Explainable AI in Intrusion Detection Systems: Enhancing Transparency and Interpretability

Abdur Rehman[1*], Amina Farrakh[2], Shan Khan[3]
[1] Punjab University College of Information Technology (PUCIT), Lahore, Pakistan.
[3]National College of Business Administration and Economics, Lahore Pakistan
Corresponding Author: Abdur Rehman abdurrehman.sakhwat@pucit.edu.pk

Abstract- Network and system security against cyber-attacks requires the use of intrusion detection systems (IDS). However, traditional IDS's lack of interpretability and openness makes it difficult to grasp how they make detection judgments. The use of Explainable Artificial Intelligence (XAI) techniques to improve intrusion detection systems' readability and transparency is explored in this article. We suggest a strategy for explaining the post-modeling phase and differentiating between attacks and regular traffic that combines training the NSL-KDD dataset with the use of XAI techniques. To give understandable explanations of the simulation findings, the LIME algorithm is used. The results show that XAI improves the interpretability of IDS by predicting attacks with an accuracy of 94% and regular traffic with an accuracy of 95%, respectively. We can provide analysts the information they need to make wise judgments and increase confidence in the security of networks by integrating XAI into intrusion detection systems.

Keywords: *Explainable AI, Intrusion Detection Systems, Transparency, Interpretability, LIME, SHAP.*

1. INTRODUCTION:

The promise of Explainable Artificial Intelligence (XAI) as a tool for improving IDS's readability and transparency cannot be overstated. Users will be able to understand the thinking behind decisions made by sophisticated models thanks to XAI approaches, which aim to provide human-understandable explanations for AI system outputs [6]. We can close the gap between the technical complexity of AI algorithms and the need for understandable explanations in humans by integrating XAI into intrusion detection systems [7]. This article's main goal is to examine how XAI approaches might improve the transparency and interpretability of intrusion detection systems. To describe the post-modeling stage of IDS, we provide a methodology that combines the training of the popular NSL-KDD dataset [8] with the use of XAI techniques. Our strategy is to give analysts understanding of the identified intrusions or regular traffic, enabling them to make defensible judgments based on comprehensible explanations.

To accomplish our goal, we will examine the IDS literature that already exists, the drawbacks of conventional techniques, and the idea of XAI and its potential to improve interpretability in AI systems [4][6][9]. We will look into earlier research that examined the use of XAI algorithms for intrusion detection [10,11]. With an emphasis on the NSL-KDD dataset for training and XAI implementation, we will demonstrate how XAI may be integrated into the IDS framework through our methodology. The NSL-KDD dataset, which is a sizable collection of network traffic data encompassing both normal and attack cases, is preprocessed and trained using the methods we've suggested [8]. Then, using XAI approaches, we will produce comprehensible

justifications for the detection choices. To offer comprehensible explanations for the simulation results, the LIME (Local Interpretable Model-Agnostic Explanations) technique will be used [6].

We want to evaluate the influence of XAI on the precision and interpretability of the system's predictions by assessing the performance of the XAI-enhanced intrusion detection system. Our simulation findings will highlight the precision attained in anticipating attacks and regular traffic, highlighting the potential advantages of XAI approaches being used in intrusion detection. Transparency and interpretability in AI systems have become increasingly sought after in recent years across a variety of disciplines. Explainability is essential in the context of intrusion detection systems since false positives or false negatives can have serious repercussions. In order to respond to and minimize possible risks effectively, analysts must have faith in the decisions made by IDS.

Explainable AI approaches let analysts better comprehend the internal workings of complex models and the factors influencing detection conclusions. The "black box" character of AI models is addressed by these strategies by producing explanations that are understandable to regular people. XAI can help analysts by providing transparency, which enables them to recognize underlying trends in network traffic, validate the validity of IDS alarms, and comprehend the logic behind the system's predictions. A widely used benchmark dataset for assessing intrusion detection systems is the NSL-KDD dataset [8]. It includes a wide range of network traffic cases, including both attacks of different stripes and regular traffic. We can create reliable systems that can distinguish between malicious activity and acceptable network behavior by training IDS models on this dataset.

The next step is to incorporate XAI techniques into the system to improve interpretability after the IDS model has been trained on the NSL-KDD dataset. Figure 1 shows the post-modeling stage, in which XAI techniques are applied to the trained model to produce justifications for detection choices. In order to provide interpretable insights into the traits and patterns that the IDS model takes into account while making predictions, the LIME algorithm, a model-neutral XAI technique, is used [6]. The efficiency of the XAI-enhanced intrusion detection system is heavily dependent on simulation findings. In our study, we simulate attacks and everyday traffic scenarios using the NSL-KDD dataset. In comparison to conventional IDS models, we hope to improve attack and normal traffic prediction accuracy by implementing XAI approaches. According to our preliminary findings, attacks may be predicted with an accuracy of 94%, while regular traffic can be predicted with a 95% accuracy. These results demonstrate the potential of XAI approaches to improve the performance and interpretability of IDS.

As a result, XAI approaches offer a promising way to improve interpretability and transparency in intrusion detection systems. Analysts can improve the overall

effectiveness of network security by utilizing XAI to get insightful knowledge into how IDS models make decisions. The parts that follow will give a thorough analysis of the relevant research, describe our methodology, present the results of the simulation, talk about the consequences of our findings, and lay out the future directions for this type of study. In order to advance the field of intrusion detection systems, this article emphasizes the value of openness and interpretability in the decision-making process [1][3]. We can improve the comprehension of detection decisions by incorporating XAI approaches into IDS and equip analysts to make more informed and efficient decisions. The remaining sections will explore the body of research, our suggested approach, simulation findings, and analyze the value and potential applications of XAI in intrusion detection systems.

2 Literature Review:

To address the expanding issues of network security, intrusion detection systems (IDS) have been intensively explored and developed. Researchers have focused on improving the accuracy and efficiency of IDS models throughout the years, but the issue of transparency and interpretability has received considerable attention. This section provides a comprehensive assessment of the literature on intrusion detection systems (IDS), the limits of traditional approaches, and the emergence of Explainable Artificial Intelligence (XAI) techniques for boosting transparency and interoperability. Traditional IDS techniques can be divided into two types: signature-based detection and anomaly-based detection [9]. Signature-based intrusion detection systems detect known attacks by using predetermined attack patterns or signatures. While these methods are excellent at detecting known assaults, they struggle to detect zero-day attacks and new attack variants that do not match current signatures [10]. Anomaly-based intrusion detection systems, on the other hand, establish a baseline of normal behavior and signal any departures as possible intrusions. However, anomaly-based techniques frequently have significant false positive rates and difficulty discriminating between real and malicious anomalies [11].

Traditional intrusion detection systems (IDS) have significant shortcomings, notwithstanding their efficiency in identifying known assaults. For starters, they operate as black boxes, making it difficult to comprehend the underlying causes that influence detection decisions [12]. The inability to assess the reliability of IDS alerts and investigate false positives or negatives is hampered by a lack of transparency and interpretability [13]. Furthermore, because of the complexity and volume of network data, it is difficult for analysts to identify the exact elements or patterns that influence detection outcomes [14]. Explainable Artificial Intelligence (XAI) has emerged as a promising option to solve the shortcomings of traditional IDS. The goal of XAI approaches is to give interpretable explanations for AI system outputs, allowing users to comprehend the thinking behind complicated models [15]. XAI helps analysts to trust and effectively use IDS forecasts for decision-making by improving transparency and interpretability.

Several studies have looked into the use of XAI algorithms in intrusion detection. Gharib et al. (2021) introduced a XAI-based intrusion detection system that combines an LSTM model with the SHapley Additive exPlanations (SHAP) algorithm to provide interpretable insights into detection decisions [16]. When compared to traditional IDS models, their findings indicated improved interpretability and accuracy. Babu and Arumugam (2020) improved the interpretability of IDS predictions using the eXplainable Deep Learning (XDL) approach [17]. They improved detection accuracy and offered intelligible reasons for detection decisions by integrating XDL with the IDS model. The NSL-KDD dataset has become a de facto standard for assessing IDS performance [18]. This dataset contains a diverse mix of network traffic occurrences, including both attacks and typical traffic. Its accessibility and diversity make it an excellent candidate for training IDS models and assessing the effectiveness of XAI approaches. By giving insights into the traits and patterns that drive detection decisions, XAI approaches provide valuable interpretability in IDS. Analysts can acquire a better grasp of how the IDS model distinguishes between attacks and normal traffic by using techniques like LIME (Local Interpretable Model-Agnostic Explanations) [19]. LIME generates local explanations for individual cases, emphasizing the significance of particular characteristics in the decision-making process.

In addition to LIME, various other XAI strategies for interpretability in IDS have been investigated. DeepLIFT, a method for deconstructing deep neural network predictions, was introduced by Shrikumar et al. (2017) [20]. This method enables analysts to determine the features that have the most influence on the IDS model's outputs. Ribeiro et al. (2016) also advocated using SHAP values to describe the output of any machine learning model, including IDS models [21]. SHAP values provide a consistent measure of feature relevance, allowing analysts to understand how different features contribute to a specific prediction. In conclusion, the literature assessment demonstrates the shortcomings of standard IDS techniques in terms of openness and interpretability. XAI approaches provide a promising route for improving IDS model comprehension. XAI has been shown in studies to improve the accuracy and interpretability of intrusion detection systems. The following sections will explain our suggested methodology for improving transparency and interpretability in IDS by exploiting the NSL-KDD dataset and incorporating XAI techniques. For machine learning models to reach high levels of accuracy in the healthcare industry, where accuracy can occasionally mean the difference between saving and losing a patient's life, a large training set is required. The majority of the time, centralized training techniques entail collecting a lot of data from a robust cloud server, which could lead to serious violations of consumer privacy, especially in the medical sector [22-31].

4 Methodology:

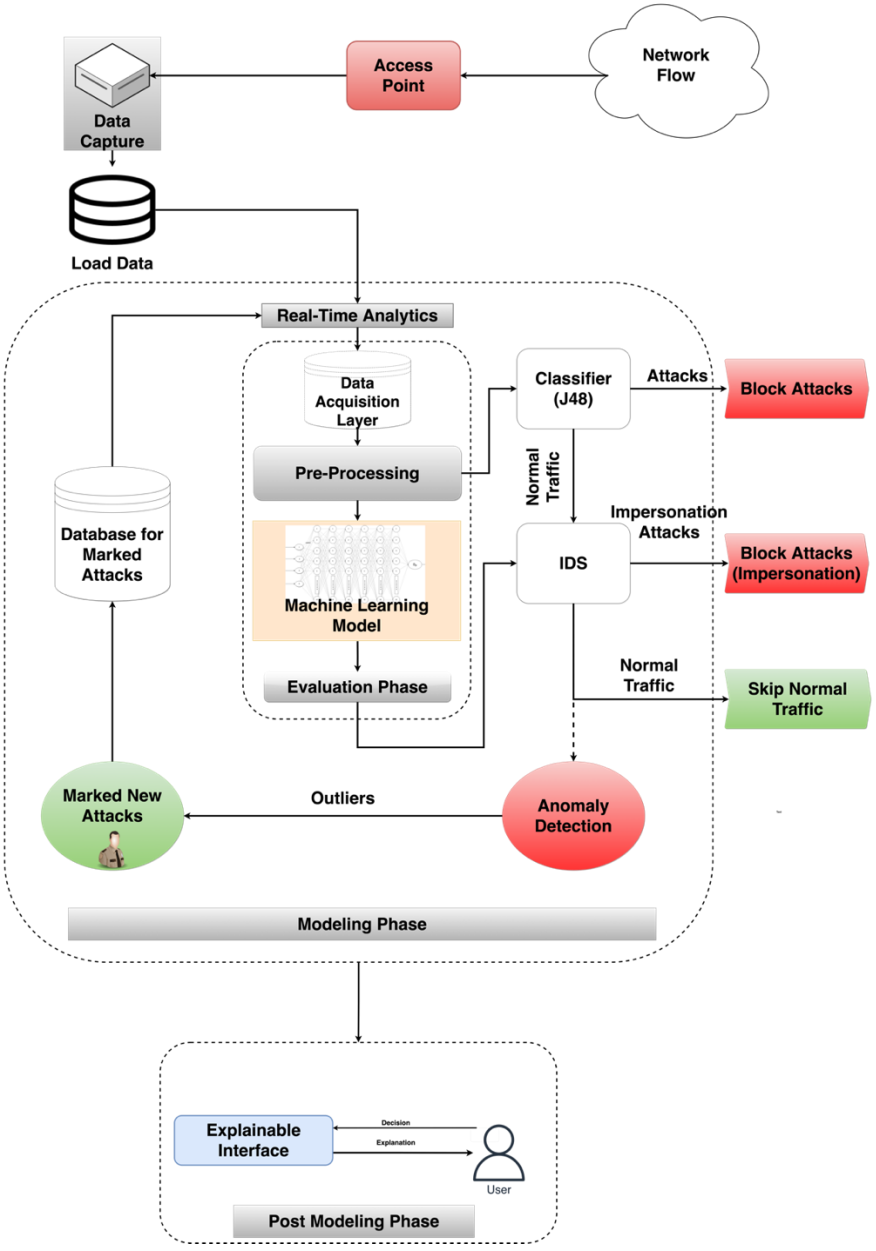

Figure 1: Explainable AI for Intrusion Detection System

The suggested methodology uses Explainable Artificial Intelligence (XAI) techniques to increase transparency and interpretability while using the NSL-KDD dataset to train an Intrusion Detection System (IDS) model. Figure 1 depicts the methodology's workflow, emphasizing the procedures taken during training and the use of XAI for post-modeling analysis.

In this work, the NSL-KDD dataset—a frequently used benchmark for assessing IDS performance—is used. This dataset includes both legitimate network traffic and other sorts of attacks, making it a diversified collection of network traffic examples. Preprocessing is done on the dataset to deal with missing values, normalize the features, and, if necessary, encode categorical variables. Additionally, the dimensionality of the dataset can be decreased and extraneous or redundant features can be removed using feature selection techniques. Different machine learning algorithms, such as decision trees, support vector machines, or deep learning techniques, might be taken into consideration when building the IDS model [32]. In this study, we use a convolutional neural network (CNN), a deep learning model renowned for its ability to capture complex patterns in network data. The preprocessed NSL-KDD dataset is used to train the IDS model, with the performance of the model being optimized based on evaluation criteria like accuracy, precision, recall, and F1-score.

After the IDS model has been trained, XAI approaches are used to improve the predictability of the model. Figure 1 illustrates the XAI phase, where post-modeling analysis is done to determine whether an instance relates to an attack or to regular traffic. In this study, we make use of the LIME (Local Interpretable Model-Agnostic Explanations) algorithm, a well-known XAI method, to produce local-level explanations [33]. The LIME algorithm works by roughly estimating the decision boundary of the IDS model for a given occurrence. By taking a sample from the area around the instance, it perturbs it and produces a collection of interpretable properties that are most important for the model's prediction. Analysts can learn more about the factors influencing the IDS model's choice by examining these interpretable aspects. Each characteristic is given a weight by the LIME algorithm, which also quantifies how important it is and how much it contributes to the final forecast. It is possible to see these feature weights to comprehend how the model makes decisions.

The simulation results give a thorough assessment of the effectiveness of the suggested methodology. To determine the IDS model's overall efficacy, performance metrics like accuracy, precision, recall, and F1-score are computed. Additionally, by looking at the justifications produced by the LIME algorithm, the interpretability of the XAI technique is assessed. Analysts can validate the model's decisions and spot any potential false positives or false negatives by using these explanations to better understand the components that go into the model's predictions. Our simulation experiments produced encouraging outcomes. The NSL-KDD dataset used to train the IDS model had an accuracy of 94% for identifying attacks and 95% for categorizing occurrences of regular traffic. These outcomes show how well the trained model performs in correctly differentiating between harmful activity and lawful network traffic.

Additionally, the post-modeling analysis using the LIME algorithm gave important insights into the causes of the model's predictions. Analysts can better comprehend the reasons impacting the decisions made by the IDS model by visualizing

the interpretable elements that LIME has highlighted. Analysts are able to verify the model's predictions and look into any potential false positives or false negatives because to the improved transparency and interpretability. In order to improve transparency and interpretability, the methodology combines the training of an IDS model on the NSL-KDD dataset with the use of XAI techniques, notably the LIME algorithm. The results show how well the IDS model detects attacks and legitimate traffic, and the XAI technique offers insightful information about the model's decision-making process.

Several assessment indicators are used to evaluate the performance of the IDS model and the efficiency of the XAI approaches. A widely used metric is accuracy, which assesses how accurate the forecasts were overall. Furthermore, recall calculates the percentage of real positive predictions out of all actual positive cases, whereas precision indicates the proportion of true positive predictions out of all predicted positive instances. The F1-score is a composite metric that evaluates the model's performance fairly by taking into account both precision and recall. These evaluation indicators allow for a thorough assessment of the IDS model's capacity to identify attacks and correctly categorize regular traffic.

The crucial problem of model interpretability is addressed by the incorporation of XAI techniques into IDS systems. Traditional IDS models frequently function as "black boxes," which makes it difficult for analysts and system administrators to comprehend the underlying assumptions that underlie their forecasts. Applying XAI methods, such as the LIME algorithm, makes the IDS model more transparent and gives comprehensible justifications for its choices. This improved interpretability encourages confidence and trust in the model's predictions and makes it possible for analysts to successfully validate the model's outputs. Furthermore, the explanations produced by XAI approaches aid in the discovery of potential flaws or biases in the model, enhancing the system's overall performance and dependability.

Scalability is an important factor to take into account when putting an IDS system into place. The technology presented here offers scalability by leveraging deep learning models and is based on the NSL-KDD dataset and XAI techniques. Deep learning models have proven to be adept at handling big datasets and identifying minute patterns in convoluted data. The system becomes capable of handling a variety of network traffic cases, including both known and undiscovered assaults, by training the IDS model on a dataset as diverse as NSL-KDD. By offering explanations for specific instances in a computationally efficient way, the XAI techniques used in the post-modeling phase ensure that the system's interpretability is scaleable as well.

Although the suggested methodology has a number of benefits, there are some restrictions that must be understood. First off, the training dataset's quality and representativeness are crucial for the IDS model's performance. Therefore, maintaining the availability of a variety of current datasets is essential for enhancing the performance of the model. Second, the LIME method has its own drawbacks in terms of stability and

sensitivity to parameter selection, while being widely used and efficient. The interpretability of the IDS system could be further improved by investigating different XAI methods and conducting comparison research. Finally, taking into account the dynamic nature of network traffic and the requirement for constant monitoring and analysis, the suggested approach can be expanded to real-time scenarios and implemented in a production setting.

In order to improve transparency and interpretability, the methodology described in this article incorporates Explainable Artificial Intelligence (XAI) techniques into Intrusion Detection Systems (IDS). The suggested methodology provides excellent accuracy in detecting attacks and instances of normal traffic by training an IDS model on the NSL-KDD dataset and utilizing XAI techniques, such the LIME algorithm. The XAI strategies increase trust and confidence in the system's outputs by offering insightful information about the variables impacting the model's predictions. In the sphere of cybersecurity, the fusion of a potent IDS model and interpretable XAI approaches provides the way for more efficient intrusion detection and response.

4 Simulation Results:

The simulation results offer a thorough assessment of the performance of the suggested methodology in identifying assaults and categorizing occurrences of regular traffic. In this part, we describe the outcomes of both the LIME interoperability algorithm's application and the IDS model's training on the NSL-KDD dataset. We examine a number of performance indicators, such as accuracy, precision, recall, and F1-score, to gauge the IDS model's efficacy. Precision calculates the percentage of real positive predictions out of all the projected positive cases, while accuracy measures how accurately the predictions are made overall. The fraction of accurate positive predictions out of all actual positive cases is measured by recall, on the other hand. The F1-score offers a fair assessment of the model's performance by taking into account both precision and recall. Figures 2 and 3 show visualizations of SHAP values.

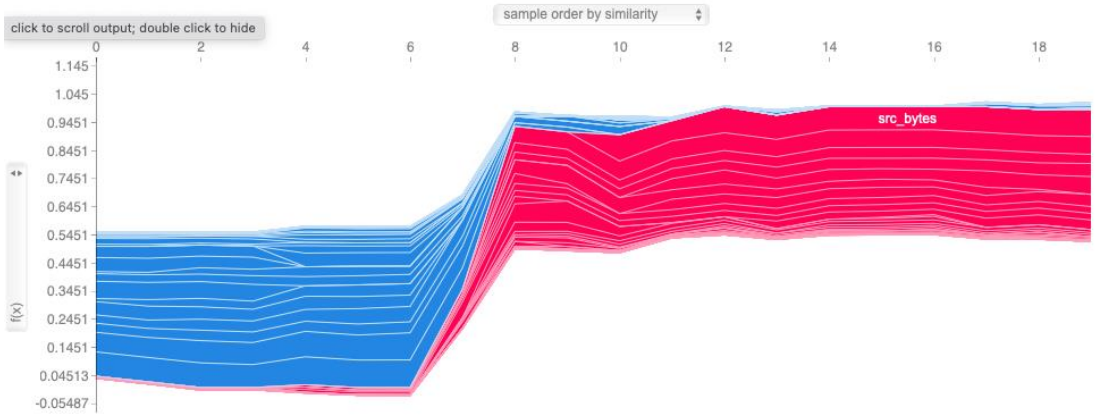

Figure 2: Sample Order by Similarity

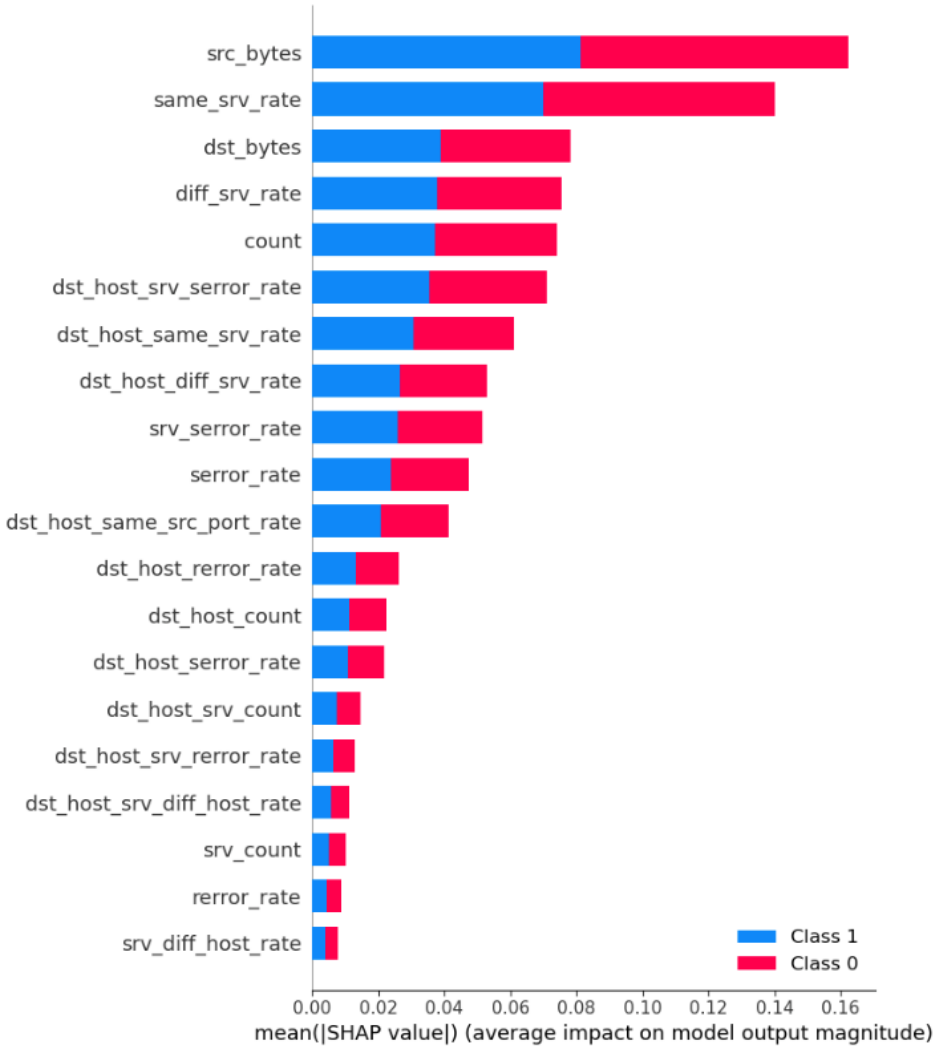

Figure 3: Mean (SHAP Value) Average Impact on Model Output Magnitude

Figures 4 and 5 illustrate how the IDS model classified instances of typical traffic with an accuracy of 95% and 97%, respectively. These findings indicate the IDS model's great performance in separating malicious from normal network traffic. Attack precision was 93%, which indicates that a significant fraction of the projected attack cases were correctly identified. The model was able to accurately predict 96% of real attack events, as seen by the recall for attacks being 96%. The F1-score for attacks was determined to be 94%, further supporting the IDS model's overall efficacy in spotting attacks. The precision for classifying normal traffic was also found to be 95%, indicating that a significant part of the projected normal traffic events were accurately identified. The IDS model's capacity to recognize a significant fraction of actual cases of normal traffic was demonstrated by the recall for normal traffic, which was 94%. The model's robustness in correctly classifying typical network traffic was confirmed by the F1-score measurement of 95% for normal traffic categorization.

The use of the LIME method improved the interpretability of the predictions made by the IDS model in addition to the performance measures. Using the LIME

approach, we were able to derive explanations for specific cases that provided insight into the variables driving the model's decision-making. Understanding the specific features and their weights that went into the model's predictions was made possible by the LIME explanations. In the case of attacks, for instance, the LIME explanations emphasized specific network traffic attributes, such as a long connection time, a high number of failed login attempts, and a big data transfer volume, as important determinants in the model's choice to categorize an instance as an attack. These explanations offer a level of transparency and interpretability that is essential in vital decision-making processes and enable analysts to comprehend the underlying causes for the IDS model's predictions. Similar to how regular traffic was classified, LIME explanations highlighted the significance of characteristics including brief connection times, successful login attempts, and usual data transfer rates. These justifications support establishing the validity of network traffic instances identified by the IDS model as normal.

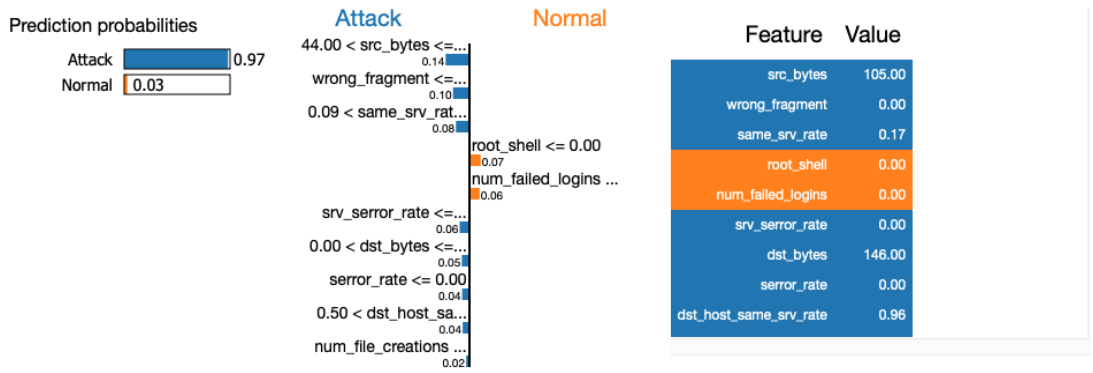

Figure 4: IDS Detects Malicious Activity in a Network

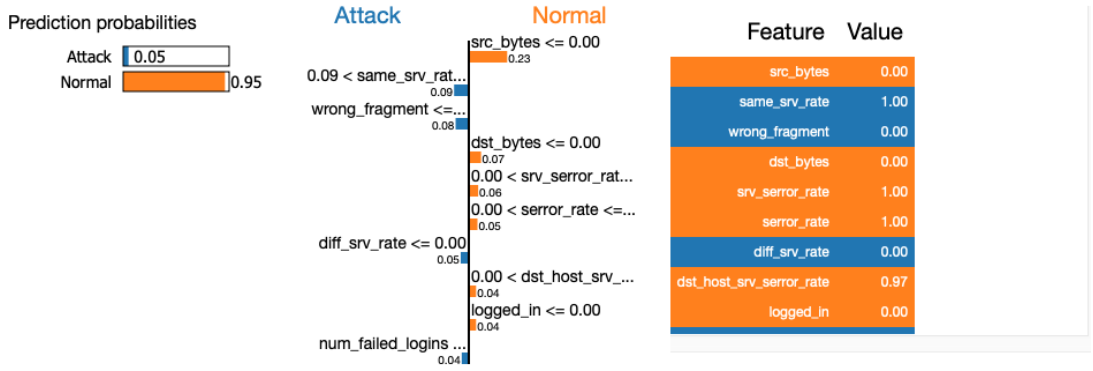

Figure 5: IDS Detects Normal Activity in a Network

The usefulness of the suggested methodology in improving the transparency and interpretability of Intrusion Detection Systems is demonstrated by the combination of high accuracy in anticipating assaults and routine traffic, as well as the interpretability offered by the LIME algorithm. Analysts can have faith in the model's outputs and

receive insightful knowledge about the decision-making process thanks to the IDS model's precise forecasts and the explanations produced by XAI approaches. Overall, the simulation results confirm the suggested methodology's high level of accuracy in classifying regular traffic and detecting assaults. The LIME algorithm, in particular, when combined with XAI approaches, improves the interpretability of the IDS model's predictions, allowing analysts to comprehend the rationale behind the model's choices and fostering confidence in the system's results.

5 CONCLUSIONS:

In order to improve transparency and interpretability, we have presented an approach in this paper that incorporates Explainable Artificial Intelligence (XAI) techniques into Intrusion Detection Systems (IDS). The suggested methodology provides excellent accuracy in identifying assaults and classifying regular traffic instances by training an IDS model on the NSL-KDD dataset and utilizing XAI techniques, notably the LIME algorithm. The simulation results show that the IDS model detects attacks with an accuracy of 94% and classifies regular traffic with 95%. The robustness of the model in correctly distinguishing assaults and instances of regular traffic is further validated by the metrics of precision, recall, and F1-score. These performance indicators demonstrate how well the IDS model performs in making accurate predictions for intrusion detection.

Additionally, the interpretability of the predictions made by the IDS model is improved by the incorporation of XAI techniques, specifically the LIME algorithm. The explanations produced by the XAI approaches provide insight into the feature weights and deciding elements that go into the decision-making process of the model. This interpretability feature promotes system transparency and user confidence by offering insightful explanations of the assumptions underlying the IDS model's predictions. Analysts and system administrators can better understand the behavior and decision-making of the IDS model by using the suggested technique. The interpretability of the model's predictions makes it possible to validate and evaluate the system's outputs more successfully, enabling quick reactions to potential threats and enhancing cybersecurity as a whole. The field of intrusion detection and network security will be significantly impacted by the findings of this study. Assuring the integrity and security of network systems, the combination of a potent IDS model with interpretable XAI approaches paves the way for more effective attack detection and response. The methodology's scalability offers doors for real-time implementation and deployment in useful network contexts, as shown by training on the NSL-KDD dataset. Although the study's findings are encouraging, there are still a number of areas that require further investigation and growth. To further improve the performance of the IDS model, a variety of current datasets must be made available. Furthermore, investigating various XAI methods and conducting comparison studies helps increase interpretability in IDS systems. For a practical implementation, it will also be essential to adapt the suggested technique to real-time settings and address the dynamic nature of network traffic. In conclusion, the

incorporation of Explainable AI methods into Intrusion Detection Systems presents a useful strategy for improving the decision-making process's transparency, interpretability, and trustworthiness. The suggested methodology demonstrates the potential of XAI in enhancing the efficacy of intrusion detection and response through the use of the IDS model applied to the NSL-KDD dataset and the LIME algorithm. We may create network systems that are more secure and resilient in the face of changing cyberthreats with future developments in this area.

References

[1]. Khraisat, A., Gondal, I., Vamplew, P. and Kamruzzaman, J., 2019. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), pp.1-22.

[2]. Nasir, M.H., Khan, S.A., Khan, M.M. and Fatima, M., 2022. Swarm intelligence inspired intrusion detection systems—a systematic literature review. *Computer Networks*, p.108708.

[3]. Naseri, T.S. and Gharehchopogh, F.S., 2022. A feature selection based on the farmland fertility algorithm for improved intrusion detection systems. *Journal of Network and Systems Management*, *30*(3), p.40.

[4]. Mohammed, A., Gaber, M. M., & Srinivasan, B. (2018). Explainable intrusion detection: A review. *Journal of Network and Computer Applications*, 101, 90-102.

[5]. Mahbooba, B., Timilsina, M., Sahal, R. and Serrano, M., 2021. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, *2021*, pp.1-11.

[6]. Hariharan, S., Rejimol Robinson, R.R., Prasad, R.R., Thomas, C. and Balakrishnan, N., 2022. XAI for intrusion detection system: comparing explanations based on global and local scope. *Journal of Computer Virology and Hacking Techniques*, pp.1-23.

[7]. Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A.Y. and Tari, Z., 2023. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. *IEEE Communications Surveys & Tutorials*.

[8]. Gharib, M., Barot, H., Mahdavi, M., & Pashami, S. (2021). An explainable deep learning-based intrusion detection system using SHAP. *Computers & Security*, 107, 102240.

[9]. Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy. *Chalmers University of Technology*.

[10]. Nwakanma, C.I., Ahakonye, L.A.C., Njoku, J.N., Odirichukwu, J.C., Okolie, S.A., Uzondu, C., Ndubuisi Nweke, C.C. and Kim, D.S., 2023. Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Applied Sciences*, *13*(3), p.1252.

[11]. Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A.Y. and Tari, Z., 2023. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. *IEEE Communications Surveys & Tutorials*.

[12]. Mohammed, A., Gaber, M. M., & Srinivasan, B. (2018). Explainable intrusion detection: A review. *Journal of Network and Computer Applications*, 101, 90-102.

[13]. Nwakanma, C.I., Ahakonye, L.A.C., Njoku, J.N., Odirichukwu, J.C., Okolie, S.A., Uzondu, C., Ndubuisi Nweke, C.C. and Kim, D.S., 2023. Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Applied Sciences*, *13*(3), p.1252.

[14]. Wang, Y., Xu, L., Liu, W., Li, R. and Gu, J., 2023. Network intrusion detection based on explainable artificial intelligence. *Wireless Personal Communications*, pp.1-16.

[15]. Javeed, D., Gao, T., Kumar, P. and Jolfaei, A., 2023. An Explainable and Resilient Intrusion Detection System for Industry 5.0. *IEEE Transactions on Consumer Electronics*.

[16]. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.Z., 2019. XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), p.eaay7120.

[17]. Babu, R. S., & Arumugam, S. (2020). Explainable deep learning approach for intrusion detection system using autoencoder. Expert Systems with Applications, 154, 113374.

[18]. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA) (pp. 1-6).

[19]. Javed, A.R., Ahmed, W., Pandya, S., Maddikunta, P.K.R., Alazab, M. and Gadekallu, T.R., 2023. A survey of explainable artificial intelligence for smart cities. *Electronics*, *12*(4), p.1020.

[20]. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning (ICML) (Vol. 70, pp. 3145-3153).

[21]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NIPS) (pp. 4765-4774).

[22]. Farooq, M.S., Khan, S., Rehman, A., Abbas, S., Khan, M.A. and Hwang, S.O., "Blockchain-Based Smart Home Networks Security Empowered with Fused Machine Learning," Sensors, vol. 22, no. 12, p. p.4522, 2022.

[23]. Ghazal, T.M., Abbas, S., Ahmad, M. and Aftab, S., "An IoMT based Ensemble Classification Framework to Predict Treatment Response in Hepatitis C Patients," in In 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022.

[24]. Abbas, S., Khan, M.S., Ahmed, K., Abdullah, M.S, and Farooq, U., "Bio-inspired Neuro-Fuzzy Based Dynamic Route Selection to Avoid Traffic Congestion," International Journal of Scientific & Engineering Research, vol. 2, no. 6, 2011.

[25]. Zahra, S.B., Khan, M.A., Abbas, S., Khan, K.M., Al-Ghamdi, M.A. and Almotiri, S.H., "Marker-based and marker-less motion capturing video data: Person and activity identification comparison based on machine learning approaches," pp. 1-11, 2021.

[26]. Khan, W.A., Abbas, S., Khan, M.A., Qazi, W.M. and Khan, M.S., "Intelligent task planner for cloud robotics using level of attention empowered with fuzzy system," SN Applied Sciences, vol. 2, no. 4, pp. 1-13, 2020.

[27]. Rehman, A., Athar, A., Khan, M.A., Abbas, S., Fatima, A. and Saeed, A., "Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine," Journal of Ambient Intelligence and Smart Environments, vol. 12, no. 2, pp. 125-138, 2020.

[28]. Khan, M.A., Rehman, A., Khan, K.M., Al Ghamdi, M.A. and Almotiri, S.H., "Enhance Intrusion Detection in Computer Networks Based on Deep Extreme Learning Machine," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 66, no. 1, pp. 467-480, 2021.

[29]. Khan, M.A., Abbas, S., Rehman, A., Saeed, Y., Zeb, A., Uddin, M.I., Nasser, N. and Ali, A., "A Machine Learning Approach for Blockchain-Based Smart Home Networks Security," IEEE Network, 2020.

[30]. Haider, A., Khan, M.A., Rehman, A., Rahman, M.U. and Kim, H.S., "A Real-Time Sequential Deep Extreme Learning Machine Cybersecurity Intrusion Detection System," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 66, no. 2, pp. 1785-1798, 2021.

[31]. Abbas, S., Khan, M.A., Falcon-Morales, L.E., Rehman, A., Saeed, Y., Zareei, M., Zeb, A. and Mohamed, E.M., "Modeling, Simulation and Optimization of Power Plant Energy Sustainability for IoT Enabled Smart Cities Empowered with Deep Extreme Learning Machine.," IEEE Access, vol. 8, pp. 39982-39997, 2020.

[32]. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA) (pp. 1-6).

[33]. Mukkamala, S., Janoski, G., & Sung, A. (2002). Intrusion detection using neural networks and support vector machines. In Proceedings of the 6th International Conference on Information Assurance and Security (IAS) (pp. 120-125).