ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152

www.ijmece .com
Vol 12 Issue.1Jan 2024

False Positive Identification in Intrusion Detection Using XAI

[1]Tasneem Rahath, [2]Tadakamalla Sahasra, [3]C Radhika, [4]Palle Harika

[1]Assistant professor in Department of Information Technology Bhoj Reddy Engineering College for Women

[2,3] UG Scholars in Department of Information Technology Bhoj Reddy Engineering College for Women

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.

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.

II LITERATURE SURVEY

***1999 DARPA INTRUSION DETECTION EVALUATION DATASET*:**

There were two parts to the 1999 DARPA Intrusion Detection Evaluation: an off-line evaluation and a real-time evaluation. Intrusion detection systems were tested in the off-line evaluation using network traffic and audit logs collected on a simulation network. The systems processed these data in batch mode and attempted to identify attack sessions in the midst of normal activities. Intrusion detection systems were delivered to the Air Force Research Laboratory (AFRL) for the real-time evaluation. These systems were inserted into the AFRL network test bed and attempted to identify attack sessions in real time during normal activities. Intrusion detection systems were tested as part of the off-line evaluation, the real-time evaluation or both.

Cost-Effective Valuable Data Detection Based on the Reliability of Artificial Intelligence

Many previous studies have investigated applying artificial intelligence (AI) to cyber security. Despite considerable performance advantages, AI for cyber security requires final confirmation by an analyst, e.g. malware misdetection can cause significant adverse side effects. Thus, a human analyst must check all AI predictions, which poses a major obstacle to AI expansion. This paper proposes a reliability indicator for AI prediction using explainable artificial intelligence and statistical analysis techniques. This will enable analysts with limited daily workload to focus upon valuable data, and quickly verify AI predictions. Analysts generally make decisions based on several features that they know exactly what they mean, rather than all available features. Since the proposed reliability indicator is calculated using features meaningful to analysts, it can be easily understood and hence speed final decisions. To verify the performance of the proposed method, an experiment was conducted using the IDS dataset and the malware dataset. The AI error was detected better than the existing AI model at about 114% in IDS and 95% in malware. Thus, cyberattack response could be greatly improved by adopting the proposed method.

***A Quality Framework to Improve IDS Performance Through Alert Post-Processing*:**

An intrusion detection system is one of the network security tools installed to monitor suspicious activity in the network and act as a

last line of defense. It normally notifies about the skeptical activity occurred in the network using sensors by sending alarms to the administrator. However, the IDS present in the large network generates not only a large number of alerts but also abundant false alerts. These generated alerts are very difficult to handle as it increases the burden for the network administrator and also pulls down the performance of the defense system. In order to overcome the issue, various countermeasures have been proposed. Commonly, to increase the quality of alerts, the alerts are post-processed in such a way that the false alerts are filtered out thereby refining the performance of the IDS defense. In this paper, we propose an IDS quality framework using alert post-processing techniques to separate out the false alerts generated by various sensors in the network. At low level alert post-processing, the priority scores are assigned based on the quality measures to filter the irrelevant alerts having less significance. At high level alert post-processing, higher level operations such as alert aggregation, clustering, and hyper alert correlation have been carried out to minimize the number of alerts and the high level report consisting of significant alerts is presented to the administrator. Experiments have been conducted using DARPA 2000 dataset to assess the performance of the proposed system. The system has produced pleasing results than many of the existing methods with 95% of alert reduction rate, 99% of completeness and 100% of soundness towards enlightening the quality of the alerts generated by the IDS.

From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods:

Over the last five years there has been an increase in the frequency and diversity of network attacks. This holds true, as more and more organizations admit compromises on a daily basis. Many misuse and anomaly based intrusion detection systems (IDSs) that rely on either signatures, supervised or statistical methods have been proposed in the literature, but their trustworthiness is debatable. Moreover, as this paper uncovers, the current IDSs are based on obsolete attack classes that do not reflect the current attack trends. For these reasons, this paper provides a comprehensive overview of unsupervised and hybrid methods for intrusion detection, discussing their potential in the domain. We also present and highlight the importance of feature engineering techniques that have been proposed for intrusion detection. Furthermore, we discuss that current IDSs should evolve from simple detection to correlation and attribution. We descant how IDS data could be used to reconstruct and correlate attacks to identify attackers, with the use of advanced data analytics techniques. Finally, we argue how the present IDS attack classes can be extended to match the modern attacks and propose three new classes regarding the outgoing network communication.

III EXISTING SYSTEM

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 PROBLEM STATEMENT

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

VII ALGORITHMS USED

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-Classifier or Regressor which takes all the predictions of baseline models as an input and generate new predictions.

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 cicids2017 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.