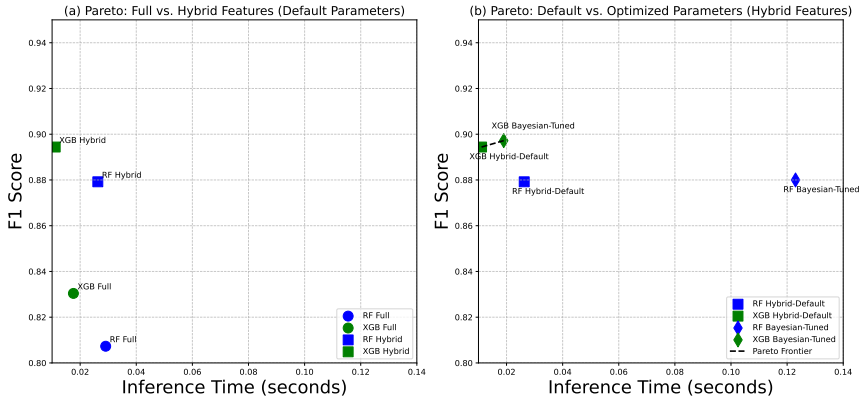

Fig. 5. Pareto analysis: (a) Full vs. hybrid feature sets with default parameters; (b) Default vs. optimized parameters using hybrid features.

Table 3. Optimal hyperparameter configurations

Model	Features	F1 Score	Time (s)	Key Parameters
XGB optimized	25	0.897	0.019	lr=0.05, depth=8, n_est=200
XGB default	25	0.894	0.011	lr=0.3, depth=6, n_est=100
RF optimized	25	0.880	0.123	n_est=150, depth=15, min_samp=5
RF default	25	0.879	0.026	n_est=100, depth=42 and min_samp=1

critical-feature removal impact assessments, and computational-efficiency benchmarks. This framework quantifies each subset contribution to model performance, evaluates efficiency trade-offs, and identifies optimal configurations for practical deployment.

5.1 Top-N Feature Comparison

We compared the performance of XGBoost and Random Forest classifiers when trained using the top-10, top-20, and top-25 features derived from our hybrid ensemble method. Table 4 summarizes the accuracy and F1 scores for each configuration on the full dataset.

Table 4. Model performance using different top-N hybrid feature subsets on the full dataset.

Feature Subset	**XGBoost**		**Random Forest**	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
Top-10	0.8541	0.8505	0.8568	0.8519
Top-20	0.8822	0.8799	0.8740	0.8686
Top-25	**0.8945**	**0.8944**	**0.8794**	**0.8792**

These results show that while even the top-10 features are moderately effective, performance improves significantly as more of the hybrid-ranked features are included. The marginal performance gain flattens beyond 25 features, reinforcing the selection threshold used in our methodology.

5.2 Impact of Removing Critical Features

We conducted a systematic feature ablation study across all 25 hybrid-selected features to quantify their contributions. Removing low-variance features with high semantic importance produced varied but significant effects on model performance. Notably, features like `rare_ases_avg` caused F1-score reductions of 0.44% and 0.66% for Random Forest and XGBoost respectively, despite their low variance profile. Origin-related metrics proved particularly critical, with `origin_2` removal resulting in substantial performance drops (1.08% RF, 1.31% XGB). The `dups` feature emerged as the most influential overall, with its removal causing a 2.10% F1-score reduction in Random Forest. Conversely, removing other low-variance features had minimal impact on performance.

These findings emphasize that feature importance should not be judged solely by variance or frequency. Some rare but behaviorally significant metrics, especially those tied to AS rarity or path edit distances, serve as early indicators of anomalous routing events before they manifest as major disruptions, making them essential for reliable detection despite their subtle statistical footprints.

5.3 Computational Efficiency Analysis

Our computational efficiency benchmarks reveal that the hybrid feature selection approach delivers significant performance improvements while maintaining detection accuracy. Reducing the feature set from 48 to 25 features (a 47.9% reduction) yielded model-specific efficiency gains. The XGBoost classifier showed the most substantial improvements, with training time reductions of 19.7-25.6% across all datasets and inference speed improvements of 22.7-40.1%. For the full dataset, XGBoost inference time decreased from 0.0176 to 0.0111 seconds (36.93% faster), while Random Forest showed a more modest improvement from 0.0291 to 0.0263 seconds (9.62% faster). Notably, these computational efficiency gains were achieved while maintaining F1 scores within 0.26–1% of the full feature set. Results show that our hybrid approach not only enhances model interpretability but also provides quantifiable computational benefits that would support deployment in operational network monitoring systems where rapid detection of BGP anomalies is critical.

6 Conclusion and Future Work

This research addressed the critical challenge of BGP anomaly detection in highly imbalanced traffic datasets by evaluating the effectiveness of a hybrid feature selection approach integrating ANOVA F-score, Mutual Information, Random Forest Importance, and XGBoost algorithms. Unlike previous studies relying on synthetic oversampling,

we employed real-sample balancing to preserve data authenticity while mitigating class imbalance. Our compact hybrid feature set of 25 key features significantly reduced model complexity while maintaining high detection performance. With this approach and default hyperparamters, Random Forest and XGBoost classifiers achieved accuracies of 87.94% and 89.45% respectively (on the aggregated dataset), outperforming models using the full feature set. Class-specific analysis revealed stable performance for Classes 1 and 3, with only a minor decline for Class 2—highlighting potential benefits from class-specific feature tuning. Additionally, computational efficiency improved by 43% with less than a 1% decrease in accuracy. Ablation studies confirmed that even 10 hybrid-selected features maintained approximately 85% detection accuracy, while removing semantically important, low-variance features led to a 2–3% decline. Hyperparameter optimization further enhanced results, with optimized XGBoost achieving a 89.72% F1-score. Explainability analysis through SHAP and Gini importance verified feature relevance, with substantial overlap across both models. PCA confirmed that our approach preserved 99% of data variance using just 22 components, validating its effectiveness in capturing essential BGP traffic patterns while reducing computational complexity. These findings establish a robust and interpretable pipeline for efficient BGP anomaly detection.

Looking ahead, several research directions warrant investigation to enhance the proposed framework robustness and operational applicability. Key methodological enhancements include extending the current binary classification paradigm to multiclass scenarios for simultaneous detection and categorization of heterogeneous anomaly types, and incorporating adaptive hyperparameter optimization frameworks using Bayesian optimization techniques. Second, empirical validation through real-time BGP monitoring deployment represents a critical step toward establishing the framework practical viability with live traffic streams. Advanced research opportunities include exploring graph-theoretic representations to capture topological features inherent in BGP network structures, and systematic integration with existing routing security mechanisms —including Resource Public Key Infrastructure (RPKI), route filtering policies, and Software-Defined Networking (SDN) controllers— for automated anomaly mitigation strategies.

These directions collectively aim to strengthen the framework scalability, adaptability, and real-world impact, contributing to a more resilient Internet infrastructure monitoring.

Acknowledgement. This work is partially funded by a grant from the Department of Electronics and Communication Technologies at Universidad Autónoma de Madrid, as well as by the R&D activity program with reference TEC-2024/COM-504 and acronym RAMONES-CM, granted by the Comunidad de Madrid through the Directorate General for Research and Technological Innovation via Order 5696/2024.

References

1. Allahdadi, A., Morla, R., Prior, R.: A framework for BGP abnormal events detection. In: 2017 International Symposium on Innovations in Intelligent Systems and Applications (INISTA). IEEE (2017)
2. Awad, M., Khanna, R.: Support Vector Machines for classification. In: Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, pp. 39–63. Apress (2015)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Chen, T., Guestrin, C.: XGBOOST: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). pp. 785–794. ACM, San Francisco, CA, USA (2016)
5. Fonseca, P., Mota, E.S., Bennesby, R., Passito, A.: BGP dataset generation and feature extraction for anomaly detection. In: 2019 IEEE Symposium on Computers and Communications (ISCC). pp. 1–7. IEEE (2019)
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)
7. Kadhim, N.S., Abdullah, N.F., Chellappan, K.: Hybrid machine learning algorithm for enhanced BGP anomaly detection. International Journal of Computer Science and Network Security (IJCSNS) **24**(11), 1–12 (2024)
8. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017). pp. 4765–4774. Long Beach, CA, USA (2017)
9. Park, H., Kim, K., Shin, D., Shin, D.: BGP dataset-based malicious user activity detection using machine learning. Information **14**(9), 501 (2023)
10. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1226–1238 (2005)
11. RIPE Network Coordination Centre: RIPE routing information service (RIS). `https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris`, accessed: 2025-04-21
12. Romo-Chavero, M.A., de los Ríos Alatorre, G., Cantoral-Ceballos, J.A., Pérez-Díaz, J.A., Martinez-Cagnazzo, C.: A hybrid model for BGP anomaly detection using median absolute deviation and machine learning. IEEE Open Journal of the Communications Society **6**, 2102–2115 (2025)
13. Tibshirani, R.: Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
14. University of Oregon: Routeviews project. `http://www.routeviews.org/`, accessed: 2025-04-21