

# FALSE POSITIVE IDENTIFICATION IN INTRUSION DETECTION USING XAI

<sup>1</sup>Chayapathi sruthi sri, <sup>2</sup>Obed UR Rahman, <sup>3</sup>K. Sai Praveen, <sup>4</sup>Rohit Vaidya, <sup>5</sup>P. Shruthi

Associate Professor in Department of CSE Sreyas Institute Of Engineering And Technology

<sup>2,3,4,5</sup>UG Scholar in Department of CSE Sreyas Institute Of Engineering And Technology

## Abstract

With the growing popularity of the Internet to access sensitive data, intrusion detection has become a necessary security measure. The evolution of Artificial Intelligence over the past few decades, particularly in Machine Learning techniques, combined with the availability of network traffic datasets, has created an immense development and research field for anomaly-based Intrusion Detection Systems. However, there is unanimity among published studies on this issue that this form of detection is more prone to false positives. In order to mitigate this problem, we propose a more effective method of identifying them, compared to using only the algorithm's confidence. For this, we hypothesize that the relevance given by the algorithm to certain attributes may be related to whether the detection is true or false. The method consists, therefore, in obtaining these features relevance through eXplainable Artificial Intelligence (XAI) and, together with a confidence measure, identifying detections that are more likely to be false. By using the LYCOS-IDS2017 dataset, it is possible to eliminate some percentage of the total false positives, with a loss of only less number of true positives. Conversely, by using only a confidence measure, the elimination of false positives is approximately just 50%, with a loss of 0.42% of true positives.

**KEYWORDS:** Intrusion detection, machine learning, explainability, XAI, false positive rate

## I INTRODUCTION

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. Explainable AI is used to describe an AI model, its expected impact and potential biases. Intrusion detection is an important activity that aims to improve the security level in computer systems. It complements other devices and techniques being considered the last line of defense. As attackers learn to circumvent firewalls, crack

passwords, steal cryptography keys, etc. Intrusion Detection Systems (IDS) become a mandatory device where sensible data is traveling. The first one compares characteristics of the monitored data against signatures or rules related to known attacks. The second one creates a model to represent normal (or benign) data and monitors deviations from it, which has the advantage of detecting unknown attacks, albeit at the price of more false positives. Advances in Machine Learning (ML) applied to anomaly IDS resulted, at least theoretically, in a sharp reduction in mistaken detections. Furthermore, it is not possible to assure the reliability of evaluations on synthetic datasets, where the highly complex open-world network traffic characteristics are hard to simulate. In this work, a post-processing method is proposed that aims to filter out false positives.

Explainable artificial intelligence (XAI) encompasses a set of processes and methods that enable human users to comprehend and trust the results produced by machine learning algorithms. XAI aims to describe an AI model, its expected impact, and potential biases, thereby fostering transparency and reliability in AI-driven systems.

## II LITERATURE SURVEY

Intrusion detection is a critical activity for enhancing the security of computer systems, acting as the last line of defense against attackers who bypass firewalls, crack passwords, or steal cryptographic keys. Intrusion Detection Systems (IDS) are essential when sensitive data is at risk. IDS generally operate in two modes: signature-based and anomaly-based detection. The signature-based approach compares monitored data against known attack signatures or rules, effectively identifying familiar threats. The anomaly-based approach, on the other hand, models normal (benign) data and detects deviations from this model, which helps identify unknown attacks but often results in more false positives.

Recent advancements in Machine Learning (ML) applied to anomaly-based IDS have theoretically reduced false positive rates. However, the reliability of evaluations on synthetic datasets remains questionable because these datasets cannot fully replicate the complex characteristics of real-world network traffic. This limitation underscores the need for more accurate and trustworthy intrusion detection methods.

In this context, a post-processing method is proposed to filter out false positives in IDS. By integrating XAI techniques, the proposed method aims to enhance the interpretability and transparency of intrusion detection outcomes. This approach allows security analysts to better understand the reasons behind alerts,

facilitating the refinement of IDS models and reducing the incidence of false alarms. Ultimately, leveraging XAI in IDS contributes to more robust, reliable, and trustworthy cybersecurity measures.

An Intrusion Detection System (IDS) maintains network traffic looks for unusual activity and sends alerts when it occurs. The main duties of an Intrusion Detection System (IDS) are anomaly detection and reporting, however, certain Intrusion Detection Systems can take action when malicious activity or unusual traffic is discovered. In this article, we will discuss every point about the Intrusion Detection System. What is an Intrusion Detection System?

A system called an intrusion detection system (IDS) observes network traffic for malicious transactions and sends immediate alerts when it is observed. It is software that checks a network or system for malicious activities or policy violations. Each illegal activity or violation is often recorded either centrally using an SIEM system or notified to an administration. IDS monitors a network or system for malicious activity and protects a computer network from unauthorized access from users, including perhaps insiders. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between 'bad connections' (intrusion/attacks) and 'good (normal) connections'.

#### Working of Intrusion Detection System(IDS)

An IDS (Intrusion Detection System) monitors the traffic on a computer network to detect any suspicious activity.

It analyzes the data flowing through the network to look for patterns and signs of abnormal behavior. The IDS compares the network activity to a set of predefined rules and patterns to identify any activity that might indicate an attack or intrusion. If the IDS detects something that matches one of these rules or patterns, it sends an alert to the system administrator. The system administrator can then investigate the alert and take action to prevent any damage or further intrusion. Classification of Intrusion Detection System(IDS) Network Intrusion Detection System (NIDS): Network intrusion detection systems (NIDS) are set up at a planned point within the network to examine traffic from all devices on the network. It performs an observation of passing traffic on the entire subnet and matches the traffic that is passed on the subnets to the collection of known attacks. Once an attack is identified or abnormal behavior is observed, the alert can be sent to the administrator. An example of a NIDS is installing it on the subnet where firewalls are located in order to see if someone is trying to crack the firewall.

**Host Intrusion Detection System (HIDS):** Host intrusion detection systems (HIDS) run on independent hosts or devices on the network. A HIDS monitors the incoming and outgoing packets from the device only and will alert the administrator if suspicious or malicious activity is detected. It takes a snapshot of existing system files and compares it with the previous snapshot. If the analytical system files were edited or deleted, an alert is sent to the administrator to investigate. An example of HIDS usage can be seen on mission-critical machines, which are not expected to change their layout.

**Protocol-based Intrusion Detection System (PIDS):** Protocol-based intrusion detection system (PIDS) comprises a system or agent that would consistently reside at the front end of a server, controlling and interpreting the protocol between a user/device and the server. It is trying to secure the web server by regularly monitoring the HTTPS protocol stream and accepting the related HTTP protocol. As HTTPS is unencrypted and before instantly entering its web presentation layer then this system would need to reside in this interface, between to use the HTTPS. **Application Protocol-based Intrusion Detection System (APIDS):** An application Protocol-based Intrusion Detection System (APIDS) is a system or agent that generally resides within a group of servers. It identifies the intrusions by monitoring and interpreting the communication on application-specific protocols. For example, this would monitor the SQL protocol explicitly to the middleware as it transacts with the database in the web server. **Hybrid Intrusion Detection System:** Hybrid intrusion detection system is made by the combination of two or more approaches to the intrusion detection system. In the hybrid intrusion detection system, the host agent or system data is combined with network information to develop a complete view of the network system. The hybrid intrusion detection system is more effective in comparison to the other intrusion detection system. **Prelude** is an example of Hybrid IDS. **Intrusion Detection System Evasion Techniques** **Fragmentation:** Dividing the packet into smaller packet called fragment and the process is known as fragmentation. This makes it impossible to identify an intrusion because there can't be a malware signature. **Packet Encoding:** Encoding packets using methods like Base64 or hexadecimal can hide malicious content from signature-based IDS. **Traffic Obfuscation:** By making message more complicated to interpret, obfuscation can be utilised to hide an attack and avoid detection. **Encryption:** Several security features, such as data integrity, confidentiality, and data privacy, are provided by encryption. Unfortunately, security features are used by malware developers to hide attacks and avoid detection.

### III EXISTINGSYSTEM

In literature they demonstrate the advantages of using a hybrid neuro-fuzzy approach to reduce the number of false alarms. The neuro-fuzzy approach was experimented with different background knowledge sets in DARPA 1999 network traffic dataset. The approach was evaluated and compared with RIPPER algorithm. Another research introduced to focused on reducing false positives in intrusion detection systems using data mining techniques. The model combines support vector machines (SVM), decision trees, and Naive Bayes to achieve their goal. The SVM is trained based on a new binary classification added to the dataset to specify if the instance is an attack or normal traffic. Attack traffic is then routed through a decision tree for classification. Finally, Naive Bayes and the decision tree vote on any unclassified attacks.

### ***Disadvantages***

The existing work does not explicitly mention considering the relevance of attributes in the detection process. This could potentially lead to false alarms not being effectively filtered out, as the approach may not take into account the specific attributes that contribute to false positives.

The existing work focuses on a hybrid neuro-fuzzy approach and compares it to the RIPPER algorithm. However, this approach's effectiveness might be limited when dealing with complex and evolving network traffic patterns, which could affect its ability to accurately reduce false alarms.

The existing work does not explicitly discuss the interpretability or explain ability of the algorithm's decisions.

The existing work relies on adding a new binary classification to the dataset for training SVM. This approach could introduce bias or might not always accurately represent the complexities of network traffic data, especially when compared to methods that focus on attribute relevance.

In the existing work, unclassified attack instances are subjected to voting by Naive Bayes and the decision tree. This approach might not always provide optimal results, as the combination of these two techniques might not effectively capture the nuances of false positives.

## **IV PROBLEMSTATEMENT**

The proposed Project for Intrusion Detection including the dataset, pre-processing, feature extraction and feature selection, algorithms, framework, and evaluation metrics, is presented and discusses the

evaluation results of the experiments performed, and finally concludes the project with framework predict of credit card fraud

## V PROPOSED SYSTEM

We propose a more effective method of identifying them, compared to using only the algorithm's confidence. For this, we hypothesize that the relevance given by the algorithm to certain attributes may be related to whether the detection is true or false. The method consists, therefore, in obtaining these features relevance through explainable Artificial Intelligence (XAI) and, together with a confidence measure, identifying detections that are more likely to be false. By using the LYCOS-IDS2017 dataset, it is possible to eliminate some percentage of the total false positives, with a loss of only less number of true positives.

### *Advantages*

1. In contrast, our work uses explainable Artificial Intelligence (XAI) to gain insights into the algorithm's decision-making process, allowing for a more transparent and understandable identification of false alarms.
2. Our work explicitly considers the relevance of attributes assigned by the algorithm, which can enhance the accuracy of identifying false alarms. This approach takes into account the importance of specific attributes in making decisions, potentially resulting in more accurate classification of false positives.
3. Our work leverages explainable Artificial Intelligence (XAI) to provide insights into the algorithm's decision-making process. This transparency can enhance trust and understanding by explaining why certain decisions are made.
4. Our work's method directly assesses the relevance of attributes assigned by the algorithm to determine the likelihood of false positives. This approach is more targeted and focused compared to the existing work.

## VI IMPLEMENTATION

**Data exploration:** using this module we will load data into system

**Processing:** Using the module we will read data for processing

**Splitting data into train & test:** using this module data will be divided into train & test

**Model generation:** Model building- Random Forest, KNN, Decision Tree, Naive Bayes, Neural Network, Voting Classifier - RF + AB, Stacking Classifier - RF + MLP with LightGBM

**User signup & login:** Using this module will get registration and login

**User input:** Using this module will give input for prediction

**Prediction:** final predicted displayed

### **Algorithms:**

**Random Forest:** Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

**KNN:** K-Nearest Neighbors Algorithm. The k-nearest neighbors' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

**Decision Tree:** Decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required splitting a node.

**Naive Bayes:** Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as: Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

**Neural Network:** Neural networks are artificial systems that were inspired by biological neural networks. These systems learn to perform tasks by being exposed to various datasets and examples without any task-specific rules. The idea is that the system generates identifying characteristics from the data they have been passed without being programmed with a pre-programmed understanding of these datasets.

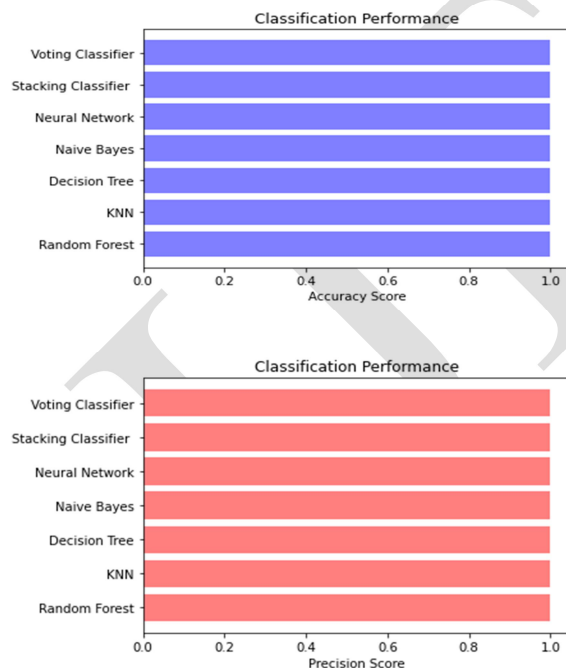


Neural networks are based on computational models for threshold logic. Threshold logic is a combination of algorithms and mathematics. Neural networks are based either on the study of the brain or on the application of neural networks to artificial intelligence. The work has led to improvements in finite automata theory.

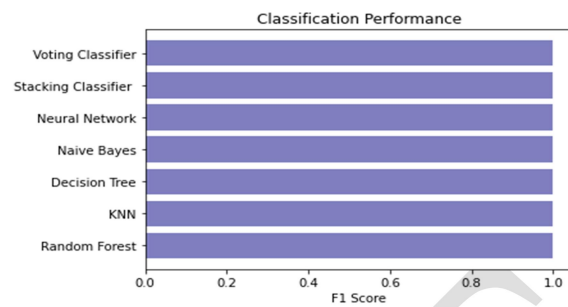
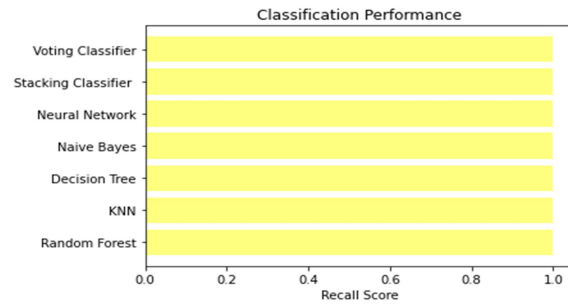
Voting Classifier - RF + AB: A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output.

Stacking Classifier - RF + MLP with LightGBM: Stacking is a way of ensembling classification or regression models it consists of two-layer estimators. The first layer consists of all the baseline models that are used to predict the outputs on the test datasets. The second layer consists of Meta-Classifer or Regressor which takes all the predictions of baseline models as an input and generate new predictions.

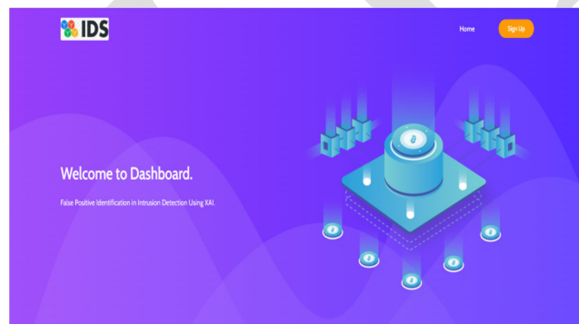
## VII RESULTS

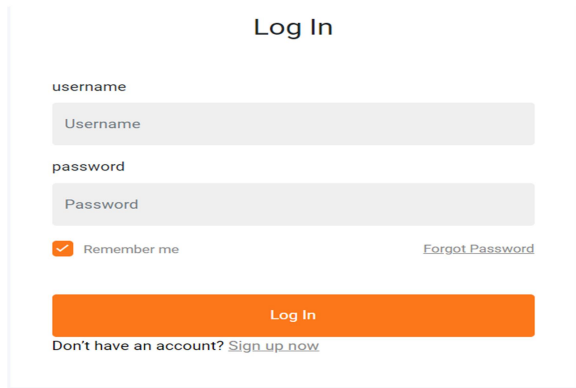






ML Model	Accuracy	f1_score	Recall	Precision
Random Forest	0.871	0.869	0.871	0.897
KNN	0.871	0.869	0.871	0.897
Decision Tree	0.871	0.869	0.871	0.897
Naive Bayes	0.566	0.598	0.566	0.863
Neural Network	0.871	0.869	0.871	0.897
Extension Stacking Classifier	0.871	0.869	0.871	0.897
Extension Voting Classifier	1.000	1.000	1.000	1.000





Log In

username  
Username

password  
Password

☒ Remember me [Forgot Password](#)

[Log In](#)

Don't have an account? [Sign up now](#)

Result: **There is an Attack Detected, Attack Type is DDoS!**

Result: **There is an No Attack Detected, it is Normal!**

## VIII CONCLUSION

An anomaly-based IDS has the potential to detect new unknown attacks, but it is also more prone to generate false positives. Unlike misuse-based IDS, whose signature in itself explains the reason for the (false) detection, it is not trivial to understand wrong detections from the IDS powered by complex ML algorithms. In this sense, XAI arises as a new possibility to handle false positives. The use of XAI attributes, especially SHAP ones, makes it possible to obtain percentages of analysis sets with a higher density of false positives. The method acts as a way of triage, shortening the number of samples where the analysts search for false positives, thus enhancing their efficiency. Even though the better performance was obtained compared to not using XAI attributes, it is not always possible to obtain percentages with a majority of false positives. This points to a need for improvement, which can be achieved in future works. One suggestion is to use other XAI techniques in order to reach better results with the confidence

combination. Improvements also can be done on the second ML algorithm (the FP detector) choice, preferably those more suitable to unbalanced sets. There is also a need for a study related to the impact of feature selection before applying XAI techniques. SHAP, for example, assumes statistical independence of the attributes, which may not happen in the general case. Then, the minimization of correlation through feature selection can result in SHAP values with better quality, which in turn can improve the method.

## 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.
- [5] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.
- [6] E. K. Viegas, A. O. Santin, and L. S. Oliveira, "Toward a reliable anomaly-based intrusion detection in real-world environments," *Computer Networks*, vol. 127, pp. 200–216, 2017.
- [7] Internet Steering Committee project in Brazil, "Total internet data traffic in brazil," 2022. <https://ix.br/agregado/>. Accessed on: Nov. 11, 2022.
- [8] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3237–3243, 2018.
- [9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, pp. 4765–4774, Curran Associates, Inc., 2017.

- [10] L. S. Shapley, A Value for n-Person Games, pp. 307–317. Princeton University Press, 1953.
- [11] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?”: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.
- [12] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” ArXiv, vol. abs/1605.01713, 2016.
- [13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PLOS ONE, vol. 10, pp. 1–46, 07 2015.
- [14] MIT Lincoln Laboratory, “1999 darpa intrusion detection evaluation dataset,” 1999. <https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset>. Accessed on: Nov. 16, 2022.
- [15] G. P. Spathoulas and S. K. Katsikas, “Reducing false positives in intrusion detection systems,” Computers & Security, vol. 29, no. 1, pp. 35–44, 2010.
- [16] P. Pitre, A. Gandhi, V. Konde, R. Adhao, and V. Pachghare, “An intrusion detection system for zero-day attacks to reduce false positive rates,” in 2022 International Conference for Advancement in Technology (ICONAT), pp. 1–6, 2022.
- [17] H. Kim, Y. Lee, E. Lee, and T. Lee, “Cost-effective valuable data detection based on the reliability of artificial intelligence,” IEEE Access, vol. 9, pp. 108959–108974, 2021.
- [18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” 2017.
- [19] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, “A survey of network-based intrusion detection data sets,” Computers & Security, vol. 86, pp. 147–167, 2019.
- [20] G. Engelen, V. Rimmer, and W. Joosen, “Troubleshooting an intrusion detection dataset: the cicsids2017 case study,” in 2021 IEEE Security and Privacy Workshops (SPW), pp. 7–12, 2021.

[21] A. Rosay, E. Cheval, F. Carlier, and P. Leroux, “Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017,” in 8th International Conference on Information Systems Security and Privacy, pp. 25– 36, SCITEPRESS - Science and Technology Publications, Feb. 2022.

[22] <http://lycos-ids.univ-lemans.fr/>

[23] M. Ring, A. Dallmann, D. Landes, and A. Hotho, “IP2Vec: Learning similarities between ip addresses,” in 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 657–666, 2017.