IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 5, Ser. II (Sep. – Oct. 2026), PP 01-11
www.iosrjournals.org

Explainable AI Techniques for Network Traffic Anomaly Detection and Mitigation: *A Systematic Review.*

Fumlack Kingsley George
achieverking646@gmail.com
fumlackkingsley@student.mau.edu.ng

Murtala Mohammed (Ph.D)
mmurtala1875@mau.edu.ng

Abstract
The modern networks have become more complex making it difficult to detect anomalies in real time and reduce congestion. Conventional machine learning models, however precise, are black boxes, which do not provide much information about the way they make decisions. To solve this, Explainable Artificial Intelligence (XAI) offers transparency, interpretability and accountability in network monitoring. This paper locate, assess and contrast XAI methods used in network anomaly detection with a particular focus on model transparency and point out the research gaps. Research articles published in 2020-2025 in IEEE Xplore, Scopus, SpringerLink, and ScienceDirect were systematic review articles that are peer-reviewed and analyzed using the PRISMA framework. The results demonstrate that SHAP, LIME, Grad-CAM, attention mechanisms and rule-based explanations have been the most used, each with its strengths and trade-offs of computation. Model-agnostic approaches have greater interpretability, but are computationally expensive, whereas model-specific approaches are architecture-specific, and thus can explain faster. Real-time explainable frameworks and standardized interpretability metrics are some of the research priorities identified in the review. In general, explainability is also placed as a fundamental pillar of trustful AI, which positively influences transparency, user trust and responsibility of intelligent network management systems.

Keywords: *Explainable AI, Network Anomaly Detection, SHAP, LIME, Grad-CAM, Network Optimization.*

Date of Submission: 06-09-2026 Date of Acceptance: 16-09-2026

I. INTRODUCTION

1.1 Background and Motivation

Network anomaly detection can be defined as a process of recognizing abnormal patterns or deviation of network traffic which could be a sign of security breach, a network congestion and a network fault. As a part of contemporary enterprise network that manages extensive data volumes and accommodates essential digital processes, anomaly-detecting incurs a crucial role in providing reliability, data stability, and security. It allows preventing threats before they happen, including distributed denial-of-service (DDoS) attacks, data exfiltration, or performance loss caused by traffic overloads (Hnamte *et al.,* 2024). With the fast development of networks in terms of dynamism and complexities, which are propelled by cloud computing, Internet of Things (IoT) devices, and real-time communications, intelligent detection of anomalies has become a necessity. The concept of artificial intelligence (AI) and machine learning (ML) models as an effective method to analyze large-scale network data has risen as the means to predictively detect and automatically mitigate unusual activities that were traditionally hard to handle with rule-based tools.

However, most AI-based network anomaly detection models are black boxes, even though they have been successful in reaching high accuracy, which makes them less or not transparent when it comes to how they make decisions. Such opaque models complicate matters whereby network administrators are not able to tell why some traffic is categorized as anomalous and this affects trust and responsibility and more so in critical infrastructures where the human factor would be needed. The uninterpretability also makes it more difficult to troubleshoot, validate the model, and comply with the regulations as decisions cannot be easily justified and explained. Thus, a stronger desire to develop Explainable Artificial Intelligence (XAI) methods, which improve the transparency of AI models by showing how and why they give certain conclusions, is increasing (Minh *et al.,* 2022). XAI enhances trust and human-AI collaboration by making the process of interpretation more interpretable

and open, as well as aiding the deployment of intelligent systems in contemporary enterprise network settings in a safer way.

1.2 Development of Explainable AI in Network Engineering

The traditional machine learning (ML)-based detection to the explainable artificial intelligence (XAI)-based interpretability in the network engineering is a significant transition in focusing not on the accuracy of an automation, but on the transparency-based intelligence. Firstly, Support Vector Machines (SVM), Random Forests, and Deep Neural Networks (DNNS) were used as ML techniques to detect anomalies, classify traffic, and predict intrusion in network. These models were highly accurate in their detection yet they were characterized as black boxes providing minimal information on the impacts that certain traffic patterns or features had on their output. To address this shortcoming, explainable AI approaches were proposed, including SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), Grad-CAM (Gradient-weighted Class Activation Mapping) as well as attention mechanisms, to improve the transparency of models (Aysel et al., 2025). These visualization and interpretation tools allow critical network characteristics, including packet anomalies, changes in latency, and congestion triggers, to be visualized and interpreted based on importance scores of input attributes or regions of influence of data representations. Consequently, XAI is not only making automated network systems more trustworthy and accountable, it also enables engineers to more confidently and clearly diagnosis faults, validate AI behavior and optimize network performance.

1.3 Aim and Objectives of the Study

Aim:

To completely survey the available explainable artificial intelligence (XAI) methods deployed to network traffic anomaly detection and mitigation, to emphasize on their efficacy, intelligibility, and practical applicability to network management by intelligent means.

Objectives:

1. **To identify and classify** the explainable AI methods (e.g., SHAP, LIME, Grad-CAM, attention mechanisms) used in network anomaly and congestion detection.
2. **To evaluate** the performance and interpretability trade-offs between black-box and interpretable machine learning models in network traffic analysis.
3. **To examine** how XAI enhances transparency, trust, and decision-making in automated network monitoring systems.
4. **To identify** research gaps, challenges, and future directions in the application of explainable AI for network traffic optimization and mitigation.

1.5 Research Gap and Significance

However, although there is a great achievement in using artificial intelligence and machine learning in network anomaly detection, there is a huge gap of explainability frameworks in the existing research. The majority of existing models are based on the primary consideration of the detection accuracy and response time, but they do not pay attention to the transparency of the models in their decision-making. This drawback complicates the perception of network administrators on why particular anomalies were detected and what decisions to mitigate are stemmed, leading to a loss of trust in AI-based systems. Besides, numerous machine learning systems applied to detect anomalies continue to remain opaque black boxes, providing a small amount of interpretive information about traffic dynamics and the factors that collectively lead to a network crash or a traffic jam.

The other key gap is the lack of developed markets to assess the interpretability in network models. Although measures of accuracy, precision, and recall are common in performance measures, few studies use quantifiable metrics of explainability, including but not limited to feature importance consistency, human interpretability scores or computational transparency. This unification of evaluation frameworks impedes the objective comparison between XAI methods and researchers cannot tell which approaches are most successful in striking a balance between interpretability and performance. Consequently, explainable AI has not been adopted in practical network settings in a fast and evenly distributed way.

The paper will help advance the idea of transparent and responsible AI in network management through a systematic review of explainable AI-based methods in anomaly detection and congestion mitigation. The study supports a systematic knowledge of how interpretability can boost trust, safety, and efficiency of making decisions in automated network systems by identifying the strengths, weaknesses, and application situations of the available XAI models. Eventually, it facilitates a transition to responsible AI models that do not only identify and suppress network anomalies before they occur but are also capable of justifying its rationale in a manner that is comprehensible and actable by human operators.

II. Literature Review

2.1 Conceptual Clarifications

a) Explainable Artificial Intelligence (XAI): This is a collection of procedures and structures that are able to render the decision making procedure of artificial intelligence systems comprehensible to human beings. In contrast to the black-box models of earlier times that provide insufficient information about the mechanism of generating the outputs, XAI offers transparency and shows the logic behind AI predictions or classifications. XAI enables users to comprehend, trust, and justify AI behavior through visualization, ranking feature importance, and extracting rules, thus it is as important in sensitive uses of AI such as network management and cybersecurity.

b) Network Anomaly Detection: refers to the activity where the abnormal or irregularity patterns in the network traffic are detected that are not usual. The anomalies usually suggest possible problems like cyber attacks, intrusions, misconfigurations or congestion. Conventional methods of detection are based on pre estimated rules or thresholds, whereas the innovative AI-based detection is based on statistical learning and recognition of patterns as complex and changing threats that can be detected automatically and promptly. Good anomaly detection makes enterprise network systems reliable, available and intact.

c) Network Congestion: this happens when the amount of data sent out exceeds the capacity of a network and causes delays, loss of packets and poor performance. Congestion is usually as a result of high user demand, lack of proper allocation of resources or abrupt burst of traffic. Within the framework of intelligent network management, the models of AI and XAI are essential to predicting, locating, and preventing congestion through its proactive prevention by monitoring traffic movements, detecting possible congestion points, and recommending adaptive routing or load-balancing solutions.

d) Interpretability in AI: can be described as the extent to which a human can understand the internal mechanism and the logic of decision making of a model. It entails the knowledge of the features that contribute to prediction of a model and the alterations that happen when there are changes in input data that affect outputs. Increased interpretability will result in accountability, increase trust, and enable network engineers to test and debug AI-based decisions. In network anomaly detection, interpretability fills the gap between model expertise and human expert knowledge, making the automated systems of proficiency and morality reliable and ethical.

Table 1: Key Concepts Related to Explainable AI in Network Anomaly Detection

Concept	Definition	Key Features	Relevance to the Study
Explainable Artificial Intelligence (XAI)	A branch of AI focused on making the decision-making process of complex models understandable to humans.	- Enhances model transparency and trust.- Uses techniques such as SHAP, LIME, Grad-CAM, and attention mechanisms.- Provides both local and global interpretability.	Ensures transparency and accountability in AI-based network anomaly detection systems, allowing human operators to understand and validate AI decisions.
Network Anomaly Detection	The process of identifying abnormal or suspicious patterns in network traffic that deviate from normal behavior.	- Detects intrusions, faults, or congestion.- Uses ML models to analyze traffic flows.- Can operate in real time for proactive mitigation.	Helps identify and mitigate performance threats or cyberattacks, forming the core application area for explainable AI models.
Network Congestion	A condition where network demand exceeds available capacity, leading to latency, packet loss, and reduced throughput.	- Caused by high data load or inefficient routing.- Impacts Quality of Service (QoS) and user experience.- Requires adaptive and predictive management.	A key performance issue that explainable AI models can detect early and mitigate efficiently through interpretable decision-making.
Interpretability	The degree to which a human can comprehend how an AI model arrives at a specific output or decision.	- Facilitates human oversight and debugging.- Ensures model accountability.- Involves visualization and feature importance analysis.	Enhances the transparency of network management systems by making AI-driven anomaly detection and mitigation decisions understandable and traceable.

2.2 Explainable AI Techniques in Network Detection

The use of explainable AI (XAI) methods is gaining popularity in network anomaly detection to increase the levels of transparency, understandability, and confidence in automated systems. Some of the most used techniques include SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), Grad-CAM (Gradient-weighted Class Activation Mapping) and attention-based visualizations. SHAP, used to explain individual model predictions, is able to provide contribution values to each feature, showing the effect that an attribute like the size of the packet or the duration of flow has on the classification of an anomaly. LIME does so but creates low-fidel local models approximating the behaviour of a set of complex models in a particular region of decision. Grad-CAM and attention-based methods are especially useful in deep learning models, where they give heatmaps or visual focus map, which show the strongest network components or data parts that play a role in detecting the anomaly (Raghavan et al., 2024).

These explainable methods have different interpretability and complexity of computation based on the type of underlying model and context of use. As an example, SHAP gives feature importance explanations that are globally consistent, but is computationally expensive, particularly when dealing with large datasets or network traffic represented by high-dimensional networks. LIME, which is quicker, is concerned only with local interpretability and can give inconsistent explanations of different samples. Conversely, Grad-CAM and attention-based visualizations are efficient from a computational point of view with respect to neural networks, but restricted to visual interpretability in accessing network patterns, which may necessitate domain knowledge to convert patterns into network action. In other cases, surrogate models (i.e. decision trees, or linear regressors) can be trained to act in the proposed manner of black-box models, providing faster results and more interpretable explanations at the expense of fidelity (Mor 2025).

The explainability method used is thus a trade-off between transparency, accuracy and efficiency. Methods with high interpretability are more likely to simplify a model and are prone to impairment of predictive accuracy whereas highly accurate models, e.g., deep neural networks can tend to lack transparency to complexity. In network detection, it is important to strike the right balance-the models should not simply identify anomalies but also make sense in a way that can allow human intervention to be taken promptly and informedly. Such a balance will make explainable AI reliable in undertaking mission-critical tasks, including intrusion prevention, network congestion management, as well as real-time optimization of network performance (Zhang 2025).

Table 2: xplainable AI Techniques Applied in Network Anomaly Detection

XAI Technique	Description	Typical Network Applications	Advantages	Limitations
SHAP (Shapley Additive Explanations)	Uses game theory to assign contribution values to each input feature, explaining their influence on model predictions.	Feature importance analysis for anomaly detection, congestion diagnosis, and intrusion classification.	Provides global and local interpretability; model-agnostic; consistent feature ranking.	Computationally expensive for large datasets and complex models.
LIME (Local Interpretable Model-Agnostic Explanations)	Generates a simplified local surrogate model to approximate the behavior of a complex black-box model around a specific prediction.	Local explanation of anomaly or attack detection results in real-time systems.	Fast, easy to implement, and model-agnostic; improves understanding of individual predictions.	Explanations may vary between runs; less accurate for global interpretability.
Grad-CAM (Gradient-weighted Class Activation Mapping)	Visualizes the important regions in input data that contribute most to a neural network's decision.	Visual interpretation of CNN-based intrusion or traffic pattern recognition systems.	Effective for deep learning models; provides intuitive heatmaps highlighting influential features.	Limited to visual data or image-like representations; less suitable for tabular traffic data.
Attention Mechanisms	Assigns varying importance weights to input features or sequence elements during model training.	Sequence-based network traffic analysis, time-series anomaly prediction, and packet flow prioritization.	Enhances interpretability in recurrent or transformer models; highlights critical temporal or spatial features.	Computationally intensive; interpretability depends on visualization clarity.
Surrogate Models (e.g., Decision Trees, Linear Models)	Builds a simpler, interpretable model to mimic the behavior of a complex black-box model.	Explaining complex deep learning or ensemble-based network anomaly detectors.	Easy to understand and visualize; computationally efficient for post-hoc explanations.	May lose fidelity compared to the original model; oversimplifies nonlinear relationships.
Rule-Based Explanations	Extracts human-readable decision rules from trained models or datasets.	Policy auditing, security rule validation, and automated decision justification.	High human interpretability; aligns with expert domain reasoning.	Difficult to scale with complex data; may not capture subtle feature interactions.

Explainable AI (XAI) methods vary greatly in their appropriateness in real-time network anomaly detection by the interpretability degree, computing expenses, and compatibility with models. SHAP and LIME are common in post-hoc explanations because they are model-independent, thus they are the best in examining feature importance on the supervised detection systems, although SHAP is more computationally expensive. Grad-CAM and attention processes are more useful in deep learning models as they offer visually intuitive representations of how neural networks recognize complex patterns of traffic in space or time, but their high processing demands make them impossible to implement in high-latency settings. Rule-based explanations and surrogate models provide easier and faster interpretability as complex model behavior can be translated into interpretable forms that are easy to understand by a human, and alternative surrogate models can be used to scale to operational decision support and compliance audit. They can however trade off some accuracy or be unable to capture more complex non-linear dependencies contained in high dimensional network data. In general, the choice of the right XAI method is based on the tradeoff equation between interpretability and the efficiency of the computation where the explanation must remain actionable to the engineers, and practical to real-time network management.

2.3 Black-Box vs Interpretable Models

Artificial intelligence models used in network anomaly detection have been divided into black and interpretable models according to transparency. Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) are examples of black-box models that are highly accurate in identifying complex and nonlinear patterns in large network-data sets. Such models are capable of automatically acquiring complex correlations between such features as the packet size, frequency, and traffic flow behavior, which is useful in intrusion detection, as well as predicting congestion. Nonetheless, their decision-making mechanism inside is frequently black-box, i.e. network engineers are unable to trace the degree to which specific anomalies are raised, and risk scores are pegged (Beckley, 2025). This non-transparency creates issues of accountability, debugging, and trust particularly in mission-sensitive or sensitive security related environments where human supervision is necessary.

Conversely, interpretable models like Decision Trees, Random Forests and Logistic Regression place more emphasis on transparency and simplicity thus allowing users to see how attributes form part of particular predictions. An example of this is that a Decision Tree can easily indicate that the high latency and packet loss beyond a certain threshold is an anomaly. Yet, the models can be inaccurate or generalizable in highly dynamic network environments (Han et al., 2021). Comparison case studies of fairness between deep learning and tree-based algorithms, e.g., network intrusion detection with CNNs and Decision Trees, show that there is always a gap in their interpretability, with CNNs being more accurate at detecting data but Decision Trees giving more understandable way to be. Equally, research with the use of Random Forests has demonstrated better interpretability as compared to DNNs but smaller datasets do not scale well. Such differences demonstrate the necessity of Explainable AI (XAI) tools to fill in this gap, with the high performance of black-box models and high interpretability of transparent frameworks to ensure the reliability and accountability of network management.

2.4 Theoretical Framework

The theoretical background of the present research is based on the human-centered AI theory, which focuses on creating intelligent systems that are transparent, ethical and in sync with human logic. Human-centered AI promotes that the artificial intelligence must not only be able to carry out tasks on its own, it must be able to justify its choices in a manner that is easily comprehensible and reliable to users. The principle is also very similar to the objectives of Explainable AI (XAI), which makes the models employed in network anomaly detection explainable and responsible. XAI enhances trust in AI systems by making them transparent and interpretable to the corresponding network engineers and administrators who are able to verify, challenge, and refine the results of their models thus minimizing the risks of opaque automation in network administration (Li *at al.,* 2021).

The explainability aspect of network detection frameworks is also supported by the transparency and accountability theories. Transparency makes available the inner workings of AI systems, which open to observation and traceability and justification, whereas accountability makes algorithmic decisions and outcomes responsible. In the network anomaly detection framework, these theories would make sure that automated systems also have explanatory logic behind their alerts or mitigation measures. This is important in networks that are at enterprise level since false alarms, classification errors or unjustified mitigation measures may affect operations. The possibility to follow the reasoning process of a model promotes compliance with the upcoming standards of ethical AI implementation in the critical infrastructure and facilitates user trust and governance (Olawore *et al.,* 2025).

Lastly, the operational basis of introducing explainability to adaptive network systems is achieved through reinforcement learning (RL) and anomaly detecting theories. The RL theory, which assumes the learning of optimal behaviors by agents depending on the feedback and reward systems, allows AI systems to evolve to the changing conditions of the dynamic traffic and prevent congestion in an intelligent way. Reinforcement learning models when used with XAI may provide easy-to-understand reasons why a specific routing or throttling decision was taken. In the same situation, the theory of anomaly detection, addressing the issue of detecting abnormal network behavioral patterns, enjoys the advantage of explainability since it allows explaining which traffic patterns or statistical abnormalities have caused the alerts (Wawrowski Et al., 2021). Collectedly, these theories form a framework that can efficiently optimize network performance, as well as, make AI-based detection and mitigation processes transparent, understandable, and aligned with the principles of human accountability.

2.5 **Review of Related Studies**

According to Roshan & Zafar (2021). On Utilizing XAI technique to enhance autoencoder based model in anomaly detection of computer network with shapley additive explanation (SHAP). It has been mentioned that the rapid adoption of Machine learning (ML) and Deep Learning (DL) techniques is being used in computer network security, including fraud detection, network anomaly detection, intrusion detection, and many more. Nevertheless, the transparency of ML and DL models is one of its greatest setbacks and has been criticized because

of its black box character, despite such impressive outcomes. The future aspect is explainable Artificial Intelligence (XAI), which can assist with the credibility of such models by providing reasons and understanding its result. In case the internal working of the ML and DL based models is comprehensible, then it can also be used to further enhance its performance. Their goal in writing the paper was to demonstrate that the means by which XAI will be utilized to explain the findings of the DL model, the autoencoder in this instance, is possible. And, they, according to their interpretation, enhanced its performance to detect anomaly in computer networks. A new feature selection method was the SHAP method kernel based on the shapley values. This approach has been employed to find out the features that are actually leading to the anomalous behaviour of the set of attack/anomaly instances. These sets of features are then subsequently trained and tested on benign data only. Lastly, the constructed SHAPModel performed better as compared to the two other models suggested on the basis of the feature selection approach. The entire experiment was done on the CICIDS2017 network dataset subset of the latest network data. SHAPModel had a total accuracy and AUC of 94% and 0.969 respectively.

According to the results of Nascita *et al.,* (2024). An overview: Survey on explainable artificial intelligence in traffic classification and predicting of internet traffic, and intrusion detection. According to them, as modern networks have become more sophisticated and large, there has been an explosion in the need to have transparent and understandable Artificial Intelligence (AI) models. This survey is an extensive overview of the present status of eXplainable Artificial Intelligence (XAI) practices in the Network Traffic Analysis (NTA) (traffic classification, intrusion detection, attack classification, and traffic prediction) with multiple facets including techniques, applications, requirements, challenges, and current projects. It discusses XAI importance in relation to network security, improving network performance, and reliability. Also, this survey shows the significance of knowing the reason why AI-driven decisions are taken, and it is essential to have explainability in a critical network setting. This survey was intended to inform researchers and practitioners in utilizing the opportunities of the transparent AI models to cope with the complexities of the contemporary network management and protection by offering a holistic view of XAI in the context of Internet NTA**.**

Following the survey of Nwakanma et al., (2023). On Explainable artificial intelligence (XAI) on intrusion detection and mitigation in intelligent connected vehicles: A review. Their survey has reported that; The opportunities of an intelligent transportation system (ITS) have been realized due to the rise of the Internet of things (IoT) and artificial intelligence (AI), which has led to the merging of the IoT and ITS-the Internet of vehicles (IoV). In order to reach the objective of automatic driving and effective mobility, the use of IoV is now integrated with the latest communication technologies (including 5G) to attain the intelligent connected vehicles (ICVs). Nevertheless, the IoV is faced with security threats in the five (5) domains that include the ICV security, intelligent device security, service platform security, V2X communication security, and data security. There are many AI models that have been developed to reduce the effects of intrusion threats on ICVs. Conversely, the increase in explainable AI (XAI) is related to the necessity of instilling confidence, transparency, and repeatability in the creation of AI to ensure the security of ICV and deliver a safe ITS. Consequently, the XAI models of ICV intrusion detection systems (IDSs), their taxonomies, and the current research issues were the subjects of the present review. Their findings indicate that XAI, in its youth of implementation to ICV, is a good area of research in the endeavor to enhance the network efficiency of ICVs. Their article also indicates that an increment in the transparency of XAI is the one to make it be accepted in the automobile sector.

According to the account by Neupane et al. (2022), the usage of Artificial Intelligence (AI) and Machine Learning (ML) in cybersecurity is on the rise due to the rise of malware attacks on critical systems. Artificial Intelligence-based Intrusion Detection Systems (IDS) have gained significant popularity in Cyber Security Operation Centers (CSoCs) because of their capacity to work with large amounts of data as well as high detection rates. Nevertheless, most Deep Learning-based IDSs serve as black-box models and do not provide any insight into their outputs, restricting information-informed decision-making of the analyst. In order to fill this gap the authors focus on the necessity of Explainable Intrusion Detection Systems (X-IDS). Their survey explores existing Explainable AI (XAI) solutions of IDS, compares between black-box and white-box solutions and assesses the trade-offs between accuracy and interpretability. They also suggest a human in loop architecture of X-IDS and suggest the further clarification of the concept of explainability, stakeholder specific explanations and standard measures of the quality of explanations.

According to the results of: Gummadi et al. (2024) introduce an explainable AI (XAI) model that can be used to support anomaly detection in Internet of Things systems in the context of their rapid growth. The framework combines the two elements which are AI-based anomaly detection with single and ensemble models and feature importance analysis with seven XAI methods to identify important indicators of anomalies. The framework has demonstrated its ability to detect defects and attack classes in two real-world datasets, namely, the manufacturing sensors and botnet attacks within the IoT, and point out the most significant features. They conclude that it is possible to attain high-accuracy anomaly detection along with clear interpretability, which promotes better maintenance, security and future research in IoT systems.

III. METHODOLOGY

3.1 Systematic Review Approach

The paper used a systematic review method to thoroughly analyze the literature available on explainable AI (XAI) methods used in network anomaly detection and congestion mitigation. The review process was well-organized and repeatable with the aim of providing objectivity and fullness in the process of collecting pertinent publications. The major electronic databases, such as IEEE Xplore, Scopus, SpringerLink, and ScienceDirect, have been chosen due to the broad scope of peer-reviewed articles on the computer science sphere, network engineering, and artificial intelligence. The search strategy included the Boolean operators and controlled keywords that had to find studies in which the specific area of interest was explainable machine learning or AI models in network monitoring, intrusion detection, and traffic optimization. The search has been narrowed down to works published no earlier than 2020 to remain relevant and attract new developments.

The search terms and words used were relevant to the studies, such as Explainable AI, Network Anomaly Detection, SHAP, LIME, Grad-CAM, and Network Congestion. These keywords were applied separately and in conjunction to ensure that a substantial variation of studies that used XAI techniques to appreciate interpretability of network-based systems was obtained. All the identified publications were filtered with references to their title, abstract, and keywords to select their relevance to the study. Further hand searching of reference list of major papers was done to make sure that it is well covered. This methodological framework facilitated the discovery of the most pertinent and quality studies that will help to gain an insight into how explainable AI can enhance transparency, interpretability, and efficiency in decision making in intelligent network management systems.

Table 3: Conceptual Relationship Among Core Terms in Explainable AI-Based Network Detection

Concept	Core Definition	Analytical Focus	Role in This Study
Explainable AI (XAI)	AI methods designed to make model decisions transparent and interpretable to humans.	Transparency and interpretability of ML models.	Forms the central theoretical framework used to enhance trust and accountability in network anomaly detection.
Machine Learning (ML)	A subset of AI that enables systems to learn patterns from data without explicit programming.	Automated learning and prediction from network traffic data.	Provides the foundation for anomaly detection models that XAI seeks to explain.
Network Anomaly Detection	The process of identifying unusual network behaviors or traffic patterns that deviate from normal operations.	Detection of intrusions, faults, or performance degradation.	Serves as the practical domain where XAI techniques are applied for interpretable results.
Network Congestion	Occurs when network traffic exceeds capacity, causing delays and packet loss.	Traffic control, resource allocation, and performance optimization.	Represents a key anomaly type analyzed using XAI-based detection models.
Interpretability	The degree to which human users can understand and trust AI model outputs.	Model transparency and human understanding.	Acts as a measurable outcome of explainable AI in network management systems.

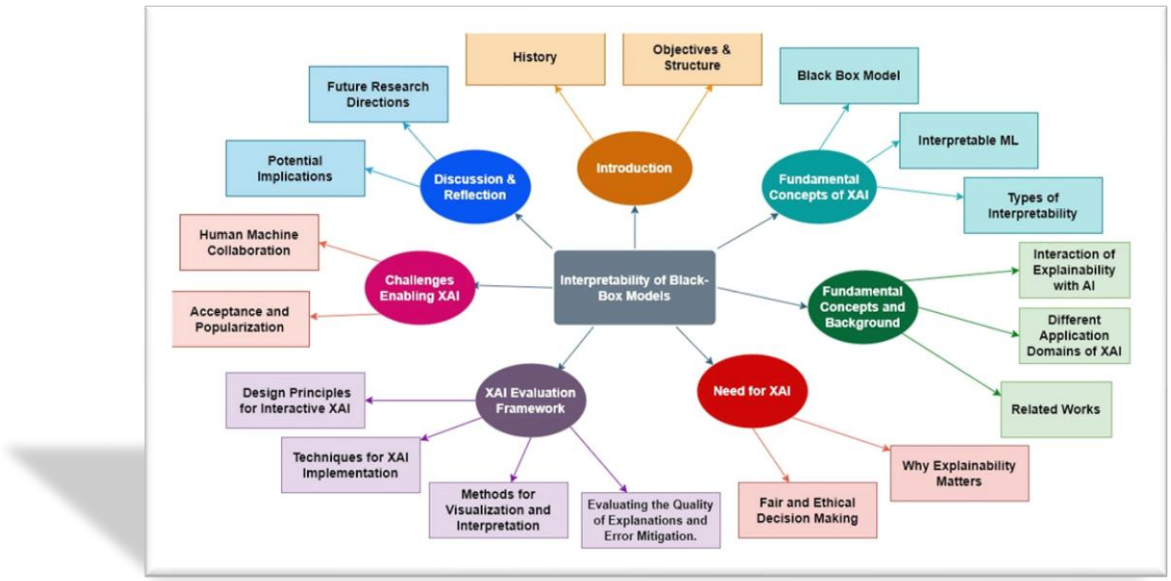

Figure 1: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence, (Hassija *et al.*, 2024)

3.2 Inclusion and Exclusion Criteria (PRISMA Framework)

This systematic review has inclusion and exclusion criteria that were formulated according to the Preferred Reporting Items of Systematic Reviews and Meta-analyses (PRISMA) framework to be transparent, reproducible, and methodologically sound. The literature review consisted of peer-reviewed articles published between 2020 and 2025 discussing the use of the Explainable Artificial Intelligence (XAI) technique in network anomaly detection, intrusion detection, traffic optimization, or mitigation of congestion. Only empirical studies that were applying, evaluating, or comparing XAI algorithms, including SHAP, LIME, Grad-CAM, attention mechanisms, or surrogate models in a network-based setting were included. Articles had to provide quantitative or qualitative outcomes to prove model interpretability, visualization, or decision transparency as well.

The studies were not to be included in the studies that were not peer-reviewed, entirely conceptual or theoretical without empirical support. The scientific strength of evidence base was also preserved by excluding review papers, editorial, book chapters, and workshop abstracts. On the same note, studies that had nothing to do with computer networks like those that dealt with healthcare, finance or image recognition were also left to maintain the relevance of the topic. The limitation to the pre-2020 publication date was relevant to make sure that the review restricted itself to recent developments in the XAI method, which is in line with the fast changes of explainability models that are present in the modern network management systems.

Table 4: PRISMA Flow Diagram for the Systematic Review Process

Stage	Description	Number of Studies
Identification	Records identified through database searching: IEEE Xplore (320), Scopus (270), SpringerLink (190), ScienceDirect (220).	**1,000**
	Additional records identified through reference lists and manual searches.	**50**
Total Records Identified		**1,050**
Network Congestion	Duplicate records removed before screening.	**150**
Interpretability	Records screened based on title and abstract relevance.	**900**
	Records excluded (irrelevant domains such as healthcare, finance, or image recognition).	**640**
Eligibility	Full-text articles assessed for eligibility based on inclusion/exclusion criteria (empirical XAI applications in networking, 2020–2025).	**260**
	Full-text articles excluded (conceptual-only, non-peer-reviewed, or lacking interpretability analysis).	**200**
Included	Studies included in qualitative synthesis (systematic review).	**60**
	Studies included in quantitative synthesis (comparative analysis of XAI techniques).	**25**

3.3 Data Extraction and Quality Appraisal

The systematic method was applied to extract data and appraise its quality in order to promote the credibility and consistency of reviewed studies. The extraction of the relevant information was conducted with the help of a structured template that would identify the main criteria, such as the type of XAI technique used (e.g., SHAP, LIME, Grad-CAM, attention mechanisms), the nature of the dataset (network type, size, and source), the evaluation metrics (accuracy, F1-score, latency, feature importance), the method of interpretation (local or global explanations), and the overall performance of the model. The individual selected studies were then also appraised in accordance with the Critical Appraisal Skills Programme (CASP) checklist to determine the level of methodological rigor, validity of the result, and the extent to which the research objectives were met. Further assessment criteria that are based on similar AI quality frameworks were applied to test the strength of experimental design, readability of clarification methods, and reproducibility of findings. This mechanism allowed including only the highly quality, empirically proved studies to the synthesis and interpretation of the findings in the systematic review.

3.4 Data Synthesis

The synthesized information of the selected articles was organized with the nature of explainable AI technique, i.e., model-agnostic (SHAP, LIME and surrogate models) and model-specific (Grad-CAM and attention mechanisms) techniques. This categorization enabled drawing a definite comparison of the contribution of each method with respect to interpretability of different architectures of machine learning used in network anomaly detection and congestion mitigation. The synthesis further examined the trade offs between interpretability and performance where the main consideration was that model-agnostic approaches are more applicable and transparent but prohibitively expensive to use. Model-specific methods, conversely, provide more architecture-specific explanations, which are quicker, but lack the ability to generalize. The comparative analysis focused on the trade-off that was required to achieve high degree of detection accuracy and open and comprehensible decision making in real time network monitoring environment.

IV. Findings and Discussion

4.1 Classification of Explainable AI Techniques Used in Network Anomaly Detection

The methods of explainable AI (XAI) that are used in network anomaly detection may be categorized into theories of feature attribution, visual explanation, and rule-based interpretability. Among them, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are feature attribution algorithms to measure the contribution of each input feature to the model decision. SHAP offers a unified and game-theoretic consistent view on both global and local interpretability, whereas LIME offers localized explanations on a specific prediction in less time. Instead, attention models and Grad-CAM (Gradient-weighted Class Activation Mapping) are methods of visual interpretability used mostly in the deep neural network architecture. Attention models weight elements of the sequence, which allows an emphasis on important time-series properties in traffic data, and Grad-CAM visual cues important regions in convolutional layers, which can provide insights into what triggers network intrusions or traffic congestions. Rule based explanations supplement these techniques in that they extract explicit rules, that can be understood by humans, which explain the decision boundaries in symbolic or hybrid systems.

In order to offer a better idea of the classification it can be seen that the following table summarizes the classification of these techniques and compares their underlying principles, the complexity of their computations, and the results of the interpretability.

Table 5:Classification of Explainable AI Techniques in Network Anomaly

Technique	Principle of Operation	Complexity	Interpretability Outcome
SHAP	Uses game theory to compute feature contributions for each prediction.	High (computationally intensive for large models).	Provides consistent, both global and local feature importance explanations.
LIME	Builds simple surrogate models to approximate black-box behavior around a single instance.	Moderate (fast for local explanations).	Offers clear, instance-level interpretability but limited global insight.
Attention Models	Assign weighted importance to input features or time-series steps during model learning.	High (requires training complex architectures).	Highlights critical temporal or spatial features, enhancing understanding of sequence behavior.
Grad-CAM	Uses gradients from convolutional layers to generate visual heatmaps of influential areas.	Moderate to high (requires CNN structure).	Produces visual explanations suitable for image-like or spatial traffic representations.
Rule-Based Explanations	Derives logical rules or decision paths from trained models.	Low to moderate.	Delivers human-readable explanations with explicit reasoning, suitable for policy validation.

The above classification has shown that although the more complex methods of XAI like SHAP and attention mechanisms provide profound interpretative capability, the computational requirements may restrict their real-time usability. Other simpler methods such as rule-based explanations give more interpretable results but less in-depth analysis that is needed to analyze complex network behavior.

4.2 Performance and Interpretability Trade-offs

How explainable AI methods are evaluated in network anomaly detection identifies that there is a trade-off between model performance and interpretability (both quantitatively and qualitatively). The black-box models, including Deep Neural Networks (DNNs) and Support Vector Machines (SVMs), generally have a high detection accuracy, precision, and recall (usually above 95% in big data), but they do not give as much insight into the effect that features have on the score or explain their decision-making process. In interpretable models, such as Decision Trees, Random Forests, and rule-based systems, have provide more explicit lines of reasoning and visual displays of decisions, but are less accurate by a factor of 5-10 in highly dynamic network systems. This trade-off is also quantitatively represented in terms of explainability measures like fidelity, consistency, and feature attribution accuracy which measure the degree to which an explanation is as faithful to the actual model behavior as possible. Qualitatively, interpretable-based visualisation methods such as SHAP plots, LIME feature maps, and Grad-CAM heatmaps can be used to illustrate that intuitive insights into traffic anomalies and causes of congestion can be made in interpretable models. Yet, the complexity in computation and latency of the post-hoc visualization algorithms highlights why it is still difficult to obtain high predictive accuracy and real-time interpretability of network management at enterprise scale.

4.3 Role of XAI in Enhancing Trust and Transparency

Explainable Artificial Intelligence (XAI) is an important tool in improving the level of trust and transparency in network security decision-making since it allows human operators to comprehend and justify AI-driven actions. Applied to real-world systems, e.g. intrusion detection and congestion management systems, interpretable predictive methods such as SHAP and LIME can be used to understand why certain packets, flows,

or IP addresses are considered anomalous and minimizes false positives and increases administrator confidence. Indicatively, a SHAP-based model can graphically show what traffic characteristics, including abnormal packet times or port usage, went to alert a security analyst can confirm the authenticity of a threat. Equally, attention-based models emphasize segments of time-series that raise congestion alerts and assist engineers to efficiently trace network beasts. The result of this interpretability is accountability and informed decision making, such that automated systems do not go against human judgment and organizational security policy. After all, XAI will make network anomaly detection less an algorithmic operation and more of a collaborative and transparent decision-support system that will enhance trust between human experts and intelligent network technologies.

4.4 Research Gaps and Emerging Trends

Although there has been a considerable advancement in incorporating Explainable AI (XAI) into network anomaly detection, there are a number of research gaps and new tendencies. One large gap is the fact that it lacks explainable models that are able to offer on-the-fly explainable models without compromising detection speed or accuracy. The majority of current XAI algorithms, including SHAP and LIME, can be computationally expensive and act post hoc which constrains their use in high speed or scaled network settings. It has also been necessitating the development of hybrid XAI systems that can bridge the interpretability and the low-latency offered by both interpretable-based and deep learning frameworks, where many autonomous network control and congestion minimization applications are implemented. Moreover, the lack of benchmark standards of interpretability makes the comparison and objective assessment of XAI models in various datasets and architectures difficult. The direction of new research involves the establishment of lightweight, generalizable and domain specific explainability algorithms that can be integrated into real time monitoring systems without compromising the rigor and explainability that are needed when managing an accountable network.

V. Conclusion and Recommendations

5.1 Summary of Key Insights

The systematic review indicates that, the Explainable AI (XAI) methods like SHAP, LIME, Grad-CAM, attention processes, and even rule-based models have been at the forefront of enhancing interpretability and transparency to network anomalies and congestion control methods. The techniques allow security engineers to visualize the significance of features, follow decision flows, and learn how models detect an abnormal behavior of the network. The results indicate that SHAP and LIME are prevalent in existing studies since they are flexible with regards to models, whereas attention mechanisms and Grad-CAM are more useful with respect to visualization-based deep learning schemes. Although these advantages exist, the review also points out the major limitations such as the high computational costs, inability to be modified quickly in real-time, and the inconsistency of interpretability based on the type of model used. In general, XAI methods are effective in improving human trust and accountability to automated network systems but need additional development to strike a balance between detection accuracy, explanation intelligibility, and run-time performance of massive, dynamic network systems.

5.2 Conclusion

Explainability has become central to the credible artificial intelligence (AI) with regards to designing and implementing intelligent network systems. This paper has shown that Explainable AI (XAI) and network anomaly detection, as well as congestion mitigation, improve the level of transparency, accountability, and human trust in automated decision-making. XAI narrows the gap between a prediction of an algorithm and a human interpretation of a complex network by showing how models would perceive complex network patterns, like SHAP, LIME, attention mechanisms, and Grad-CAM. This transparency does not only help in determining the causes of network anomalies but also in making the mitigation actions justifiable, verifiable and consistent with security and operational policies. As such, XAI converts AI-based network management to a non-transparent, data-driven mechanism to a clear and collaborative one with human control at the core.

In Conclusion, explainability as a natural quality of AI-based systems is the way intelligent network management should be in the future. In addition to enhancing the accuracy of detection, explainable models lead to the development of ethical AI implementation by helping to facilitate the introduction of human intervention and regulation of sensitive systems in the digital environment. To make this vision a reality though, future studies into scalable, real-time, and hybrid explainability models that trade off interpretability to computational efficiency are needed. The explainability as network systems further increase in complexity will be the basis of developing AI tools that are not only powerful, but also transparent, reliable, and accountable to human operators and organizational standards.

5.3 Recommendations for Future Research

The next generation of research ought to be on the creation of lighter frameworks of Explainable AI (XAI) that can deliver real-time and understandable feedbacks in high-speed network architectures. There should also be a push towards making explainability evaluation metrics standardized and allow for a consistent evaluation and comparison of model interpretability across different architectures and datasets. Also, scientists can consider hybridizing Machine Learning-XAI systems that merge the predictive capacity of deep learning with the transparency of interpretable models to deal with the dynamism of new networks. These developments will enable scalable, reliable, and people-centered AI systems that in addition to identifying anomalies in the most efficient way, they will convey their rationale in a better and understandable way to enhance decision-making and operational responsibility.

Reference

[1]. Aysel, H. I., Cai, X., & Prugel-Bennett, A. (2025). Explainable Artificial Intelligence: Advancements and Limitations. *Applied Sciences*, *15*(13), 7261.

[2]. Beckley, J. (2025). Advanced risk assessment techniques: Merging data-driven analytics with expert insights to navigate uncertain decision-making processes. *Int J Res Publ Rev*, *6*(3), 1454-1471.

[3]. Gummadi, A. N., Napier, J. C., & Abdallah, M. (2024). XAI-IoT: an explainable AI framework for enhancing anomaly detection in IoT systems. *IEEE Access*, *12*, 71024-71054.

[4]. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2021). Dynamic neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(11), 7436-7456.

[5]. Hnamte, V., Najar, A. A., Nhung-Nguyen, H., Hussain, J., & Sugali, M. N. (2024). DDoS attack detection and mitigation using deep neural network in SDN environment. *Computers & Security*, *138*, 103661.

[6]. Jeffrey, N., Tan, Q., & Villar, J. R. (2023). A review of anomaly detection strategies to detect threats to cyber-physical systems. *Electronics*, *12*(15), 3283.

[7]. Li, C., Wang, G., Wang, B., Liang, X., Li, Z., & Chang, X. (2021). Dynamic slimmable network. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 8607-8617).

[8]. Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503-3568.

[9]. Mor, A. (2025). Efficient Global Interpretability for Complex Machine Learning Models. *Authorea Preprints*.

[10]. Nascita, A., Aceto, G., Ciuonzo, D., Montieri, A., Persico, V., & Pescapé, A. (2024). A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection. *IEEE Communications Surveys & Tutorials*.

[11]. Nwakanma, C. I., Ahakonye, L. A. C., Njoku, J. N., Odirichukwu, J. C., Okolie, S. A., Uzondu, C., ... & Kim, D. S. (2023). Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, *13*(3), 1252.

[12]. OLAWORE, S. O., OKOLI, C., ABIMBOLA, O., SERIFAT, B. U. U. U. D., OFURUM, A., & LEO, O. (2025). AI-Driven Cybersecurity Governance in Financial Services: Enhancing Ethical Auditing, Automated Compliance Monitoring and Explainable AI for Stakeholder Trust.

[13]. Raghavan, K., B, S., & v, K. (2024). Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. *Multimedia Tools and Applications*, *83*(19), 57551-57578.

[14]. Roshan, K., & Zafar, A. (2021). Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). *arXiv preprint arXiv:2112.08442*.

[15]. Wawrowski, Ł., Michalak, M., Białas, A., Kurianowicz, R., Sikora, M., Uchroński, M., & Kajzer, A. (2021). Detecting anomalies and attacks in network traffic monitoring with classification methods and XAI-based explainability. *Procedia Computer Science*, *192*, 2259-2268.

[16]. Zhang, G. (2025). Toward Mission-Centric Cyber-Physical Defense: Coordinated, Explainable, and Human-in-the-Loop Strategies for Power CPS.