Jurnal Kejuruteraan 38(3) 2026: 1449-1461
https://doi.org/10.17576/jkukm-2026-38(3)-35

A Comparative Study of Machine Learning Deep Learning and Hybrid Approaches to Enhance BGP Traffic Security

Nassir S. Kadhim[a,b], Nor Fadzilah Abdullah[a] & Kalaivani Chellappan[a*]

[a]*Faculty of Engineering and Built Environment; The National University of Malaysia (UKM), Malaysia*

[b]*Ministry of Communication (MOC) / ITPC Company -Iraq.*

[*]*Corresponding author: kckalai@ukm.edu.my*

Received 6 May 2025, Received in revised form 7 January 2026
Accepted 7 February 2026, Available online 30 May 2026

ABSTRACT

The Border Gateway Protocol (BGP) is important for internet routing, enabling the exchange of routing information between autonomous systems. However, it remains vulnerable to cyberattacks such as hijacking, Denial-of-Service (DoS) attacks, and network outages. Although the recent advancements in machine learning (ML) hold promise for accurate BGP anomaly detection, the existing publicly available datasets often contain outdated information regarding past BGP cyberattacks, hindering models of novel threat detection. Furthermore, the network topology criteria are also often neglected for most anomaly detection models. In a multi-stage approach, this work employs a real-topology simulation to analyze BGP traffic under attack scenarios, deriving 24 features to create datasets for a machine learning-based anomaly detection system. A comparative evaluation of eight machine learning (ML) algorithms determined that Random Forest (RF) was the most effective, achieving an accuracy of 94.6%. Among four deep learning (DL) models, Bidirectional Long Short-Term Memory (BiLSTM) demonstrated the highest accuracy at 98.9%. To further improve detection performance, hybrid ML models integrating RF-SGD, KNN-LR, and RF-QDA were developed, with the RF-SGD model achieving the highest accuracy of 99.3% and an Area Under Curve (AUC) of 0.988. The results indicate that hybrid models outperform standalone ML and DL approaches, providing a more robust and efficient solution for enhancing BGP security against advancing BGP cyberattacks.

Keywords: BGP security; machine learning; deep learning; anomaly detection; cyberattacks

INTRODUCTION

The Border Gateway Protocol (BGP), the backbone of inter-domain routes on the Internet, has long been recognized for its susceptibility to security threats. Despite numerous proposed countermeasures, BGP vulnerabilities persist due to its foundational design, which lacks intrinsic mechanisms for authentication or route validation. The common key vulnerabilities include prefix hijacking, where malicious actors advertise unauthorized IP prefixes, and route leaks, which disrupt intended routing paths. These exploits can result in traffic misdirection, eavesdropping, or complete service disruption for targeted entities (Scott, Johnstone & Szewczyk 2024). Recent studies highlight the evolution of attack strategies against BGP. For instance, the manipulation of the Autonomous System (AS) paths allows attackers to contaminate routing tables and propagate invalid updates across networks. BGP hijacking incidents, such as those affecting major technological companies and government institutes, underline the protocol's susceptibility to direct and indirect attacks. Furthermore, advanced techniques like multi-vector attacks exploit single-point detection weaknesses in existing security frameworks, rendering many anomaly detection methods ineffective against sophisticated adversaries (Al-Musawi et al. 2017). Efforts to mitigate these vulnerabilities, including cryptographic techniques like BGPsec and real-time monitoring systems, have seen partial success but often face challenges such as high computational costs and limited scalability. The ongoing evolution of attack methodologies necessitates the adoption of more robust approaches, such as machine learning and deep learning,

to enhance anomaly detection and improve protection to inter-domain routing systems (P. Edwards 2019).

Traditional BGP security methods face several limitations that hinder their effectiveness in securing the Internet's routing infrastructure. Solutions such as Secure BGP (S-BGP) involve complex cryptographic operations that are costly and challenging to implement. These solutions often require significant computational resources, which can increase convergence times and are not easily deployable. While cryptographic techniques like BGPSec aim to ensure authentication and integrity, they do not fully prevent attacks that exploit the BGP protocol's complexity. These measures are insufficient to address all types of attacks, such as those manipulating routing policies (Li et al. 2015).

Despite numerous proposals for enhancing BGP security, BGP was not designed with security in mind, leading to vulnerabilities such as prefix hijacking, route leaks, and denial-of-service (DoS) attacks (Butler et al. 2010; Mitseva et al. 2018). The protocol's reliance on trust among networks makes it susceptible to various attacks (Bezawada et al. 2008). With the rapid growth of Internet traffic, BGP has become a critical component of network infrastructure management. Therefore, early detection and mitigation of network anomalies are essential for maintaining Internet stability and security. Various methodologies have been proposed for detecting network anomalies, including statistical methods, data mining techniques, hegemony analysis, and machine learning approaches (Al-Daweri et al. 2021). Over the past decade, several ML- and DL-based techniques have been implemented to improve the ability of the attack detection system to identify malicious activities (Liao et al. 2024), (Shtayat et al. 2023). The significant growth in internet network traffic and ensuing security risks in parallel create difficulties for these systems to identify cyberattacks effectively. The key idea is to provide up-to-date information on recent ML- and DL-based attack detection systems to provide a baseline for new researchers exploring this important domain (Azam et al. 2023). Various methodologies are employed in signature-based and anomaly-based approaches.

In this study, machine learning (ML), deep learning (DL), and hybrid model approaches were employed to enhance BGP security by leveraging a simulated data-based Iraq BGP topology architecture (Internet Gateways Network). By employing a realistic and context-specific network simulation, this research aims to address the unique challenges faced by regions with limited resources and high deployment costs. Furthermore, it presents a comparative analysis of ML and DL models for anomaly detection, assessing their scalability and feasibility in addressing security threats, including prefix hijacking, DoS attacks, and network outages. This evaluation is based on key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

RELATED WORK

Various approaches have been proposed to address BGP security vulnerabilities and security challenges. (Yang & Jia 2023) introduced a path-based algorithm combining attributes and topological information from BGP routing tables. (Mujtaba et al. 2012) employed the Failure Quality Control method for real-time intrusion detection. Machine learning techniques have also been applied, with (Arai et al. 2019) identifying the top 10 features sufficient for accurate classification of BGP anomalies. A novel approach using matrix profile data mining was presented by (Scott et al. 2024), offering advantages such as domain agnosticism, assumption-free analysis, and scalability. These methods aim to protect against BGP hijacking attacks, which can intercept sensitive data and disrupt services, as well as other threats like DoS attacks and routing misconfigurations.

Furthermore, Edwards (2019) introduced a new method for anomaly detection in the BGP using unsupervised learning techniques. Subsequently, the occurrences of anomalies were collected and triggered; the possible data pipeline methods were explored. For historical events, the authors detected the recognized anomalies using k-means and DBSCAN. Dai et al. (2019) proposed a support vector machine-based BGP anomaly detection method (SVM-BGPAD). They employed a feature selection algorithm based on Fisher linear analysis and Markov random field technology. The SVM algorithm parameters were optimized using the cross-validation and grid search methods. The results of the introduced model were compared based on accuracy and F1-score. Tripathy and Behera (2023) assessed and analyzed the performance of several machine learning algorithms for detecting BGP anomalies using RIB. Naive Bayes (NB), Decision Tree (J48), and Support Vector Machine (SVM) classifiers were employed to detect BGP network traffic anomalies. The researchers evaluated feature discretization and feature selection using Slammer, Nimda, and Code Red I data sets of known Internet anomalies.

Deep learning methods, in addition to machine learning, define complex models that can capture BGP behaviors that statistical pattern recognition models, with their shallow nature, cannot handle. Cheng et al. (2018) proposed a novel multi-scale LSTM model (MSLSTM) to model and capture the anomalous behaviors from BGP traffic, which uses discrete wavelet transform to generate

wide-ranging information of different scales explicitly from time sequences. Ha et al. (2022) pre-processed the three datasets to create training and experimental data. The data were input into a CNN–LSTM hybrid model, with the LSTM handling temporal information and the CNN handling spatial information. (Sunita & Mallapur 2022) introduced a novel BGP anomaly detection model via hybrid classifiers that combine LSTM and optimized CNN over two processes: (1) feature extraction and (2) anomaly detection-classification. During the feature extraction step, certain features, like the number of exterior gateway protocol (EGP) packets, were extracted.

This study provides a comparison of machine learning and deep learning models for the binary classification-based BGP anomaly detection system. Our work focuses on evaluating the performance metrics of eight ML, four DL, and three hybrid ML models, as presented in the performance metrics evaluation section.

METHODOLOGY

This work employs a structured approach to evaluate and compare the effectiveness of machine learning (ML), deep learning (DL), and advanced hybrid models for BGP traffic security using simulated data reflective of actual BGP network topology. The methodology consists of multiple sequential stages designed to simulate realistic BGP attack scenarios, deriving relevant features, preparing datasets that utilize inputs in the ML, DL, and hybrid models for assessment performance metrics of BGP anomaly detection, and finally, comparative evaluation of the performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC). Figure 1 outlines the methodology process in a flowchart.

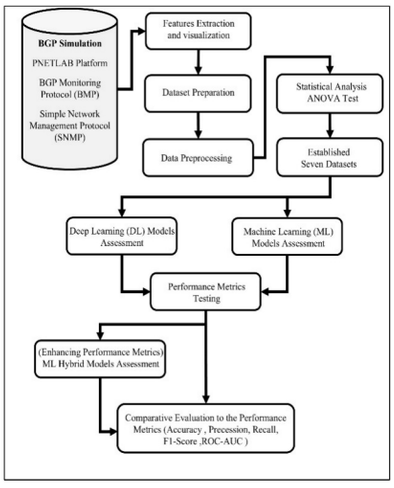

FIGURE 1. Methodology Flow chart

The following sections provide a detailed discussion of the methodological phases utilized in this work.

SIMULATION ENVIRONMENT

The simulation was conducted using the PNETLAB (Packet Network Emulator Tool Lab) platform, designed to simulate, test, and improve network robustness. It has been widely employed for evaluating network protocols, such as BGP, under various conditions to understand their vulnerabilities, performance, and reliability (Vinod & Mahesh 2023). PNETLAB was chosen for its scalability and support for advanced BGP configurations. The simulated environment modeled a distributed routing network with multiple Autonomous Systems (ASes) for the several internet service providers (ISPs) interconnected via routers across the BGP network.

The network design simulation served as the foundational platform for analyzing BGP security mechanisms. This simulation environment was constructed to emulate real-world networking scenarios, enabling the controlled study of security threats, anomaly detection, and mitigation strategies in a BGP context. A systematic approach was employed to ensure the simulation framework closely aligned with typical network configurations and traffic patterns encountered in operational networks. Iraq's BGP network topology was emulated by creating the High Level Design (HLD) of the MOC IGW peers' topology, as shown in Figure 2, and setting up a lab environment in PNETLAB to simulate a cyberattack.

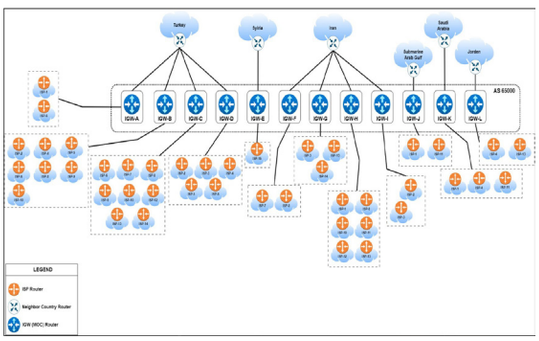

FIGURE 2. Iraqi gateway network high level design (HLD)

DATA COLLECTION

Using OpenBMP as the data collector, BGP update information was imported into a PostgreSQL database, and SQL queries were applied to derive the BGP features most impacted by cyberattacks (Cosovic & Slobodan 2018). Several BGP anomaly detection models commonly utilize

these features. In this study, three cyberattack scenarios—sub-prefix hijacking, denial-of-service (DoS), and link outage were simulated in addition to the normal operation scenario. BGP updates were recorded at one-minute intervals for each scenario throughout the simulator's designated operation period to generate the corresponding BGP feature datasets.(Kadhim et al. 2025).

FEATURE DERIVATION AND VISUALIZATION

A total of twenty-four BGP features were derived from the AS-PATH attributes and the volume of BGP update messages for both normal and cyberattack scenarios. These features play a crucial role in distinguishing abnormal routing behaviors and detecting deviations indicative of network instability.

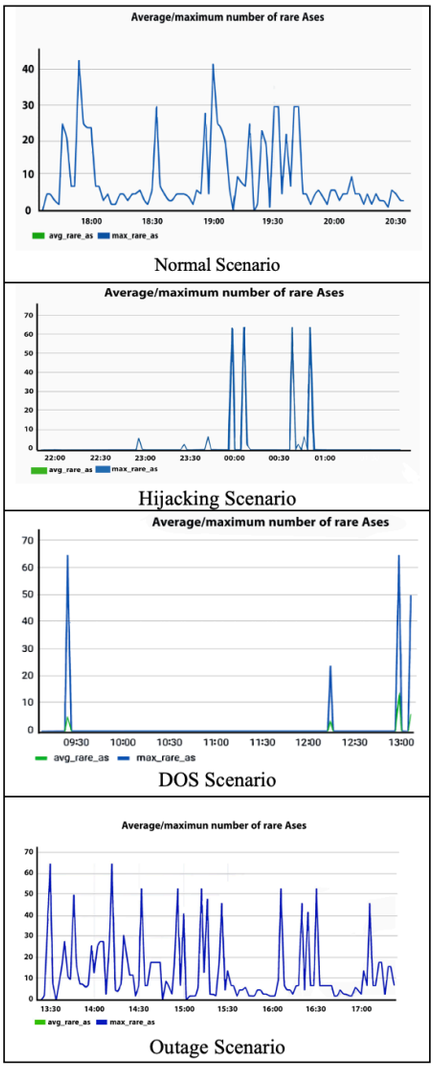

FIGURE 3. Average and maximum number of rare (Ases) Feature Visualization

Additionally, the PNETLab dashboard was utilized to visualize the derived features, generating graphical representations of BGP dynamics to enhance interpretability and support detailed analytical assessment. The visualization of these features enables an in-depth exploration of their statistical properties, distribution patterns, and potential correlations, offering valuable insights into their role in network stability and routing efficiency. (Idowu et al. 2013). For example, Figure 3 shows the average and maximum number of rare Autonomous Systems (Ases), where a BGP rare AS feature identifies (Ases) that appear infrequently in routing paths or updates under normal and attack conditions. This visualization illustrates feature behavior across scenarios, and the same methodology was applied to all BGP features to assess their relevance in distinguishing normal and attack scenarios.

FEATURES STATISTICAL ANALYSIS

The analysis of variance (ANOVA) test was utilized to generate the F-statistic and corresponding p-values (Ostertagova & Ostertag 2013) for all the features that were derived, as shown in Table 1.

The F-statistics allowed us to measure the variance between different groups compared to the variance within the groups, while the p-values provided statistical significance for each feature. Features with a p-value < 0.05 were considered statistically significant, facilitating the identification of those most affected by cyberattacks. Based on the findings of the ANOVA statistical analysis, it was observed that six BGP features are the most significant in relation to BGP cyberattacks within the range of F-statistic values ($F > 90$), which we termed "Features Set A". The second group of nine features showed moderate significance within the range of F-statistic values ($30 < F < 90$); this group is referred to as "Features Set B". The third group of seven, referred to as "Features Set C", showed a lower significance than the rest of the features ($F < 30$). Finally, two features were identified as statistically insignificant, with p-values exceeding 0.05, indicating that they were unaffected by cyberattacks. These findings highlight the varying levels of sensitivity among various BGP features, providing a structured framework for significant feature selection in the context of BGP anomaly detection.

TABLE 1. Features set creation based on ANOVA test

Feature Name	P-Value	F - Statistic	Set Name
Average/ Maximum edit distance	0.00000	661.37812	Features Set (A)
Prefix origin change	0.00000	294.58378	
Number of ORIGIN changes	0.00000	162.35400	
Number of implicit withdrawals with same/different path	0.00000	132.80608	
Maximum unique AS-PATH length	0.00000	131.12853	
Number of rare Ases	0.00000	92.37245	
Number of plain new announcements	0.00000	59.31361	
Number of new paths announced after withdrawing an old path	0.00000	59.10333	
Maximum/average announcements per AS	0.00000	56.58267	
Announcements to Longer paths	0.00000	52.32693	
Number of announcements/withdrawals	0.00000	50.00072	Features Set (B)
Maximum/average announcements per prefix	0.00000	42.16644	
Number of IGP/EGP/ INCOMPLETE messages	0.00000	37.36794	
Number of announced prefixes	0.00000	33.89654	
Average number of rare Ases	0.00000	31.74219	
maximum number of rare Ases	0.00000	27.96660	
Announcements to Shorter paths	0.00000	19.96247	
Average AS-PATH Length	0.00000	18.11514	
Average unique AS-PATH length	0.00000	15.67355	
Number of new-path announcements	0.00001	8.59153	Features Set (C)
AS-Path changes according to geographic location	0.00006	7.41663	
Maximum AS-PATH Length	0.00022	6.50876	
Number of new announcements after withdrawal	0.18574	1.60700	Not Significant
Number of duplicate announcements/withdrawals	0.28569	1.26237	

FEATURE SETS CATEGORIZATION

The statistical analysis results were adopted to classify the 22 most impacted features into three groups (A, B, and C) based on their significance. Additionally, three supplementary feature sets were generated by combining each pair of primary feature groups, along with a fourth comprehensive feature set incorporating all three groups, as presented in Table 2. The primary aim of employing multiple feature sets was to expand the input space of machine learning models, enabling them to learn more effectively from the data and produce accurate predictions. All constructed datasets were evaluated using machine learning and deep learning methods based on binary classification.

TABLE 2. BGP significant combination feature sets

Combining sets	Number of significant features
Set (A) + (B)	15
Set (A) + (C)	13
Set (B) + (C)	16
Set (A) + (B) + (C)	22

PERFORMANCE METRICS EVALUATION

This research focused on evaluating the performance metrics of machine learning (ML) and deep learning (DL) models for anomaly detection using binary classification. The primary goal was to assess the classifier's efficacy in accurately detecting anomalies while minimizing false classifications. Key performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were calculated to measure the models' reliability. Table 3 presents the confusion matrix, illustrating the classifier's decision-making process in distinguishing anomalies from normal instances (Kevin Hoarau 2021).

TABLE 3. ML Confusion Matrix

		Predicted Class	
		Anomaly	Regular
Actual Class	Anomaly (True)	TP	FN
	Regular (False)	FP	TN

The factors of the classification are termed as TP (true positive), which is the number of correctly classified attacks; TN (true negative), the number of normal flows correctly classified; FP (false positive), the number of normal instances misclassified as attacks; and FN (false negative), the number of attack instances misclassified as normal (Park et al. 2023). Based on these elements, the performance metrics can be defined as the following (Butler et al. 2010).

1. Accuracy: Measures the classification model's ability to correctly categorize data examples, regardless of whether they are positive or negative. It is determined as the ratio of the total number of correctly categorized records in a dataset to the total number of rows in the dataset:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. Precision: A metric that quantifies the correctness of positive predictions provided by a model. The calculation involves determining the proportion of accurate positive predictions produced by the model relative to the total number of true positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3. Recall: A metric that quantifies the proportion of positive instances correctly detected by the model. It is computed by dividing the number of true positive predictions by the total number of real positive occurrences in the dataset:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

4. F1-score: Commonly referred to as the F-score or F-measure, it is an ML evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The formulation of the harmonic mean of precision and recall is given by:

$$\text{F1-score} = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{4}$$

5. ROC-AUC Score: The ROC-AUC value indicates the classifier's efficacy in distinguishing between positive and negative classifications. It may assume values ranging from 0 to 1. An elevated ROC AUC signifies a higher performance.

$$AUC - ROC = \int_0^1 TPR(FPR)d(FPR)$$

Where:

$$TPR = \frac{TP}{TP+FN}, \qquad FPR = \frac{FP}{FP+TN}$$

ML MODELS EVALUATION

Eight ML models were evaluated (SVM, LR, NB, DT, QDA, KNN, RF, and SGD), producing the metric scores shown in Table 4 for accuracy, Table 5 for precision, Table 6 for recall, Table 7 for F1 scores, and Table 8 for ROC-AUC scores, highlighting the best results for each metric. The comparative analysis of these models for BGP anomaly detection reveals varying levels of efficacy across different techniques and created datasets. Since the random forest (RF) exhibited the highest accuracy of 94.60% on datasets A+B+C and A+C, along with a high ROC-AUC of 0.944, it demonstrates its exceptional BGP anomaly detection ability. Logistic Regression (LR) and Decision Tree (DT) also performed well, achieving accuracies of 93.80% on A+B+C and A+C, respectively, with ROC-AUC values slightly above 0.94, indicating reliable sensitivity and specificity. K-Nearest Neighbors (KNN) demonstrated strong performance, with an accuracy of 93.50% on dataset A, while Support Vector Machine (SVM) and Naive Bayes (NB) delivered comparatively lower results, reflecting limitations in handling complex or dynamic relationships. Stochastic Gradient Descent (SGD) and Quadratic Discriminant Analysis (QDA) provided moderate accuracy and ROC-AUC values, rendering them less consistent than RF or DT. Overall, RF is the most robust machine learning model for anomaly detection, while dataset selection significantly impacted performance across all techniques.

TABLE 4. Accuracy of the ML models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
SVM	0.5485	0.5926	0.5235	0.5857	0.5955	0.5681	0.5882
LR	0.9078	0.8338	0.6475	0.9333	0.9127	0.8720	0.9387
NB	0.8651	0.8142	0.5480	0.8872	0.7990	0.6960	0.8475
DT	0.8980	0.8303	0.7460	0.8264	0.9382	0.8857	0.9098
QDA	0.8480	0.8269	0.75	0.9240	0.8014	0.8892	0.8627
KNN	0.9357	0.9122	0.8519	0.9264	0.9343	0.9196	0.9235
RF	0.9426	0.8818	0.7563	0.9450	0.9460	0.8426	0.9460
SGD	0.8612	0.7573	0.6004	0.8813	0.9014	0.8269	0.9230

TABLE 5. Precision of the ML models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
SVM	0.5174	0.5501	0.4959	0.5503	0.5739	0.5425	0.5411
LR	0.9088	0.8065	0.6235	0.9153	0.9141	0.8450	0.9282
NB	0.8001	0.7902	0.5106	0.8324	0.7125	0.6154	0.7757
DT	0.9351	0.8119	0.6866	0.7589	0.8994	0.9173	0.9190
QDA	0.7690	0.7465	0.6628	0.8702	0.7723	0.8324	0.9404
KNN	0.9404	0.8839	0.8057	0.9099	0.9469	0.8932	0.8988
RF	0.9678	0.9089	0.6990	0.9796	0.9743	0.8331	0.9818
SGD	0.8308	0.7657	0.5626	0.9061	0.9058	0.7978	0.9059

TABLE 6. Recall of the ML models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
SVM	0.6010	0.7375	0.7729	0.6541	0.5458	0.525	0.8218
LR	0.8937	0.8510	0.6333	0.9458	0.8989	0.8916	0.9427
NB	0.9510	0.8239	0.9531	0.9520	0.9604	0.9437	0.9510
DT	0.8416	0.8322	0.8468	0.925	0.9781	0.8322	0.8864
QDA	0.9677	0.9572	0.9541	0.9854	0.8197	0.9572	0.7562

continue...

...cont.

KNN	0.9218	0.9364	0.9031	0.9364	0.9114	0.9416	0.9437
RF	0.9083	0.8322	0.8468	0.9020	0.9093	0.8322	0.9020
SGD	0.8854	0.6979	0.6781	0.8343	0.8822	0.8468	0.9333

TABLE 7. F1-Score of the ML models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
SVM	0.5561	0.6301	0.6042	0.5978	0.5595	0.5336	0.6526
LR	0.9012	0.8281	0.6284	0.9303	0.9065	0.8677	0.9354
NB	0.8691	0.8067	0.6649	0.8882	0.8181	0.7450	0.8544
DT	0.8859	0.8220	0.7583	0.8338	0.9371	0.8727	0.9024
QDA	0.8570	0.8388	0.7822	0.9242	0.7953	0.8905	0.8383
KNN	0.9310	0.9094	0.8516	0.9229	0.9288	0.9168	0.9207
RF	0.9371	0.8689	0.7658	0.9392	0.9407	0.8327	0.9402
SGD	0.8572	0.7302	0.6150	0.8687	0.8939	0.8216	0.9194

TABLE 8. ROC-AUC Score of the ML models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
SVM	0.5514	0.6006	0.5373	0.5895	0.5928	0.5657	0.6012
LR	0.9070	0.8347	0.6467	0.9340	0.9119	0.8731	0.9389
NB	0.8699	0.8147	0.5705	0.8908	0.8079	0.7098	0.8532
DT	0.8949	0.8304	0.7516	0.8319	0.9404	0.8828	0.9085
QDA	0.8546	0.8342	0.7613	0.9274	0.8024	0.8929	0.8568
KNN	0.9350	0.9136	0.8548	0.9270	0.9330	0.9208	0.9246
RF	0.9407	0.8791	0.7614	0.9427	0.9440	0.8420	0.9436
SGD	0.8626	0.7540	0.6048	0.8787	0.9004	0.8280	0.9236

DL MODELS EVALUATION

Deep learning (DL) is a learning model that performs calculations used in machine learning over many layers at once and reveals highly complex patterns inside the data (Lecun et al. 2015). This work evaluation demonstrates that the four assessed deep learning models (ANN, CNN, LSTM, and BiLSTM) for BGP anomaly detection exhibit strong performance metrics, particularly in terms of accuracy and ROC-AUC values, underscoring their ability to capture complex patterns in the data (Alaghbari et al. 2022). Table 9 presents accuracy results for the tested datasets, Table 10 for precision, Table 11 for recall, Table 12 for F1-score, and Table 13 for ROC-AUC score. The CNN achieved a high accuracy of 96.40% on dataset A+B with a strong ROC-AUC value of 0.9613 on A+C, whereas ANN demonstrated competitive performance with an accuracy of 94.95% on A+B and an ROC-AUC of 0.9616 on dataset A. Additionally, the LSTM model performed exceptionally well with an accuracy of 98.60% and a comparable ROC-AUC of 0.9854 on dataset A+B+C, showcasing its capacity to handle relationships with time. The BiLSTM model emerged as the best-performing model, achieving the highest accuracy of 98.70% on dataset A+B and the top ROC-AUC value of 0.9857 on dataset B+C, reflecting its strength in anomaly data detection.

TABLE 9. Accuracy of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
ANN	0.9446	0.8936	0.8074	0.9495	0.9426	0.9059	0.9475
CNN	0.9471	0.9113	0.8461	0.9642	0.9471	0.9299	0.9632
LSTM	0.9569	0.9711	0.8377	0.9828	0.9735	0.9745	0.9868
BiLSTM	0.9657	0.9725	0.8299	0.9873	0.9804	0.9853	0.9868

TABLE 10. Precision of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
ANN	0.9389	0.8237	0.7189	0.9149	0.9323	0.8397	0.9113
CNN	0.9358	0.8715	0.7775	0.9484	0.9369	0.8956	0.9557
LSTM	0.9544	0.957	0.7703	0.9741	0.9687	0.9637	0.9873
BiLSTM	0.9634	0.9565	0.7666	0.9852	0.9677	0.9772	0.9791

TABLE 11. Recall of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
ANN	0.946	0.9879	0.9717	0.9838	0.9477	0.9884	0.9859
CNN	0.9544	0.9521	0.9404	0.9773	0.9523	0.9633	0.9669
LSTM	0.954	0.9834	0.9313	0.9895	0.9744	0.9826	0.9848
BiLSTM	0.9625	0.9871	0.9189	0.9878	0.9906	0.9918	0.9931

TABLE 12. F1-Score of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
ANN	0.9415	0.8978	0.8255	0.9479	0.9394	0.9078	0.9466
CNN	0.9445	0.9093	0.8507	0.9624	0.9442	0.9275	0.9611
LSTM	0.9538	0.9697	0.8427	0.9817	0.9713	0.9728	0.986
BiLSTM	0.9627	0.9714	0.8345	0.9864	0.9789	0.9844	0.986

TABLE 13. ROC-AUC Score of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
ANN	0.9616	0.8807	0.7896	0.9569	0.9159	0.8952	0.9468
CNN	0.9610	0.9033	0.8325	0.9414	0.9613	0.9319	0.9560
LSTM	0.9610	0.9756	0.8159	0.9804	0.9848	0.9613	0.9854
BiLSTM	0.9804	0.9706	0.7988	0.9851	0.9854	0.9857	0.9759

HYBRID ML MODELS EVALUATION

Hybrid models were advanced by combining pairs of machine learning (ML) algorithms to use their complementary strengths and improve performance metrics, particularly accuracy, compared to individual ML (Markkandeyan et al. 2025). A hybrid model incorporates methodologies to cluster the strengths of different methods while mitigating their weaknesses, leading to improved overall performance for the individual methods (Flamia Azevedo et al. 2024), (Elsayed et al. 2021). Achieving high accuracy is vital for BGP anomaly detection systems, as it ensures reliable identification of anomalies while minimizing false positives and negatives, which is essential for maintaining the stability and security of network infrastructure. The evaluated hybrid combinations include:

1. RF + QDA
2. KNN + LR
3. RF + SGD

The mentioned hybrid models demonstrated a significant enhancement in accuracy by combining the strengths of individual machine learning, hence offering

enhanced performance for BGP anomaly detection. The evaluation results for the hybrid models are shown in Table 14 for accuracy, Table 15 for precision, Table 16 for recall, Table 17 for F1-score, and Table 18 for ROC-AUC score across all tested datasets. The RF-QDA hybrid model achieved an impressive accuracy of 98.80% on dataset A+B with a high ROC-AUC of 0.9885, leveraging the complementary strengths of RF's tree-based structure and Quadratic Discriminant Analysis (QDA)'s probabilistic framework. The KNN-LR hybrid model exhibited the capability to enhance performance by combining K-Nearest Neighbors (KNN) with Logistic Regression (LR), although it achieved a slightly lower accuracy of 94.60% on datasets A+B+C and A+B, respectively, with an ROC-AUC of 0.944. The RF-SGD hybrid model achieved the highest accuracy of 99.30% on dataset A+B+C with a remarkable ROC-AUC of 0.9933, highlighting its ability to integrate the robust decision-making of Random Forest (RF) with the efficient optimization of Stochastic Gradient Descent (SGD). These findings indicate that hybrid models improve accuracy compared to individual machine learning models and provide a more reliable solution for complex anomaly detection tasks by effectively mitigating their respective limitations.

TABLE 14. Accuracy of the hybrid models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
RF - QDA	0.9666	0.9691	0.8666	0.9887	0.9073	0.9745	0.8838
KNN - LR	0.9279	0.9014	0.6563	0.9460	0.925	0.9196	0.9460
RF - SGD	0.9681	0.9730	0.8774	0.9906	0.9813	0.9784	0.9936

TABLE 15. Precision of the hybrid models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
RF -QDA	0.9642	0.9618	0.9429	0.9870	0.8589	0.9622	0.8215
KNN - LR	0.9139	0.8772	0.6224	0.9246	0.9006	0.8961	0.9239
RF - SGD	0.9601	0.9697	0.9133	0.9844	0.9727	0.9683	0.9898

TABLE 16. Recall of the hybrid models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
RF -QDA	0.9731	0.9805	0.7962	0.9916	0.9870	0.9907	0.9972
KNN - LR	0.9537	0.9462	0.8916	0.9777	0.9648	0.9592	0.9787
SGD -RF	0.9805	0.9796	0.8490	0.9981	0.9925	0.9916	0.9981

TABLE 17. F1-Score of the hybrid models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
RF - QDA	0.9686	0.9711	0.8634	0.9893	0.9185	0.9762	0.9008
KNN - LR	0.9333	0.9104	0.7331	0.9504	0.9316	0.9266	0.9505
RF - SGD	0.9702	0.9746	0.8800	0.9912	0.9825	0.9798	0.9940

TABLE 18. ROC-AUC Score of the DL models evaluated for each dataset

	Dataset A	Dataset B	Dataset C	Dataset A +B	Dataset A + C	Dataset B + C	Dataset A + B+ C
RF - QDA	0.9662	0.9684	0.8673	0.9885	0.9023	0.9734	0.8767
KNN - LR	0.9263	0.8986	0.6416	0.9440	0.9225	0.9171	0.9440
RF - SGD	0.9673	0.9726	0.8792	0.9902	0.9806	0.9776	0.9933

Hybrid machine learning models enhance the accuracy of BGP anomaly detection systems by effectively combining the strengths of multiple algorithms to overcome individual model limitations (Atefi et al. 2020). The optimal accuracy model RF-SGD clusters the robust feature selection and interpretability of Random Forest (RF) with the optimization efficiency of Stochastic Gradient Descent (SGD), enabling the hybrid model to achieve high accuracy based on binary classification. Hybrid models can better address the high dimensionality and the nature of BGP datasets than standalone models by combining complementary capabilities, such as handling missing data and improving boundary definitions between anomaly and normal traffic classes (Bengani 2024).

DISCUSSION

The evaluation of machine learning (ML), deep learning (DL), and hybrid models for BGP anomaly detection across different datasets (A, B, C, A+B, A+C, B+C, and A+B+C) provided valuable insights into their performance. The analysis highlights the impact of dataset combinations on model accuracy, precision, recall, F1-score, and ROC-AUC, clarifying how data variety and volume influence anomaly detection capabilities. Combining datasets like A+B+C allowed hybrid models to achieve optimal accuracy, precision, and recall. This highlights the importance of designing comprehensive training datasets and employing hybrid approaches for highly accurate and reliable BGP anomaly detection systems.

Hybrid models significantly outperformed ML and DL models across all datasets, with the most significant improvements observed on challenging individual datasets such as C and its combinations. While dataset A presented a beneficial set with high-level accuracy, in contrast, dataset C emphasized the need for additional preprocessing or feature engineering to handle minor anomalies. Combining datasets mitigated these challenges by offering a more extensive context for model training. This demonstrates the ability of hybrid models to integrate complementary strengths and address complex scenarios more effectively.

Moreover, the deep models are able to produce higher performance metrics than the individual machine learning models. However, the most important factor is that execution time will increase dramatically, particularly when handling high-dimensional data such as BGP traffic. So, the computational cost, need for large labelled datasets, and challenges in interpretability and real-time application (Tyagi & Rekha 2020) render using deep learning models unsuitable for real-time detection systems. Therefore, considering the enhancement achieved in performance metrics, the hybrid models based on machine learning models are suggested as an optimal model for BGP anomaly detection systems.

CONCLUSION

The multi-stage methodology provides a comprehensive framework for analyzing BGP security using ML and DL approaches. The study identifies the most effective models for anomaly detection, emphasizing the importance of hybrid techniques in addressing the complexity of modern BGP attacks. The approach employed in this study has demonstrated a robust and systematic approach for enhancing the accuracy and reliability of BGP anomaly detection systems. This methodology ensures an optimized framework for anomaly detection by segmenting the process into several stages: data preprocessing, feature derivation, model development, evaluation, and hybrid integration. Each stage builds upon the previous one, facilitating the systematic identification of relevant features and the selection of the most effective models for integration. The feature selection process based on statistical analysis allowed for the isolation of significant BGP features, thereby enhancing the accuracy of both machine learning and deep learning models in subsequent stages.

Additionally, the study results underscore the potential of hybrid models to adapt to diverse data types within the BGP networks, as seen in the superior performance metrics of RF-SGD, which achieved 99.30% accuracy with a remarkable ROC-AUC of 0.9933. These findings make hybrid approaches suitable for large-scale networks where single-model solutions may struggle with computational or performance limitations. Hybrid machine learning models are particularly advantageous for BGP anomaly detection systems as they offer higher accuracy along with a robust, scalable, and adaptive framework for detecting complex network anomalies in dynamic environments. In future works, a comparative analysis will be conducted using the same methodological framework to allow direct comparison with prior research on global BGP events at the corresponding BGP network layer.

ACKNOWLEDGEMENT

We would like to thank the Iraqi Ministry of Communication (MOC) - ITPC company-Iraq for their assistance in collecting the information and data for this manuscript.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Al-Daweri, M. S., Abdullah, S. & Zainol Ariffin, K. A. 2021. An adaptive method and a new dataset, UKM-IDS20, for the network intrusion detection system. *Computer Communications* 180: 57–76.

Al-Musawi, B., Branch, P. & Armitage, G. 2017. BGP anomaly detection techniques: A survey. *IEEE Communications Surveys & Tutorials* 19(1): 377–396.

Alaghbari, K., Md Saad, M. H., Hussain, A. & Alam, M. R. 2022. Activities recognition, anomaly detection and next activity prediction based on neural networks in smart homes. *IEEE Access* 10: 28219–28232.

Arai, T., Nakano, K. & Chakraborty, B. 2019. Selection of effective features for BGP anomaly detection. *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*: 1–6.

Atefi, K., Hashim, H. & Khodadadi, T. 2020. A hybrid anomaly classification with deep learning (DL) and binary algorithms (BA) as optimizer in the intrusion detection system (IDS).

Azam, Z., Islam, P. M. & Huda, M. 2023. Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree. *IEEE Access* PP: 1–1.

Bengani, V. 2024. Hybrid learning systems: Integrating traditional machine learning with deep learning techniques.

Bezawada, B., Kulkarni, S. S. & Liu, A. X. 2008. Symmetric key approaches to securing BGP—A little bit trust is enough. *IEEE Transactions on Parallel and Distributed Systems* 22: 1536–1549.

Butler, K., Farley, T. R., McDaniel, P. & Rexford, J. 2010. A survey of BGP security issues and solutions. *Proceedings of the IEEE* 98(1): 100–122.

Cheng, M., Li, Q., Lv, J., Liu, W. & Wang, J. 2018. Multi-scale LSTM model for BGP anomaly classification. *IEEE Transactions on Services Computing* 14(3): 765–778.

Cosovic, M. & Slobodan, O. 2018. BGP anomaly detection with balanced datasets. *Tehnicki Vjesnik* 25: 766–775.

Dai, X., Wang, N. & Wang, W. 2019. Application of machine learning in BGP anomaly detection. *Journal of Physics: Conference Series* 1176: 032015.

Elsayed, M., Le-Khac, N.-A., Albahar, M. & Jurcut, A. 2021. A novel hybrid model for intrusion detection systems in SDNs based on CNN and a new regularization technique. *Journal of Network and Computer Applications* 191: 103160.

Flamia Azevedo, B., Rocha, A. & Pereira, A. 2024. Hybrid approaches to optimization and machine learning methods: A systematic literature review. *Machine Learning* 113.

Ha, A., Gunawan, T., Habaebi, M., Halbouni, M., Kartiwi, M. & Ahmad, R. 2022. CNN-LSTM: Hybrid deep neural network for network intrusion detection system. *IEEE Access* PP: 1–1.

Idowu, R., Maroosi, A., Muniyandi, R. & Othman, Z. 2013. An application of membrane computing to anomaly-based intrusion detection system. *Procedia Technology* 11: 585–592.

Kadhim, N., Chellappan, K. & Abdullah, N. 2025. BGP security analysis using network simulation: An impact study of cyber attacks. *Jurnal Kejuruteraan* 37: 1349–1362.

Hoarau, K., Tahiry Razafindralambo. 2021. BML—An efficient and versatile tool for BGP datasets. *IEEE International Conference on Communications*.

LeCun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *Nature* 521: 436–444.

Li, Q., Zhang, X., Zhang, X. & Su, P. 2015. Invalidating idealized BGP security proposals and countermeasures. *IEEE Transactions on Dependable and Secure Computing* 12(3): 298–311.

Liao, H., Murah, M. Z., Hasan, M. K., Aman, A., Fang, J., Hu, X. & Khan, A. U. R. 2024. A survey of deep learning technologies for intrusion detection in Internet of Things. *IEEE Access* PP: 1–1.

Markkandeyan, S., Ananth, A. D., Rajakumaran, M., Gokila, R. G., Venkatesan, R. & Lakshmi, B. 2025. Novel hybrid deep learning-based cyber security threat detection model with optimization algorithm. *Cyber Security and Applications* 3: 100075.

Mitseva, A., Panchenko, A. & Engel, T. 2018. The state of affairs in BGP security: A survey of attacks and defenses. *Computer Communications* 124.

Mujtaba, M., Nanda, P. & He, X. 2012. Border gateway protocol anomaly detection using failure quality control method. *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*: 1239–1244.

Ostertagova, E. & Ostertag, O. 2013. Methodology and application of one-way ANOVA. *American Journal of Mechanical Engineering* 1: 256–261.

Edwards, P., C., L. C. & Kadam, G. 2019. Border gateway protocol anomaly detection using machine learning techniques. *SMU Data Science Review* 2.

Park, H., Kim, K., Shin, D. & Shin, D. 2023. BGP dataset-based malicious user activity detection using machine learning. *Information* 14(9).

Scott, B., Johnstone, M. & Szewczyk, P. 2024. A survey of advanced border gateway protocol attack detection techniques. *Sensors* 24: 6414.

Scott, B. A., Johnstone, M. N., Szewczyk, P. & Richardson, S. 2024. Matrix profile data mining for BGP anomaly detection. *Computer Networks* 242: 110257.

Shtayat, M., Hasan, M. K., Sulaiman, R., Islam, S. & Rehman, A. 2023. An explainable ensemble deep learning approach for intrusion detection in industrial Internet of Things. *IEEE Access* PP: 1–1.

Sunita, M. & Mallapur, S. 2022. Optimal trained hybrid classifier for border gateway protocol anomaly detection. *International Journal of Swarm Intelligence Research* 13: 1.

Tripathy, S. & Behera, B. 2023. Performance evaluation of machine learning algorithms for intrusion detection system. *Journal of Biomechanical Science and Engineering*: 621–640.

Tyagi, A. & Rekha, G. 2020. Challenges of applying deep learning in real-world applications. In: 92–118.

Vinod, B. K. & Mahesh, S. C. 2023. A review of Petri net tools and recommendations. *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*: 710–721.

Yang, C. & Jia, W. 2023. BGP anomaly detection: A path-based approach.