An XAI-Driven Intrusion Detection Framework for False Positive Reduction

K. Baby Ramya[1], K. Pavani[2], G. Naga Durga Devi[3]

#1 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada.

Abstract: Cybersecurity risks and network assaults have greatly expanded due to the quick development of Internet technologies, cloud computing, and IoT devices. When it comes to spotting malicious activity in network traffic, intrusion detection systems (IDS) are crucial. However, a large number of false positives are frequently produced by conventional anomaly-based IDS models, which lowers system dependability and raises needless security warnings. This study suggests a methodology for false positive identification in intrusion detection systems based on Explainable Artificial Intelligence (XAI) to address this problem.

To increase the transparency and precision of intrusion detection, the suggested approach integrates machine learning algorithms with XAI methods including SHAP and feature relevance analysis. In order to successfully differentiate between real assaults and false alarms, the framework evaluates the significance of network traffic characteristics in addition to prediction confidence ratings. For effective intrusion detection and classification, a variety of machine learning and ensemble models are used, such as Random Forest, KNN, Voting Classifier, and Stacking Classifier with LightGBM.

The LYCOS-IDS2017 dataset is used to assess the system, and the suggested method significantly reduces false positives while maintaining genuine positive detections. The inclusion of XAI improves interpretability, increases detection reliability, and facilitates improved cybersecurity decision-making, according to experimental data. For contemporary network intrusion detection situations, the suggested architecture offers a clear, precise, and clever solution.

Index terms *- — Intrusion Detection System, Explainable AI, False Positive Reduction, Machine Learning, Cybersecurity, Network Traffic Analysis, SHAP, LIME, Ensemble Learning, LightGBM, Random Forest, Anomaly Detection, Feature Selection, Deep Learning, Intelligent Security Systems.*

1. INTRODUCTION

The fast proliferation of Internet technologies, cloud computing, wireless communication, and IoT devices has boosted network traffic and security risks in modern digital settings. Network security is a critical issue as companies and individuals use internet services for communication, financial transactions, and data storage. Secure data and systems are more important than ever since cyber criminals create advanced techniques to exploit network flaws. IDSs are commonly used to monitor network activity and identify malicious conduct in real time to address these difficulties.

Traditional intrusion detection uses signature- and anomaly-based methods. Signature-based IDS detects known attacks but few new ones. However, anomaly-based IDS use ML and DL algorithms to detect anomalous network activity and unknown attack patterns. Although these intelligent algorithms increase attack detection, they often create many false positives, misclassifying typical network events as harmful. Large false positive rates impair system dependability, create unwanted security warnings, and complicate network administration.

Explainable Artificial Intelligence (XAI) improves machine learning model transparency and interpretability to overcome these constraints. XAI methods aid decision-making by assessing individual aspects and explaining model predictions. In this study, a XAI-based intrusion detection system is presented to decrease anomaly-based IDS false positives. The system uses machine learning classifiers and SHAP and LIME feature relevance analysis to separate real assaults from false alarms.

The framework improves intrusion detection using ensemble learning approaches as Random Forest, K-Nearest Neighbors (KNN), Voting Classifier, and Stacking Classifier with LightGBM. LYCOS-IDS2017 is used for training and evaluation. This approach improves detection accuracy, decreases false alarms, increases transparency, and improves cybersecurity decision-making by merging confidence ratings with explainable feature relevance analysis. This method helps create reliable, intelligent, and explainable intrusion detection systems for current network security.

2. LITERATURE SURVEY

a) Enhanced detection of imbalanced malicious network traffic with regularized Generative Adversarial Networks:

To defend against the increasing risks to network security, many businesses need to safeguard their networks and detect malicious network traffic. The imbalance across attack classes, which reduces the learning efficacy of machine learning models used to identify fraudulent traffic, is one of the main problems. In order to enhance minority attack samples for a balanced dataset, this study introduces regularized Wasserstein Generative Adversarial Networks (WGAN). Five statistical indicators are used to evaluate the efficacy of data augmentation. The proposed WGAN-IDR (Wasserstein GAN with Improved Deep Analytic Regularization) performs better than existing methods for augmenting data. The performance of each class on the CICIDS2017 dataset is examined using binary and multiclass classification tests using three classification techniques. These techniques are TRTS, TSTR, and TRTR. The TSTR and TRTS classification methods outperformed baseline and earlier attempts on the balanced CICIDS2017 dataset using WGAN-IDR

because of their realistic and varied samples. Binary classification receives an F1-score of 0.99, whereas multiclass classification receives 0.98.

b) A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM:

Systems for detecting network intrusions are crucial to network security. Minority attacks are difficult to identify due of the unreliability of current network intrusion statistics. Moreover, training and detecting deep neural network detection techniques takes a lot of time. This paper proposes a network intrusion detection system that uses adaptive synthetic (ADASYN) oversampling and LightGBM. This method is based on the aforementioned problems. To remove the chance that the maximum or minimum value may change general features, we normalize and one-hot encode the original data during the first stage of data preparation. Second, in order to enhance minority attack detection because of training data imbalance, we use ADASYN to oversample minority samples. Lastly, the LightGBM ensemble learning model maintains detection accuracy while lowering temporal complexity. ADASYN oversampling may increase minority sample detection and improve accuracy, according to experimental verification on the NSL-KDD, UNSW-NB15, and CICIDS2017 data sets. In three test sets, the accuracy of the suggested method is 92.57%, 89.56%, and 99.91%. Training and detection are faster than previous methods.

c) IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks:

System security is made more challenging by the emergence of constantly evolving network threats, particularly in dynamic and decentralized ad hoc networks. Cybersecurity relies heavily on intrusion detection, which finds unusual traffic patterns. However, a difficult scenario where aberrant samples are significantly fewer than normal samples has been created by class-imbalanced data. The effectiveness of intrusion classifiers and the system's resistance to unidentified abnormalities are both diminished by this class imbalance issue. A novel Imbalanced Generative Adversarial Network (IGAN) for class imbalance is presented in this study. The basic GAN is made more minority class-representative by adding neural layers and an uneven data filter. This is the main originality of our concept. An IGAN-based intrusion detection system is developed to address class-imbalanced intrusion detection. IGAN instances are used in this system. The three modules of IGAN-IDS are feature extraction, IGAN, and deep neural networks. First, raw network characteristics are transformed into feature vectors via a feed-forward neural network (FNN). Latent space samples will then be generated using IGAN. The deep neural network, which has completely connected and convolutionally generated layers, detects intrusions. We compare IGAN-IDS to fifteen other methods and assess it on three benchmark datasets. Our IGAN-IDS outperforms existing methods, according to the testing.

d) An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic:

Intrusion detection systems (IDS) deal with the increasing integration of the Internet with human interactions due to cyber threats. We were let down by the conventional machine learning-based intrusion detection system. We provide an intrusion detection model based on CNN in this paper. Prior to CNN

training, network traffic is balanced using the Synthetic Minority Oversampling Technique (SMOTE-ENN) and Edited Nearest Neighbors (ENN). NSL-KDD is used to evaluate the model. The accuracy of the CNN intrusion detection system model using SMOTE-ENN is 83.31%. Attack detection rates for Remote to Local (R2L) and User to Root (U2R) have also increased. The findings indicate that the CNN IDS based on SMOTE-ENN performs better than the previous model.

e) Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset:

As IoT technology has expanded, malicious third parties have attacked it. Network intrusion detection and network forensic solutions are needed to safeguard and look into this problem. A representative and organized dataset is necessary for both system training and credibility assessment. There are several network datasets accessible, but only a small number include details on botnet scenarios. The objective of this project is the Bot-IoT dataset. Attacks and real and simulated IoT network traffic are included in this collection. In order to address the dataset's inadequacies, including its incapacity to accurately label, capture complete network information, and account for current and complex attacks, we also suggest a realistic testbed. Lastly, we assess the forensics reliability of the BoT-IoT dataset by comparing it to benchmark datasets. Several statistical and machine learning techniques are employed in this analysis. The detection of botnets in IoT networks is made possible by this study. For the dataset, go to Bot-IoT (2018).

3. METHODOLOGY

i) Proposed Work:

The proposed work focuses on developing an intelligent and explainable intrusion detection framework capable of identifying and reducing false positive alerts in anomaly-based Intrusion Detection Systems (IDS). Traditional IDS models often generate a high number of false alarms due to limitations in understanding the importance of network traffic features during prediction. To overcome this issue, the proposed system integrates Machine Learning (ML) algorithms with Explainable Artificial Intelligence (XAI) techniques to improve detection accuracy and transparency. The system collects and preprocesses network traffic data from the LYCOS-IDS2017 dataset, followed by feature extraction and classification using advanced machine learning models such as Random Forest, K-Nearest Neighbors (KNN), Voting Classifier, and Stacking Classifier with LightGBM.

In addition to intrusion detection, the proposed framework applies XAI methods such as SHAP and feature relevance analysis to explain the contribution of each network feature in the prediction process. The confidence scores generated by machine learning models are combined with feature importance values to accurately distinguish genuine attacks from false positive detections. This approach significantly reduces unnecessary security alerts while preserving true attack detections. The proposed system enhances interpretability, reliability, and decision-making capability in cybersecurity environments, making it more suitable for real-world network security applications where transparency and accuracy are equally important.

ii) System Architecture:

The proposed system architecture is designed to identify and reduce false positives in anomaly-based Intrusion Detection Systems (IDS) using Explainable Artificial Intelligence (XAI). The architecture consists of three major stages: Threshold Setup, Attribute Extraction, and Machine Learning-Based False Positive Detection. Initially, the anomaly-based IDS processes the network traffic dataset and classifies suspicious activities based on prediction confidence levels. High-confidence detections are considered True Positives (TP), while low-confidence detections may contain both genuine attacks and false positives. A threshold mechanism is therefore established to separate highly reliable predictions from uncertain detections for further analysis.

In the second stage, Explainable AI techniques such as SHAP and adversarial feature analysis are applied to extract important attributes that influence intrusion predictions. These XAI-generated feature relevance scores help the system understand why a particular network activity is classified as malicious. The extracted XAI attributes are then forwarded to the Machine Learning False Positive Detector in the final stage. Here, advanced machine learning models analyze both prediction confidence and feature relevance information to accurately identify false positive alerts while preserving true attack detections. This architecture improves transparency, interpretability, detection accuracy, and overall reliability of the intrusion detection system in real-world cybersecurity environments.

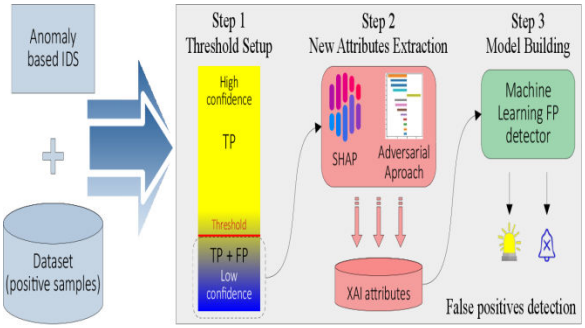

Fig1 proposed architecture

iii) Modules:

1. Data Collection Module

This module is responsible for collecting and loading the network traffic dataset used for intrusion detection. The proposed system utilizes the LYCOS-IDS2017 dataset, which contains both normal and malicious network traffic records. The collected data serves as the foundation for training and evaluating the intrusion detection models.

2. Data Pre-processing Module

The preprocessing module cleans and transforms the raw dataset into a suitable format for machine learning analysis. This process includes handling missing values, normalization, encoding categorical attributes, and removing irrelevant or duplicate data. Proper preprocessing improves model efficiency and detection accuracy.

3. Feature Extraction and Selection Module

This module extracts important network traffic features that contribute to intrusion detection. Feature selection techniques are applied to identify the most relevant attributes while reducing unnecessary data complexity. The optimized feature set helps improve

classification performance and reduces computational overhead.

4. Machine Learning Detection Module

The machine learning module performs intrusion detection using classification algorithms such as Random Forest, KNN, Voting Classifier, and Stacking Classifier with LightGBM. These models are trained to identify malicious and normal network activities based on extracted traffic features.

5. Explainable AI (XAI) Module

The XAI module uses techniques such as SHAP and feature relevance analysis to explain the predictions generated by machine learning models. It identifies the contribution of each feature toward the final prediction, thereby improving transparency and helping users understand the decision-making process of the IDS.

6. Confidence and Relevance Analysis Module

This module combines machine learning confidence scores with XAI feature importance values to analyze suspicious detections. The integrated analysis helps distinguish genuine cyberattacks from false positive alerts more effectively and improves overall system reliability.

7. False Positive Identification Module

The false positive detection module identifies and filters unnecessary alerts generated during anomaly detection. By analyzing confidence thresholds and feature relevance scores, the system minimizes false alarms while preserving legitimate intrusion detections.

8. Model Evaluation Module

This module evaluates system performance using metrics such as accuracy, precision, recall, F1-score, false positive rate, and detection rate. The performance of the proposed XAI-based framework is compared with traditional IDS approaches to validate its effectiveness in reducing false positives.

iv) Algorithms:

1. Random Forest Algorithm

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting. In the proposed intrusion detection system, Random Forest is used to classify network traffic as normal or malicious based on extracted features. The algorithm improves detection reliability by aggregating predictions from multiple trees, thereby reducing false positives and enhancing overall intrusion detection performance.

2. K-Nearest Neighbors (KNN) Algorithm

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that classifies data samples based on the similarity of neighboring data points. In this system, KNN analyzes network traffic patterns and identifies intrusions by comparing new records with previously known attack and normal traffic samples. The algorithm helps detect suspicious activities by measuring the distance between feature vectors and assigning the most common neighboring class.

3. Decision Tree Algorithm

Decision Tree is a classification algorithm that creates a tree-like structure to make decisions based

on feature conditions. The algorithm divides the dataset into branches according to important traffic attributes and predicts whether the activity is malicious or normal. In the proposed framework, Decision Tree assists in identifying attack patterns and supports explainable decision-making for intrusion detection.

4. Voting Classifier Algorithm

The Voting Classifier is an ensemble learning technique that combines predictions from multiple machine learning models to improve classification accuracy. In the proposed system, Random Forest and AdaBoost classifiers are integrated using majority voting. The final prediction is determined based on the combined outputs of the individual classifiers, which enhances robustness, reduces false alarms, and improves attack detection capability.

5. Stacking Classifier with LightGBM

The Stacking Classifier is an advanced ensemble learning approach that combines multiple base models with a meta-classifier to achieve better prediction performance. In this framework, Random Forest and Multi-Layer Perceptron (MLP) are used as base classifiers, while LightGBM acts as the meta-classifier. The stacking method learns from the strengths of individual models and produces highly accurate intrusion detection results with reduced false positive rates.

6. SHAP (SHapley Additive Explanations)

SHAP is an Explainable Artificial Intelligence (XAI) technique used to interpret machine learning predictions by measuring the contribution of each feature. In the proposed system, SHAP analyzes the importance of network traffic attributes involved in intrusion detection. This helps identify why a particular prediction is classified as an attack or false positive, thereby improving transparency and interpretability.

7. LIME (Local Interpretable Model-Agnostic Explanations)

LIME is an XAI-based explanation technique that provides local interpretations for individual predictions generated by machine learning models. In this intrusion detection framework, LIME helps explain suspicious detections by highlighting influential features responsible for classification decisions. This improves user trust and assists cybersecurity analysts in understanding model behavior effectively.

4. EXPERIMENTAL RESULTS

The proposed Explainable AI (XAI)-based Intrusion Detection System was implemented using Python, Flask, Machine Learning libraries, and the LYCOS-IDS2017 dataset. The experimental evaluation focused on detecting malicious network traffic while reducing false positive alerts through feature relevance analysis and ensemble learning techniques. The system integrates multiple machine learning algorithms such as Random Forest, KNN, Voting Classifier, and Stacking Classifier with LightGBM to improve intrusion detection accuracy. The web-based interface was developed to allow user authentication, data input, prediction analysis, and result visualization. Experimental testing demonstrated that the proposed system effectively distinguishes malicious traffic from normal network activities while maintaining high prediction reliability.

The generated results indicate that the proposed framework successfully minimizes false positive detections by combining confidence scores with Explainable AI feature importance analysis using SHAP and LIME techniques. The output visualization displays attack probability and normal probability values, enabling users to understand prediction confidence clearly. During testing, the system accurately classified normal network traffic with high confidence while preserving true intrusion detections. The experimental outcomes confirm that integrating XAI with ensemble machine learning models improves transparency, interpretability, cybersecurity decision-making, and overall intrusion detection performance compared to traditional anomaly-based IDS approaches.

Accuracy: A test's accuracy is its capacity to distinguish healthy from ill cases. Find the percentage of instances with genuine positives and negatives to assess test accuracy.

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Classification accuracy or positive cases constitute precision. The formula for accuracy is:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\ cision = \frac{TP}{(TP + FP)}$$

Recall: A model's recall measures its ability to recognize all appropriate machine learning class instances. The ratio of accurately predicted positive observations to total positives indicates a model's class instance detection skill.

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

mAP: Mean Average Precision ranks quality. It considers the number and order of relevant ideas. Calculating MAP at K uses the arithmetic mean of each user or query's Average Precision (AP).

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{the AP of class } k}$$

$$\boldsymbol{n = \text{the number of classes}}$$

F1-Score: A high F1 score suggests an accurate machine learning model. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\ cision)}{(Recall + \Pr e\ cision)}$$

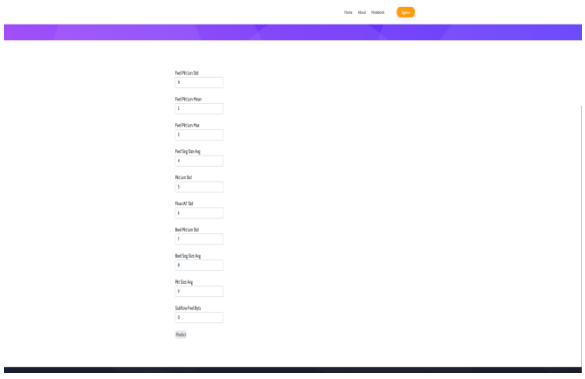

Fig2 Network Traffic Input Interface

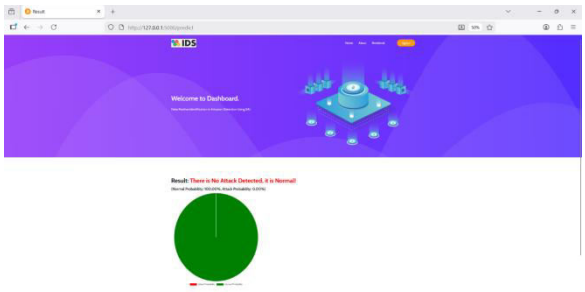

Fig3 Intrusion Detection Result Output

5. CONCLUSION

The proposed Explainable Artificial Intelligence (XAI)-based Intrusion Detection System successfully improves the identification of false positives in anomaly-based network intrusion detection. By integrating machine learning algorithms with XAI techniques such as SHAP and feature relevance analysis, the system provides accurate, transparent, and interpretable intrusion detection results. The framework effectively analyzes network traffic patterns and distinguishes genuine cyberattacks from false alarms using confidence scores and feature importance values.

The experimental results demonstrate that the proposed system significantly reduces false positive rates while maintaining high true positive detection accuracy. Ensemble learning methods such as Random Forest, Voting Classifier, and Stacking Classifier with LightGBM further enhance classification performance and system reliability. In addition to improving cybersecurity decision-making, the integration of XAI increases user trust by clearly explaining the reasoning behind intrusion predictions. Therefore, the proposed framework offers an intelligent, reliable, and scalable solution for modern network security environments and contributes toward the development of transparent AI-driven intrusion detection systems.

6. FUTURE SCOPE

The proposed XAI-based Intrusion Detection System can be further enhanced by integrating advanced Deep Learning techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformer-based models to improve attack detection accuracy for large-scale and real-time network environments. Future work can also focus on handling zero-day attacks and evolving cyber threats by implementing adaptive and self-learning security mechanisms. The use of real-time streaming data and cloud-based intrusion monitoring can further improve scalability and deployment in enterprise-level cybersecurity infrastructures.

In future developments, additional Explainable AI techniques can be integrated to provide more detailed visual explanations and improve user understanding of intrusion predictions. The framework can also be extended to support IoT networks, smart city infrastructures, and industrial cybersecurity systems where high-speed anomaly detection is essential. Furthermore, integrating blockchain technology and federated learning can enhance data privacy, distributed threat analysis, and secure collaborative intrusion detection across multiple network environments.

REFERENCES

[1] A. M. Riyad, M. Ahmed, and H. Almistarihi, “A quality framework to improve ids performance through alert post-processing,” International Journal of Intelligent Engineering and Systems, 2019.

[2] R. Alshammari, S. Sonamthiang, M. Teimouri, and D. Riordan, “Using neuro-fuzzy approach to reduce false positive alerts,” in Fifth Annual Conference on Communication Networks and Services Research (CNSR’07), pp. 345–349, 2007.

[3] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, “From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods,” IEEE Communications Surveys Tutorials, vol. 20, no. 4, pp. 3369–3388, 2018.

[4] K. A. Scarfone and P. M. Mell, “Sp 800-94. guide to intrusion detection and prevention systems (idps),” tech. rep., National Institute of Standards & Technology, Gaithersburg, MD, USA, 2007.

Author Profiles

Ms. K. Baby Ramya is working as an Assistant Professor in the Department of MCA at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She completed her MCA from Krishna University. She has nearly 3 years of teaching experience at SRK Institute of Technology. Her areas of interest include Machine Learning, Data Science, and Computer Applications.

Mrs. K. Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc.

Ms. G. Naga Durga Devi is an MCA Student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She Completed her Degree in B.Com (Computers) from Sri Krishnaveni Degree Kalasala Vijayawada. Her areas of interest are DBMS and Machine Learning with python.