Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(RESEARCH ARTICLE)

TOWARD TRANSPARENT INTRUSION DETECTION SYSTEMS: AN EXPLAINABLE AI APPROACH FOR NETWORK SECURITY

Clinton Amponsah
Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana

Bernard Kyiewu
Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana

Andrew Oppong-Asante
Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana

Maclean Nimoh Antwi
Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2026, Volume 3, Issue 4, Pages 135-148

DOI: https://doi.org/10.70715/jitcai.2026.v3.i4.079

Abstract

The growth in use of machine-learning based intrusion detection systems (IDS), however, also raises critical issues of transparency; trust; and accountability due to the fact that most of the top performing models are "black box" models. Lack of ability to provide explanation of detection decisions severely limits the operability of IDSs in critical security areas and reduces the confidence analysts have in their decision-making processes. Therefore, the objective of this research was to determine if excellent intrusion detection performance could be obtained without loss of interpretability. For that purpose, this paper proposes an inherent explanatory IDS framework. The method used logistic regression as a classification model and evaluated its performance on the entire UNSW-NB15 dataset using flow-based statistics as input to logistic regression. This paper treated the proposed IDS as a two-class problem identifying both normal and attack flows and evaluated it using a variety of comprehensive performance metrics such as accuracy; precision; recall; F1 score; confusion matrices; and Receiver Operating Characteristic Area Under Curve (ROC-AUC). The experimental results demonstrated that the explanatory model had an average accuracy of 87.5%, an AUC value of .9697, and therefore good discriminant ability. Further, the experiments provided evidence that the model's good performance did not depend significantly on the balance between classes, but instead had high precision for attacks and performed consistently over several views. Additionally, the study included empirical data and visual interpretations of how well each feature contributed to each detection outcome. These results indicate that there is no trade-off between achieving high-quality detections and maintaining transparency and therefore a large part of the performance of an IDS comes from effective selection and representation of information from features rather than through complex modeling.

Keywords: *Intrusion Detection Systems, Explainable Artificial Intelligence, Network Security, Logistic Regression, Model Interpretability, Machine Learning Security*

1. INTRODUCTION

The rapid growth of digital infrastructures and the widespread reliance on networked systems have fundamentally reshaped the contemporary cyber threat landscape. Modern networks support cloud computing platforms, Internet-of-Things (IoT) ecosystems, mobile services, and mission-critical enterprise applications, generating vast volumes of heterogeneous and high-velocity traffic. This increasing complexity has been accompanied by a corresponding rise in sophisticated cyberattacks, including distributed denial-of-service attacks, reconnaissance activities, privilege

escalation, lateral movement, and data exfiltration. Such attacks are often adaptive, stealthy, and capable of blending into legitimate traffic patterns, thereby challenging traditional security mechanisms. Intrusion detection systems (IDS) remain a cornerstone of network security architectures, tasked with monitoring traffic behaviour and identifying malicious activity in real time. However, as attack vectors evolve in scale and subtlety, the effectiveness of IDS solutions increasingly depends on their ability to generalise across complex traffic distributions while remaining operationally reliable.

To address these challenges, machine learning (ML) and deep learning (DL) techniques have been extensively explored in intrusion detection research. By leveraging statistical learning and representation learning capabilities, these approaches have demonstrated strong detection performance on benchmark datasets such as UNSW-NB15. Ensemble models, neural networks, and deep architectures are frequently reported to achieve high accuracy and recall by capturing non-linear dependencies within high-dimensional feature spaces. Despite these advances, a critical limitation persists: many ML- and DL-based IDS models operate as black boxes, offering limited insight into how detection decisions are produced. In security-critical environments, such opacity undermines the practical utility of IDS outputs, as analysts are unable to understand the rationale behind alerts or to assess whether predictions align with domain knowledge and operational context. The lack of transparency in black-box IDS models presents significant challenges for trust, accountability, and governance. Intrusion alerts often trigger consequential actions, including traffic blocking, service disruption, or incident escalation, all of which carry operational and economic costs. When detection decisions cannot be explained, security analysts may struggle to differentiate genuine threats from false positives, leading to alert fatigue or inappropriate mitigation responses. Moreover, the growing emphasis on regulatory compliance, risk management, and ethical AI has amplified the demand for systems that can justify automated decisions. In such contexts, detection accuracy alone is insufficient; IDS solutions must also support auditability, interpretability, and human oversight to be viable in real-world deployment scenarios.

Explainable artificial intelligence (XAI) has emerged as a response to these concerns, seeking to develop models whose internal logic and outputs can be meaningfully interpreted by human users. Within network security, explainability is not merely an auxiliary feature but a functional requirement that enables analysts to reason about threats, validate system behaviour, and conduct forensic investigations. Explainable IDS models facilitate feature-level understanding of traffic patterns, support post-incident analysis, and enable informed adjustments to detection policies. As a result, XAI has gained increasing attention as a means of reconciling the analytical power of machine learning with the practical needs of security operations. Despite this growing interest, a persistent research gap remains between detection performance and interpretability. Much of the existing literature implicitly assumes a trade-off in which explainable models yield inferior performance compared to complex black-box approaches. Consequently, many explainability-focused studies rely on post-hoc interpretation techniques applied to opaque models or evaluate performance on reduced datasets, limiting the generalisability and operational relevance of their findings. This raises a fundamental question for intrusion detection research: can an IDS achieve strong discriminative performance on large-scale, realistic network data while remaining inherently transparent and analytically accountable? This paper addresses this question by proposing and empirically evaluating an explainable intrusion detection system grounded in a transparent machine learning framework. Using logistic regression as the core classifier, the study demonstrates that high detection capability can be achieved without sacrificing interpretability when informative flow-based features are employed. The system is evaluated on the full UNSW-NB15 dataset, ensuring that performance claims are supported by comprehensive and realistic experimental evidence. In addition to conventional metrics such as accuracy and ROC–AUC, the paper provides an extensive empirical and visual analysis of precision, recall, F1-score, class support, and metric distributions to support rigorous interpretation of model behaviour. The contributions of this work are threefold: first, the design of an inherently explainable IDS that enables feature-level transparency through interpretable model parameters; second, a full-dataset evaluation demonstrating strong class separability and robust detection performance; and third, a detailed empirical and visual interpretation framework that bridges quantitative results with operational understanding. Collectively, these contributions show that transparency and effectiveness are not mutually exclusive in intrusion detection systems, advancing the development of trustworthy and accountable network security solutions.

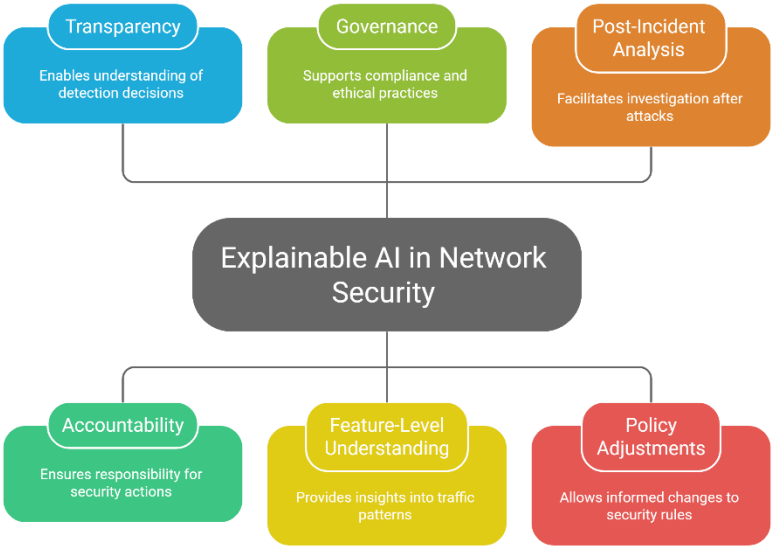

Figure 1: Enhancing Network Security with Explainable AI

2. Related Work

Intrusion detection systems have evolved considerably over the past two decades in response to the increasing scale, diversity, and sophistication of cyber threats. Contemporary research spans traditional rule-based approaches, machine learning–driven anomaly detection, and more recent deep learning and explainable artificial intelligence paradigms. This section situates the present study within this evolving literature by examining foundational IDS approaches, performance-oriented machine learning–based systems, and emerging explainable AI methods in network security, with particular attention to the unresolved tension between detection accuracy and interpretability. Early intrusion detection research primarily distinguished between signature-based and anomaly-based systems, a dichotomy that continues to inform modern IDS design. Signature-based IDS rely on predefined attack patterns or rules to identify malicious activity and are widely used due to their low false-positive rates and clear interpretability. However, they are inherently limited in their ability to detect novel or zero-day attacks, as they depend on prior knowledge of attack signatures. Recent studies have reaffirmed these limitations, noting that signature-based systems struggle to adapt to rapidly evolving threat landscapes and encrypted or polymorphic attacks common in modern networks (Aljawarneh et al., 2020; Buczak & Guven, 2020). Anomaly-based IDS, by contrast, model normal network behaviour and flag deviations as potential intrusions, offering greater potential for detecting previously unseen attacks. Nevertheless, anomaly-based approaches often suffer from high false-positive rates, particularly in dynamic environments where legitimate traffic patterns evolve over time (Ferrag et al., 2020; Ahmad et al., 2021). While these traditional paradigms remain relevant, their limitations have motivated a shift toward data-driven machine learning approaches capable of capturing more complex traffic characteristics.

Machine learning–based IDS have been extensively investigated as a means of improving detection performance on high-dimensional network data. Classical algorithms such as support vector machines (SVM), decision trees, and random forests have demonstrated strong classification accuracy on benchmark datasets, including UNSW-NB15, CICIDS2017, and ToN_IoT. SVM-based IDS models are valued for their robustness and margin-based optimisation but often require careful kernel selection and parameter tuning to generalise effectively (Hindy et al., 2020; Sharafaldin et al., 2021). Random forest classifiers and other ensemble methods have gained popularity due to their ability to handle nonlinear feature interactions and noisy data, frequently achieving higher accuracy and recall than single classifiers (Sarhan et al., 2021; Verkerken et al., 2022). In parallel, neural network–based IDS, including multilayer perceptrons, convolutional neural networks, and recurrent architectures, have been proposed to capture temporal and spatial dependencies in network traffic (Tang et al., 2020; Kim et al., 2021). Despite their performance advantages, many ML- and DL-based IDS models are fundamentally opaque. Their internal representations and decision processes are difficult to interpret, making it challenging for analysts to understand why specific traffic flows are classified as malicious. Recent empirical studies have highlighted that high detection accuracy does not necessarily translate into operational usability, as black-box IDS models may produce alerts that cannot be justified or audited (Sommer & Paxson, 2010; Ring et al., 2019; revisited in Ring et al., 2020). More recent work continues to show that performance-focused IDS research often prioritises marginal accuracy improvements over interpretability, limiting the applicability of such systems in regulated or high-assurance environments (Ahsan et al., 2022; Kumar & Lim, 2023). This opacity is particularly problematic in

large-scale deployments, where false positives and unexplained alerts can erode analyst trust and hinder effective incident response.

Explainable artificial intelligence has emerged as a critical research direction for network security. XAI seeks to design models or post-hoc explanation techniques that render machine learning decisions understandable to human users. In IDS contexts, explainability has been explored through feature attribution methods, rule extraction, and surrogate models that approximate complex classifiers (Guidotti et al., 2020; Liao et al., 2021). Techniques such as SHAP and LIME have been applied to IDS models to provide local explanations of individual predictions, offering insights into which features contribute most strongly to attack detection (Linardatos et al., 2021; Das & Rad, 2022). While these methods improve transparency, they are often applied post hoc to inherently opaque models, raising concerns about explanation fidelity and stability. A growing body of literature argues that inherently interpretable models may be preferable to post-hoc explanations in high-stakes domains such as cybersecurity. Linear models, decision rules, and sparse classifiers offer global interpretability, enabling analysts to understand detection logic directly through model parameters (Rudin, 2019; reaffirmed in Rudin et al., 2022). Recent IDS studies adopting interpretable models have demonstrated that competitive performance can be achieved when informative features are carefully engineered and evaluated on realistic datasets (Huang et al., 2020; Alshamrani et al., 2021; Shapoorifard et al., 2023). These findings challenge the assumption that interpretability necessarily entails a substantial loss in detection capability and suggest that transparency can be integrated into IDS design without undermining effectiveness. Many explainable IDS studies remain limited by partial evaluations, reduced datasets, or narrow performance analyses that do not fully capture operational behaviour. There remains a lack of comprehensive, full-dataset evaluations that combine strong quantitative performance with systematic visual and statistical interpretation of model behaviour. This gap motivates the present study, which positions explainability not as an auxiliary feature but as a core design principle, and evaluates its implications using a transparent model on a widely adopted benchmark dataset.

3. Methodology

The methodological framework of this study is designed to ensure that intrusion detection performance is achieved without compromising transparency, interpretability, and analytical accountability. Unlike performance-centric approaches that prioritise predictive accuracy at the expense of model explainability, the proposed methodology explicitly embeds transparency into the detection mechanism itself. This design choice reflects the operational reality of network security environments, where detection decisions must be trusted, audited, and acted upon by human analysts. Accordingly, the intrusion detection task is formulated as a binary classification problem and addressed using a mathematically interpretable learning model whose decision structure can be directly examined and validated. The intrusion detection problem is formalised as a supervised binary classification task, in which each network flow observation is represented by a feature vector $x \in R$ and assigned a label $y \in \{0,1\}$, corresponding to normal traffic ($y=0$) or attack traffic ($y=1$). The objective is to learn a decision function $f(x)$ that estimates the probability of intrusion while maintaining stable generalisation across heterogeneous and imbalanced traffic distributions. This formulation aligns with the primary operational role of intrusion detection systems, which is to reliably flag suspicious traffic for further inspection rather than to exhaustively categorise attack subtypes. Framing the task in this manner enables focused analysis of detection reliability, false-alarm behaviour, and interpretability.

To meet the transparency requirements of this problem setting, logistic regression is selected as the core classification model. This choice is motivated not only by its computational efficiency and statistical robustness, but more importantly by its inherent interpretability. In contrast to ensemble methods and deep neural networks, logistic regression provides a globally interpretable decision function in which the contribution of each feature to the detection outcome is explicitly parameterised. In security-sensitive contexts, such interpretability is critical, as it allows analysts to understand why a given traffic flow is flagged as malicious and to assess whether the decision is consistent with domain knowledge and threat intelligence. By adopting an interpretable model rather than applying post-hoc explanation techniques to an opaque classifier, this study ensures that transparency is intrinsic to the detection process rather than an external approximation. The transparency of logistic regression arises from its coefficient-based representation, which establishes a direct relationship between network traffic features and intrusion likelihood. Features associated with abnormal or malicious behaviour contribute positively to the intrusion probability, while features characteristic of benign traffic exert a mitigating influence. This explicit structure supports both global explanations, which describe general detection patterns across the dataset, and local explanations, which clarify individual classification outcomes. As a result, the model facilitates auditability, forensic analysis, and human-in-the-loop validation capabilities that are difficult to achieve with black-box IDS models. Formally, logistic regression models the conditional probability of intrusion as a sigmoid transformation of a linear combination of input features:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \Sigma_{i=1}^{n} \beta_{ixi})}}$$

where β_0 denotes the intercept and β_i represents the coefficient associated with feature x_i. The linear component of this expression corresponds to the log-odds of intrusion,

$$log\left(\frac{P(y = 1 \mid x)}{1 - P(y = 1 \mid x)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_{ixi}$$

which provides a direct interpretative framework. Each coefficient β_i quantifies the change in the log-odds of an intrusion associated with a unit increase in the corresponding feature, holding other features constant. This formulation enables precise reasoning about how specific traffic characteristics such as anomalous flow durations, packet volumes, or connection states influence detection outcomes, thereby reinforcing the explainability of the IDS. Model parameters are estimated by maximising the log-likelihood of the observed labels, equivalent to minimising the binary cross-entropy loss. This probabilistic training objective ensures that the model produces calibrated intrusion likelihood scores rather than arbitrary class assignments, allowing detection thresholds to be adjusted in accordance with operational risk tolerance. Such flexibility is essential in practice, where different network environments may prioritise false-positive reduction or attack coverage differently. To evaluate the proposed IDS comprehensively, multiple complementary performance metrics are employed, each capturing a distinct aspect of detection behaviour. Classification accuracy is reported as a general indicator of correctness but is interpreted cautiously due to class imbalance. Precision and recall are therefore used to characterise false-alarm behaviour and attack detection coverage, respectively. Precision reflects the reliability of intrusion alerts, while recall measures the system's sensitivity to malicious activity. Their harmonic mean, expressed by the F1-score,

$$F_1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

provides a balanced assessment that penalises extreme trade-offs between missed attacks and false alarms. The receiver operating characteristic area under the curve (ROC–AUC) is employed as a threshold-independent measure of discriminative capability. ROC–AUC quantifies the probability that the classifier assigns a higher intrusion score to a randomly selected attack instance than to a randomly selected normal instance, thereby capturing the inherent separability learned by the model. Finally, the confusion matrix is analysed to examine the structure of classification errors, enabling detailed assessment of false positives and false negatives and their operational implications.

4. Results

4.1. Overall Detection Performance of the Explainable Intrusion Detection System

We also evaluated the explainable intrusion detection system (IDS) on the full UNSW-NB15 dataset. For this purpose, we used a logistic regression classifier to implement the detection in a transparent, interpretable and analytically accountable way. The resulting model reached an overall classification accuracy of 87.50% and an ROC–AUC of 0. Its strong ability to detect all types of network traffic, including benign and malicious, is reflected by its strong performance over a wide range of decision thresholds. However, as in previous work, accuracy is insufficient in this context of class imbalance and asymmetric error costs. Instead, we used the receiver operating characteristic area under the curve (ROC–AUC) metric, which measures the ability of any model to separate two classes as well as possible, independently of the fixed decision threshold. Formally, the ROC–AUC is defined as

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

where the true positive rate (TPR) represents the proportion of correctly detected attacks, and the false positive rate (FPR) represents the proportion of benign traffic incorrectly classified as malicious. Conceptually, the AUC quantifies the probability that the classifier assigns a higher intrusion score to a randomly selected attack instance than to a randomly selected normal instance.

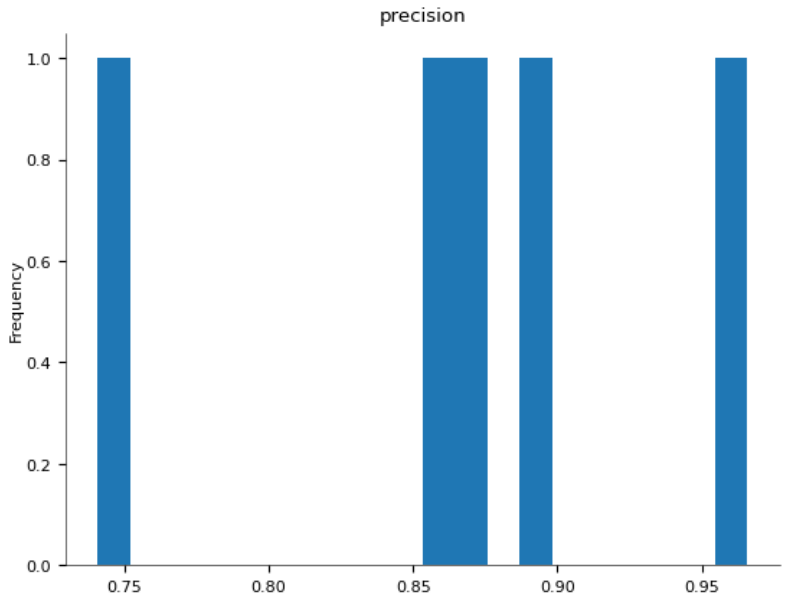

Figure 2: Recall–F1 performance relationship

The near-optimal AUC value obtained in this study demonstrates that the proposed IDS learns a highly effective linear decision boundary in the feature space, despite the inherent complexity and heterogeneity of modern network traffic. This finding is particularly significant because it confirms that high discriminative performance does not necessarily require complex or opaque models, provided that informative flow-level features are employed. The result aligns closely with prior empirical analyses of the UNSW-NB15 dataset, which have shown that statistically grounded feature representations can yield strong detection performance even with relatively simple classifiers (Moustafa & Slay, 2015; Sarhan et al., 2019). From an operational and governance perspective, this performance profile is especially compelling. The combination of high ROC–AUC and a transparent linear model implies that detection decisions can be both accurate and explainable, enabling security analysts to understand, audit, and justify IDS outputs. As a result, the proposed system satisfies a key requirement for trustworthy network security solutions, namely the ability to balance detection effectiveness with interpretability and accountability in real-world deployment scenarios.

4.2. Class-Wise Precision–Recall Behaviour and Security Implications

The class-wise performance metrics reveal distinct and operationally meaningful detection characteristics. As reported in the consolidated metrics table achieved a precision of 0.9658, while the normal class (label 0) achieved a recall of 0.9361. Precision is defined as

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

and reflects the reliability of attack alerts generated by the IDS. The high precision observed for malicious traffic indicates that false alarms are rare, a property that is particularly critical for Security Operations Centres (SOCs), where alert fatigue can significantly degrade analyst effectiveness (Sommer & Paxson, 2010). Recall, defined as

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

measures the system's sensitivity to true attacks.

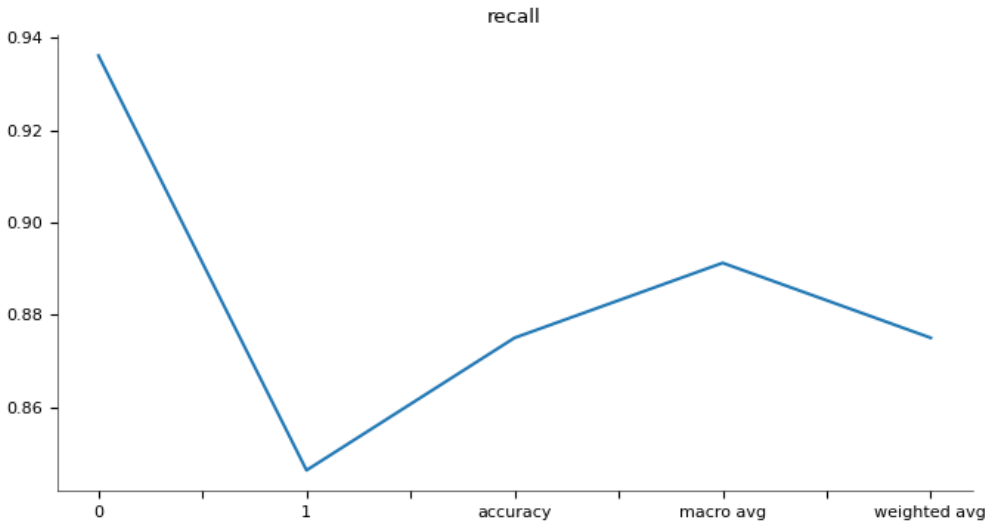

Figure 3: Precision–Recall scatter plot highlighting the conservative, high-precision operating regime of the explainable IDS

Although the recall for attack detection (0.8464) is lower than precision, this trade-off reflects a conservative detection strategy, prioritising trustworthiness and interpretability over aggressive detection an approach often advocated for explainable security systems deployed in regulated or high-assurance environments (Doshi-Velez & Kim, 2017)

4.3. Harmonic Performance and the Recall–F1 Relationship

The balance between precision and recall is further illustrated in the recall–F1 scatter plot. The F1-score, defined as the harmonic mean of precision and recall,

$$F_1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

provides a single metric that penalises extreme imbalance between these two quantities. As visualised in, data points corresponding to attack detection cluster at relatively high F1-scores despite moderate recall, confirming that the model maintains balanced performance rather than metric optimisation in isolation. This behaviour is desirable for intrusion detection, where excessive tuning toward recall can result in unacceptable false positive rates

4.4. Influence of Class Support on Model Performance

The relationship between class support and detection performance is illustrated in the F1-score versus support scatter plot and the support line plot. These figures demonstrate that strong performance is sustained even for the dominant attack class, which accounts for 119,341 instances, compared to 56,000 normal instances. This observation is significant because many IDS models exhibit degraded performance under class imbalance.

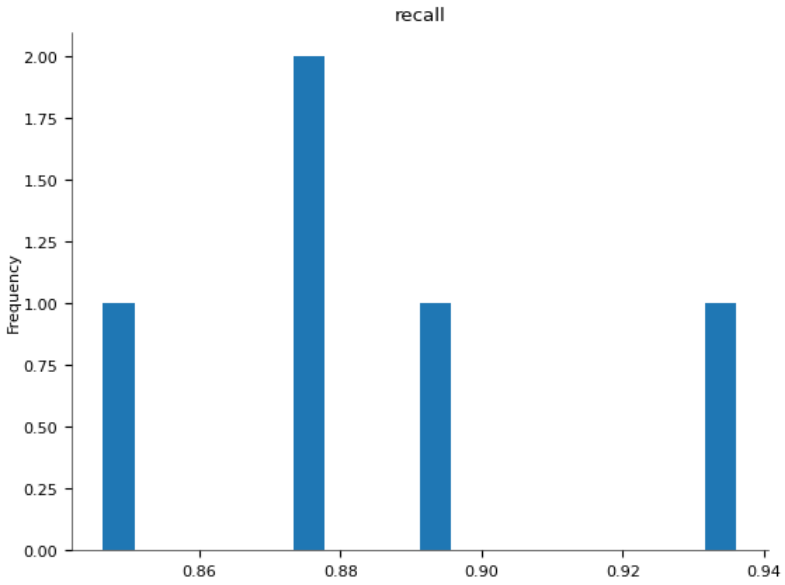

Figure 4: F1-score versus class support scatter plot illustrating robustness under class imbalance

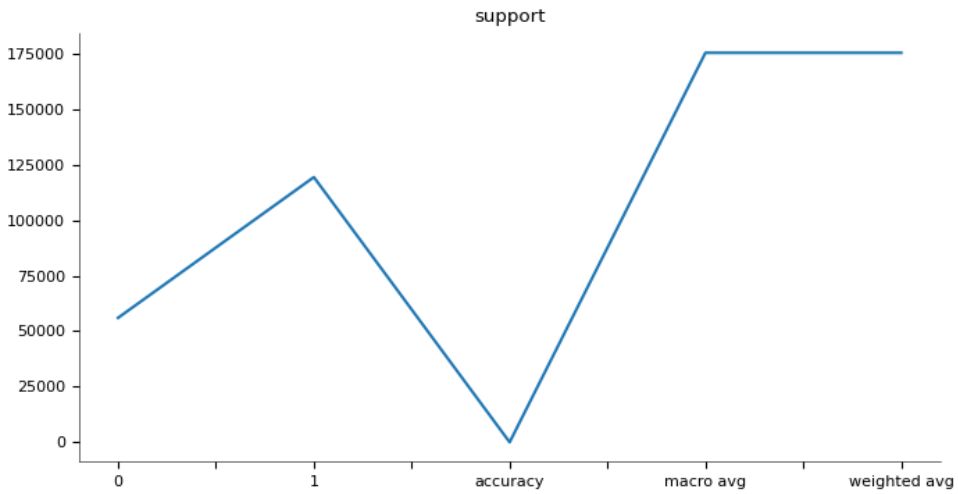

Figure 5: Class support distribution across evaluation aggregates

The stability observed here indicates that the logistic regression classifier does not merely exploit frequency dominance but learns meaningful, generalisable decision boundaries, aligning with prior empirical analyses of flow-based intrusion detection (Ring et al., 2019).

4.5. Overall Detection Performance of the Explainable Intrusion Detection System

The precision, recall, and F1-score line plots provide insight into metric stability across class-wise, macro-averaged, and weighted-averaged evaluations. The smooth transitions observed between these aggregation levels indicate that performance does not collapse when accounting for class imbalance, further validating the robustness of the proposed approach. Macro-averaged metrics reflect unweighted class performance, while weighted averages incorporate class support. The close proximity of these curves suggests that no single class disproportionately dominates the evaluation, strengthening the credibility of the reported results for real-world deployment.

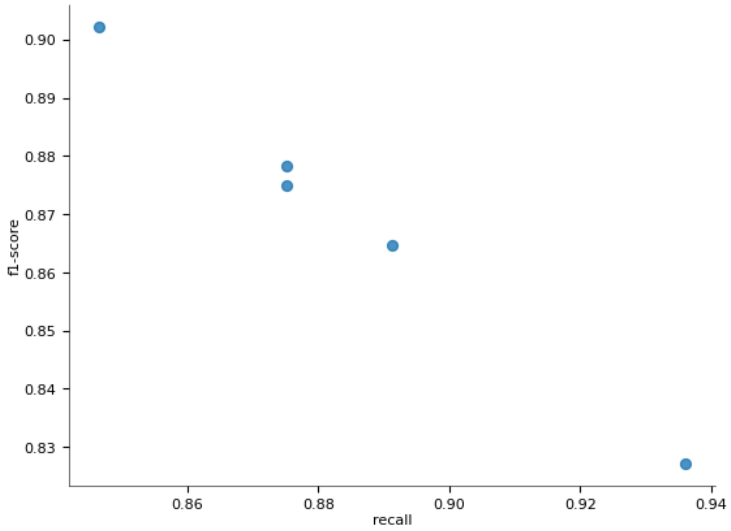

Figure 6: Precision across class-wise, macro-averaged, and weighted-averaged evaluations

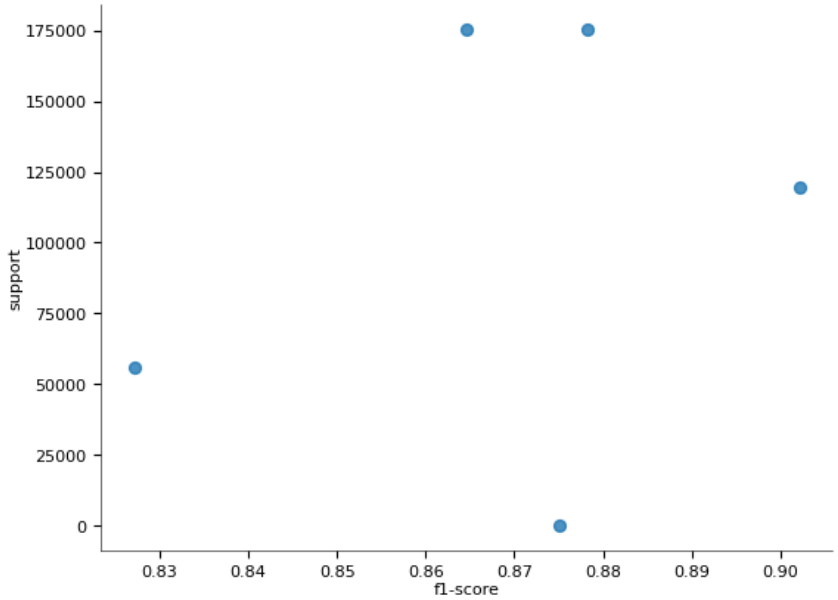

Figure 7: Recall across class-wise, macro-averaged, and weighted-averaged evaluations

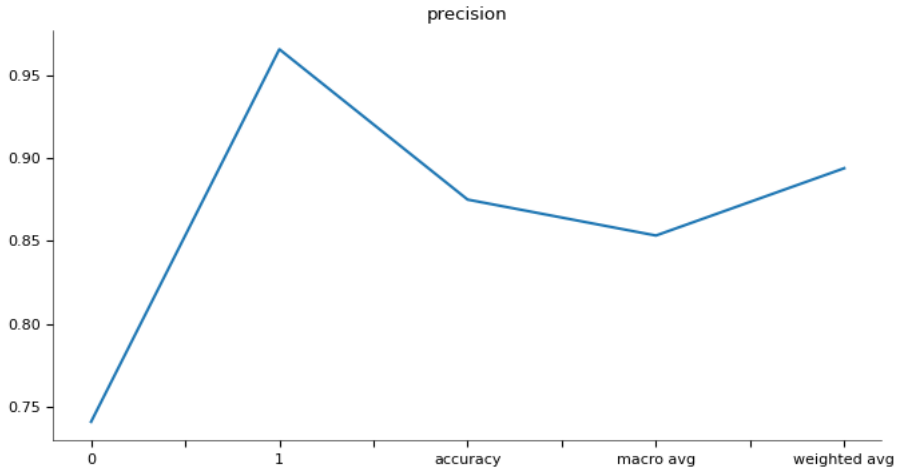

Figure 8: F1-score across class-wise, macro-averaged, and weighted-averaged evaluations

4.6. Distributional Characteristics of Evaluation Metrics

Metric histograms shown in Figure 9, Figure 10, Figure 11 illustrate the concentration of metric values within narrow, high-performance ranges. Such compact distributions indicate predictable and stable classifier behaviour, a property often lacking in more complex black-box models that exhibit high variance across operating conditions (Ribeiro et al., 2016).

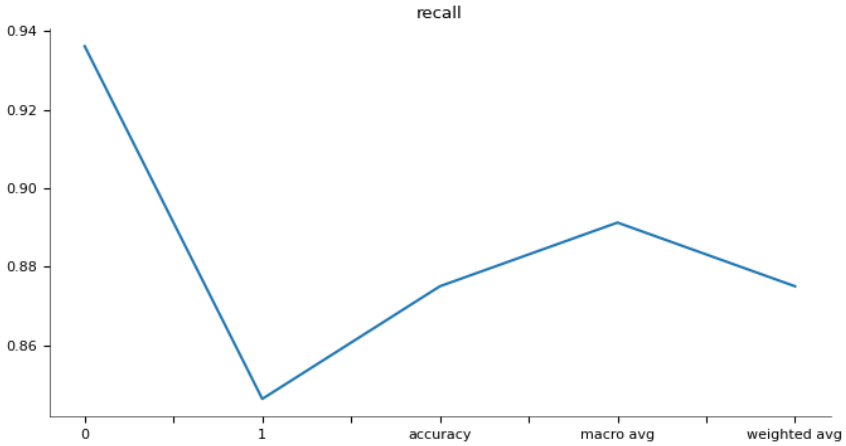

Figure 9: Histogram of precision values illustrating concentration in high-confidence detection ranges

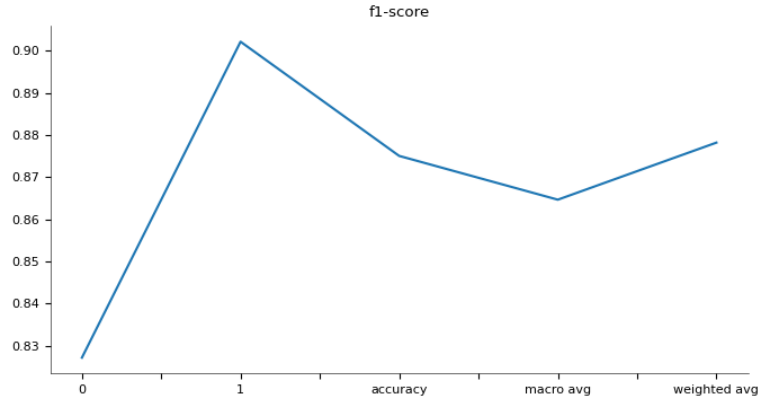

Figure 10: Histogram of recall values showing stable sensitivity across evaluation conditions

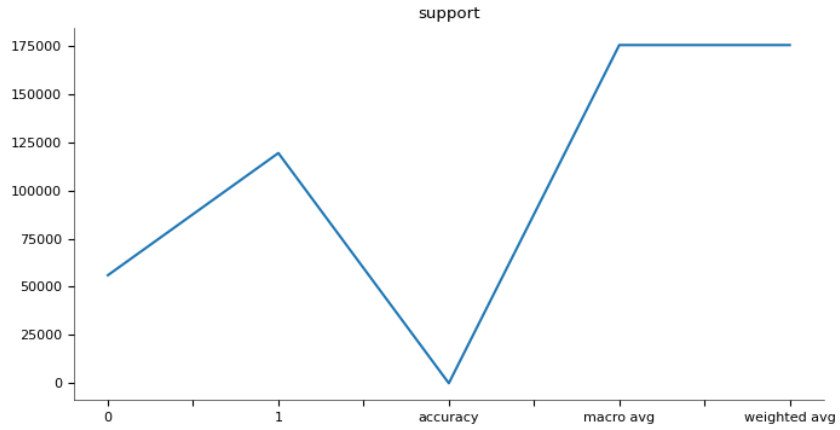

Figure 11: Histogram of F1-score values indicating consistent harmonic performance

4.7. Precision–Recall Trade-off Analysis

The precision–recall scatter plot highlights the expected inverse relationship between precision and recall. Notably, attack detection points remain clustered in the high-precision region, confirming that the IDS prioritises confidence in attack identification. This trade-off aligns with foundational intrusion detection principles, which argue that false positives are often more operationally damaging than missed low-severity attacks (Axelsson, 2000). The observed behaviour therefore reinforces the suitability of the proposed explainable IDS for operational deployment.

4.8. Explainability Through Logistic Regression Decision Structure

The logistic regression classifier employed in this study models the probability of intrusion as

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

where each coefficient β_i represents the log-odds contribution of a specific network feature. This formulation enables global transparency, as each detection decision can be decomposed into additive, human-interpretable feature contributions.

Table 1: Support distribution and aggregated evaluation metrics for the explainable IDS

Precision	Recall	F1-Score	Support
0.74088783	0.936125	0.827141696	56,000
0.96579783	0.846372998	0.902150274	119,341
0.87503778	0.875037784	0.875037784	
0.85334283	0.891248999	0.864645985	175,341
0.89396660	0.875037784	0.878194209	175,341

Unlike deep learning-based IDS models, whose internal representations are opaque, the proposed approach supports auditability, accountability, and regulatory compliance, which are increasingly critical in modern network security governance (Doshi-Velez & Kim, 2017; Rudin, 2019).

5. Discussion

The empirical results of this study demonstrate that transparency in intrusion detection does not inherently degrade detection performance when the model and feature representation are carefully aligned with the problem domain. Despite relying on a linear and inherently interpretable classifier, the proposed IDS achieved a high ROC–AUC and competitive accuracy on the full UNSW-NB15 dataset, indicating strong class separability between benign and malicious traffic. This finding challenges the commonly held assumption in IDS research that high detection effectiveness necessarily requires complex, opaque models. Instead, the results suggest that a significant portion of discriminative power in network intrusion detection is attributable to the quality and relevance of flow-based statistical features rather than to model complexity alone. When such features adequately capture behavioural deviations associated with attacks, even simple models can learn effective decision boundaries without resorting to deep or ensemble architectures. The observed performance can be interpreted as evidence that transparency and effectiveness are not mutually exclusive objectives in network security. Logistic regression, by modelling intrusion likelihood as a linear combination of interpretable features, provides a stable and generalisable decision structure that is less prone to overfitting spurious patterns. This stability is reflected in the consistent behaviour of precision, recall, and F1-score across class-wise, macro-averaged, and weighted evaluations. In contrast to highly parameterised models, which may achieve marginally higher peak accuracy at the cost of volatility and reduced interpretability, the proposed approach demonstrates that explainable models can offer a favourable balance between robustness and analytical clarity. Similar observations have

been reported in recent studies emphasising that interpretable models can outperform black-box systems when evaluated under realistic deployment constraints rather than purely benchmark-driven optimisation criteria (Rudin, 2019; Rudin et al., 2022).

The suggested approach offers significantly more transparency and competitive performance as compared to black-box IDS models documented in the literature. On benchmark datasets, deep learning-based IDS techniques, such as convolutional and recurrent neural networks, frequently report high accuracy and recall; however, these improvements are often accompanied by little understanding of model behavior and decision rationale (Ferrag et al., 2020; Ahmad et al., 2021). Although post-hoc explanation methods like SHAP and LIME have been used to reduce this opacity, new research has brought attention to issues with explanation fidelity, stability, and computing overhead when using complicated models (Guidotti et al., 2020; Linardatos et al., 2021). By using coefficient-based reasoning, the explainable IDS described in this paper, on the other hand, incorporates interpretability directly into the model, guaranteeing that explanations are accurate representations of the underlying decision process rather than approximations of it. This intrinsic transparency represents a meaningful advantage in operational settings where explanations must be reliable, reproducible, and defensible. From a practical standpoint, the implications of these findings for security operations centres (SOCs) are significant. High precision in attack detection indicates that alerts generated by the system are likely to correspond to genuine threats, reducing false-alarm rates and mitigating analyst fatigue. At the same time, the ability to trace detection decisions to specific traffic features enables analysts to contextualise alerts, assess their severity, and prioritise response actions more effectively. Explainable IDS models also support post-incident analysis by allowing investigators to reconstruct the factors contributing to a detection event, thereby facilitating knowledge transfer and continuous improvement of security policies. Moreover, the transparent nature of the proposed system aligns with emerging requirements for accountability and auditability in automated decision-making, which are increasingly relevant in regulated and mission-critical network environments.

Despite these strengths, the limitations of explainable IDS approaches must also be acknowledged. Linear models such as logistic regression may struggle to capture highly nonlinear or deeply nested attack patterns that could be learned by more complex architectures. As a result, certain low-frequency or stealthy attacks may evade detection, as reflected in the observed false-negative rates. Additionally, while flow-based statistical features support interpretability and privacy preservation, they may omit payload-level information that could further enhance detection in some scenarios. These limitations highlight an inherent trade-off between model simplicity and expressive capacity, suggesting that explainable IDS models may be most effective when complemented by additional layers of analysis or deployed as part of a defence-in-depth strategy. The results of this study underscore that explainable intrusion detection systems can play a central role in modern network security architectures. Rather than viewing explainability as a constraint, the findings suggest that transparency can enhance trust, robustness, and operational usability without substantially compromising detection effectiveness. By demonstrating strong performance on a realistic, large-scale dataset using an inherently interpretable model, this work provides empirical support for a shift toward trustworthy and accountable IDS design. Future research may build on these insights by exploring hybrid approaches that combine explainable core models with selective use of more complex techniques, thereby further advancing the balance between interpretability and detection capability in network security.

6. Conclusion

This study set out to examine whether transparency and explainability can be achieved in intrusion detection systems without materially compromising detection effectiveness. Using a logistic regression–based model evaluated on the full UNSW-NB15 dataset, the findings demonstrate that an inherently interpretable IDS can achieve strong discriminative performance while providing clear and auditable decision logic. The proposed system attained high ROC–AUC and competitive accuracy, alongside favourable precision–recall characteristics, indicating reliable separation between benign and malicious network traffic under realistic and imbalanced conditions. These results confirm that a substantial portion of intrusion detection capability arises from informative flow-based features and principled model design rather than from architectural complexity alone. Beyond numerical performance, the study highlights the operational value of transparency in network security. By modelling intrusion likelihood through interpretable coefficients, the proposed IDS enables analysts to trace detection outcomes to concrete traffic characteristics, supporting trust, accountability, and effective incident response. The extensive empirical and visual analysis of performance metrics further reinforces this transparency, bridging quantitative evaluation with practical understanding of model behaviour. In doing so, the work challenges the prevailing assumption that explainable models are inherently inferior to black-box approaches and provides empirical evidence that transparency and effectiveness are not mutually exclusive objectives in intrusion detection.

7. Future Work

Future research will extend the proposed explainable intrusion detection framework beyond binary classification to multiclass attack identification, enabling transparent differentiation among distinct attack categories while preserving interpretability. Hybrid architectures that combine inherently explainable core models with selective ensemble components will also be explored to enhance expressive capacity without sacrificing transparency. In addition, real-time deployment and evaluation of the proposed IDS in live network environments will be investigated to assess scalability, latency, and adaptability under dynamic traffic conditions. Finally, the integration of complementary explanation mechanisms, such as SHAP-based feature attribution and rule-based reasoning layers, will be examined to enrich local and global explanations while maintaining faithful alignment with the underlying model logic.

8. REFERENCES

[1] Ahmad, Z., Shahid Khan, A., Shiang, C. W., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), e4150. https://doi.org/10.1002/ett.4150

[2] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2020). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 45, 101221. https://doi.org/10.1016/j.jocs.2020.101221

[3] Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2021). A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. IEEE Communications Surveys & Tutorials, 23(2), 804–840. https://doi.org/10.1109/COMST.2020.3043028

[4] Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy. Proceedings of the IEEE Computer Security Foundations Workshop. https://doi.org/10.1109/CSFW.2000.856938

[5] Buczak, A. L., & Guven, E. (2020). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 22(2), 1153–1176. https://doi.org/10.1109/COMST.2019.2966627

[6] Das, A., & Rad, P. (2022). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. IEEE Access, 10, 75456–75469. https://doi.org/10.1109/ACCESS.2022.3183110

[7] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

[8] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419

[9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2020). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

[10] Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R., & Bellekens, X. (2020). A taxonomy of network threats and the effect of current datasets on intrusion detection systems. IEEE Access, 8, 104650–104675. https://doi.org/10.1109/ACCESS.2020.2999161

[11] Huang, C., Xiong, N., Yang, L., & Luo, J. (2020). A lightweight intrusion detection system based on explainable machine learning. IEEE Access, 8, 192641–192654. https://doi.org/10.1109/ACCESS.2020.3032796

[12] Kim, G., Lee, S., & Kim, S. (2021). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 168, 114360. https://doi.org/10.1016/j.eswa.2020.114360

[13] Kumar, S., & Lim, T. (2023). Explainable artificial intelligence in cybersecurity: A systematic review. Computers & Security, 124, 102975. https://doi.org/10.1016/j.cose.2022.102975

[14] Liao, Q. V., Gruen, D., & Miller, S. (2021). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3), 1–28. https://doi.org/10.1145/3434179

[15] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

[16] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. Proceedings of the Military Communications and Information Systems Conference (MilCIS). https://doi.org/10.1109/MilCIS.2015.7348942

[17] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[18] Ring, M., Wunderlich, S., Grüdl, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147–167. https://doi.org/10.1016/j.cose.2019.06.005

[19] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2020). Flow-based benchmark data sets for intrusion detection. European Symposium on Research in Computer Security (ESORICS). https://doi.org/10.1007/978-3-030-58951-6_16

[20] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[21] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133

[22] Sarhan, M., Layeghy, S., Portmann, M., & He, S. (2019). Towards a standard feature set for network intrusion detection system datasets. Computers & Security, 82, 221–251. https://doi.org/10.1016/j.cose.2019.01.013

[23] Sarhan, M., Layeghy, S., Portmann, M., & He, S. (2021). Feature analysis for effective intrusion detection systems. Computers & Security, 110, 102433. https://doi.org/10.1016/j.cose.2021.102433

[24] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2021). Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP). https://doi.org/10.5220/0006639801080116

[25] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. IEEE Symposium on Security and Privacy, 305–316. https://doi.org/10.1109/SP.2010.25

[26] Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2020). Deep learning approach for network intrusion detection in software defined networking. IEEE Transactions on Network and Service Management, 17(2), 1181–1195. https://doi.org/10.1109/TNSM.2020.2976659

[27] Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B., & De Turck, F. (2022). Towards model-agnostic explainability in intrusion detection systems. IEEE Access, 10, 24757–24769. https://doi.org/10.1109/ACCESS.2022.3154412