JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 02, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

MACHINE LEARNING MODEL FOR IDENTIFY AND ANALYZE INTRUSION DETECTION SYSTEMS

Macherla Vamshi

(M.Tech Student), Guide: Dr.E.BALAKRISHNA (Associate.prof)

Department of Computer Science and Engineering

Vaagdevi College of Engineering

(UGC Autonomous, Accredited by NBA, Accredited by NAAC with "A")

Bollikunta,Warangal 506005 (T.S)

ABSTRACT:

In recent years, there has been an increase in the number of networked computers, making them open to various forms of cyber attack. Machine learning-based Intrusion Detection Systems (IDS) have become increasingly popular as a means of protecting networks from these types of risks. The effectiveness of present IDSs, particularly for less common attack types, is limited by issues such as stale and unbalanced datasets. We propose developing and evaluating an IDS based on machine learning techniques including K-Nearest Neighbour, Random Forest, Support Vector Machine, and Decision Tree. We use the recent, realistic, and imbalanced CSE-CIC-IDS2018 dataset. Experimental results show that our method considerably increases the detection rate for infrequently seen intrusions, making IDSs more effective and efficient against modern cyber threats.

Keywords: *Intrusion detection.knn,svm, CSE-CIC-IDS2018 dataset,DT,RM*

INTRODUCTION:

As our daily lives become more and more dependent on Internet and networked computers, the possibility of cyber attacks has come to be as an important cause of concern. By exploiting weaknesses in computers, hackers are constantly improving their capacity to get into computer networks. Intrusion Detection Systems (IDS) have become important defences in the face of these threats. Intruder detection systems (IDSs) use machine learning methods to analyse data from a network for indications of hacking attempts. Existing IDSs face challenges, however, because they are trained on obsolete and unbalanced datasets, which reduces their efficacy, especially against infrequently seen attack types. K-Nearest Neighbour, Random Forest, Support Vector Machine, and Decision Tree are some of the machine learning methods we propose utilising to build an IDS for this project. In this work, we use the recent and skewed CSE-CIC-IDS2018 dataset to enhance the performance of intrusion detection systems, particularly with respect to infrequent incursions. The goal of this study is to improve IDS performance so that networked computers are better protected from the increasing complex and dangerous cyber threats.

LITERATURE REVIEW

[1] Title: Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset Author: Gozde Karatas; Onder Demir; Ozgur Koray Sahingoz DescriptionIn recent years, the increasing number of computers that are networked has been partly due to the widespread adoption of the Internet. Because of vulnerabilities in the servers,

Submitted: 28/07/2023 **Accepted**: 18/08/2023

criminals can breach protected computers by employing both established and novel forms of attack. One popular method of computer defence against such threats is the use of an Intrusion Detection System (IDS) that has been pre-trained on a large data set through the application of machine learning algorithms. The used datasets only contain data from a specified time period and set of networks, hence they are not up to date. They're also unbalanced and are unable to retain sufficient data to fend off any prospective attack. Current IDSs are less effective due to these old and unbalanced datasets, which is especially problematic for infrequently seen attack types. Using the K Nearest Neighbour, Random Forest, Gradient Boosting, Adaboost, Decision Tree, and Linear Discriminant Analysis algorithms, this research proposes six machine-learning-based IDSs. Rather than using earlier, mostly-worked security datasets, a more realistic IDS is implemented using the just released CSE-CIC-IDS2018. Additionally, the chosen data set is unbalanced. Therefore, a synthetic data generation model called Synthetic 6 Minority Oversampling Technique (SMOTE) is used to minimise the imbalance ratio in order to boost the system's efficiency depending on attack types and decrease missed incursions and false alerts. Using this method, we can produce data for subclasses and bring their total number of records up to those of the major classes. The suggested method greatly enhances the detection rate for infrequently seen incursions, as shown by experimental findings. [2] Title: An Intrusion Detection System for Imbalanced Dataset Based on Deep Learning Description: With the help of deep learning (DL), anomaly-based Network intrusion detection systems (NIDS) (the technology used to detect new assaults) have shown impressive performance. There are still limitations to these NIDSs, though. The majority of NIDS datasets are severely unbalanced, with far more samples from benign traffic types than malicious ones. The imbalanced class problem hinders the effectiveness of deep learning classifiers for underrepresented groups by training the classifier to favour the dominant group. This research suggests a hybrid strategy for dealing with the imbalance issue, which has the potential to increase the detection rate for minority classes while maintaining efficiency. This method combines undersampling with the noise-reduction techniques of the Tomek link to create a hybrid approach called Synthetic Minority Over-Sampling (SMOTE). Long Short-Term Memory Networks (LSTMs) and Convolutional Neural Networks (CNNs) are two deep learning models used in this research to improve the quality of the intrusion detection system. We put our suggested model through its paces on the NSL-KDD and CICIDS2017 datasets to see how useful it is. In this study, we train the learning models on a separate test set and then evaluate them using 10-fold cross validation. As can be seen from the experimental results, the suggested model achieved an overall accuracy and Fscore of 99.57% and 98.98% on LSTM and 99.70% and 99.27% on CNN in a multiclass classification task using the NSLKDD dataset. With CICICD2017, LSTM achieved an overall accuracy of 99.82% and a Fscore of 98.65%, whereas CNN achieved an accuracy of 99.85% and a Fscore of 98.98%. [3] Title: Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic Author: Razan Abdulhammed; Miad Faezipour; Abdelshakour Abuzneid; Arafat AbuMallouh 7 DescriptionAttacks on computer networks have expanded considerably in recent years, and the methods employed by cybercriminals only continue to develop in

sophistication and sophistication. In addition, the complexity and prevalence of uneven class distributions in most datasets point to the necessity of further investigation. The purpose of this paper is to use the most recent Coburg Intrusion Detection Dataset-001 (CIDDS-001) dataset to create an efficient intrusion detection system utilising several strategies for handling imbalanced datasets. Extensive research and experimental evaluations are conducted on CIDDS-001's sampling efficacy using deep neural networks, random forests, voting, variational autoencoders, and stacked machine learning classifiers. Since the suggested system can handle the imbalanced class distribution with less samples, it is more practical for use in real-time data fusion challenges that aim to classify data in real-time. [4] Title: Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning Author: Lan Liu; Pengcheng Wang; Jun Lin; Langzhou Liu Description: Malicious cyber-attacks can frequently hide in vast amounts of regular data when there is an imbalance in the network flow. In cyberspace, it demonstrates a high level of stealth and obfuscation, making it challenging for Network Intrusion Detection System (NIDS) to guarantee the accuracy and promptness of detection. In order to detect intrusions in unbalanced network traffic, this paper investigates machine learning and deep learning. To address the issue of class imbalance, a unique Difficult Set Sampling Technique (DSSTE) algorithm is suggested. First, separate the imbalanced training set into the challenging set and the easy set using the Edited Nearest Neighbor(ENN) algorithm. The majority samples in the challenging set will then be reduced using the KMeans technique. Zoom in and out on the continuous characteristics of the minority samples in the challenging set and create new samples to boost the minority population. Finally, a new training set is created by combining the easy set, the compressed set of the majority in the tough, and the minority in the difficult set. The method evens out the initial training set's imbalance and provides targeted data augmentation for the underrepresented class that needs to learn. It enables the classifier to perform better during classification and better learn the distinctions during the training stage. We do trials using the more recent and comprehensive intrusion dataset CSE-CIC-IDS2018 as well as the venerable intrusion dataset NSL-KDD to validate the suggested technique. We employ traditional classification models including AlexNet, Mini-VGGNet, Long and Short Term Memory (LSTM), Random Forest (RF), Support Vector Machine (SVM), and XGBoost. We contrast the other 24 approaches, and the experimental findings show that our suggested DSSTE algorithm performs better than the competing methods. [5]Title: An End-to-End Framework for Machine Learning-Based Network Intrusion Detection System [4] Description: Network intrusion detection systems have challenges from design to operation due to the rise in connected devices and the attackers' ongoing evolution of their tactics. As a result, machine learning methods are increasingly being used in network intrusion detection systems. The data set employed in these research, however, is no longer relevant in terms of background and attack traffic. In order to enable the full deployment of the solution, this study describes the AB-TRAP framework, which permits the usage of updated network traffic and takes operational considerations into account. The AB-TRAP framework consists of five steps: (i) creating the attack data set; (ii) creating the genuine data set; (iii) training machine learning models; (iv) realising (implementing) the models; and

(v) assessing the performance of the realised model after deployment. To stop TCP port scanning attacks, we tested the AB-TRAP in both local (LAN) and international (internet) contexts. A decision tree with minimum CPU and RAM use in kernel space produced a f1-score of 0.96 and an area under the ROC curve of 0.99 for the LAN study scenario. For the internet scenario, a single-board computer has an average user-space overhead of 1.4% CPU and 3.6% RAM, an average f1-score of 0.95, an average area under the ROC curve of 0.98, and eight machine learning algorithms. The reproducibility, usage of the most recent network traffic, attacks, and addressing of the concerns regarding the model's realisation and deployment are some of this framework's most important attributes.

Problem definition:

The current intrusion detection system is based on conventional techniques that make use of unreliable and skewed datasets for IDS training. These data sets are not up to date because they are taken from specific networks over short time periods and do not include information on newly discovered forms of attack. This has a devastating effect on the effectiveness and efficiency of contemporary IDSs, especially when it comes to identifying uncommon intrusions. In addition, there are more false positives and omissions since the datasets are not evenly distributed. Using the more realistic and up-to-date CSE-CIC-IDS2018 dataset, our proposal suggests implementing a machine learning-based IDS to increase intrusion detection performance.

METHDOLOGY:

The suggested system employs a machine learning-based methodology, specifically the K Nearest Neighbour, Random Forest, Support Vector Machine, and Decision Tree algorithms, to overcome the shortcomings of the current IDS. We forego the use of antiquated datasets in favour of the more accurate and up-to-date CSE-CIC-IDS2018 data set, which is a better reflection of today's cyber risks. We use methods to improve the system's efficiency and accuracy in recognising rarely observed incursions in order to compensate for the skewed data set. Through experimental evaluation, we show that our proposed method considerably enhances the detection rate for various attack types, boosting the IDS's overall performance and making it more resilient to more complex cyberattacks.

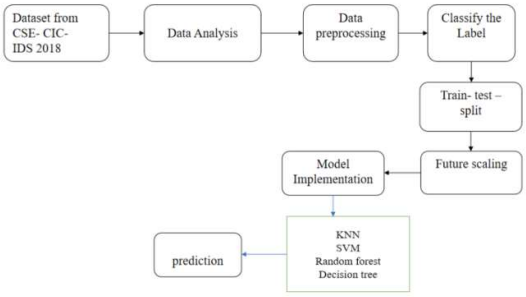


Figure 1: System Architecture

Data set Collection

A data set (or data set) is a group of related pieces of information, typically displayed in a tabular format. One variable is represented in each column. Each row represents a single record in the dataset. Values for factors like an item's height and weight are listed. A datum is a single numerical value. The data used in this study comes from the CSE-CIC-IDS2018 data set, which is a large and recently updated database of network traffic information. The data set was compiled for the purpose of training and assessing the proposed IDS, and it features a wide variety of attack types because it was created with intrusion detection research in mind. In order to analyse and detect intrusions, it is necessary to collect data on network traffic from a variety of sources and verify its accuracy and relevance.

Pre-Processing

To transform raw feature vectors into a representation that is more suited for the downstream estimators, the sklearn.preprocessing package contains various common utility functions and transformer classes. Standardising the data set is helpful for learning algorithms in general. If there are outliers in the data, robust scalers or transformers are better. Scalers, transforms, and normalizers all have varied behaviours on a dataset with marginal outliers, and these differences are highlighted in Examine the impact of various scalars on data containing outliers. The CSE-CIC-IDS2018 dataset's raw network traffic data is cleaned up by hand to get rid of extraneous details. Feature selection, normalisation, and dealing with missing data may all fall under this category.

Splitting of data set

The cleaned data set is then split into a training set and a test set. Machine learning models are trained using the training set and their efficacy is measured using the testing set.

Model implementation:

K-Nearest Neighbour, Random Forest, Support Vector Machine, and Decision Tree are only some of the proposed models that may be implemented with the help of machine learning libraries. Model architectures are defined, relevant hyperparameters are chosen, and the models are fitted to the training data during implementation. Following preprocessing, the dataset is used to train the models, which discover latent patterns and associations between the input features and intrusion labels. After the models have been trained, they may be used to analyse new data from the network and make predictions about the presence or absence of intrusions. The model's implementation uses machine learning algorithms to guarantee fast, precise identification of intrusions.

K Nearest Neighbor (KNN): This algorithm assigns categories to data points depending on how close they are to one another. Unlabeled instances are given labels based on the labels of their k nearest neighbours. As an ensemble learning technique, Random Forest builds many decision trees and uses a weighted average of their predictions to reach a conclusion. It boosts precision by preventing overfitting and enabling the capture of a wider variety of patterns in the data.

Support Vector Machine (SVM): SVM is an effective supervised learning technique that uses an optimum hyperplane to categorise data points into distinct groups. Its goal is to increase the gap between categories so that they can be more accurately sorted.

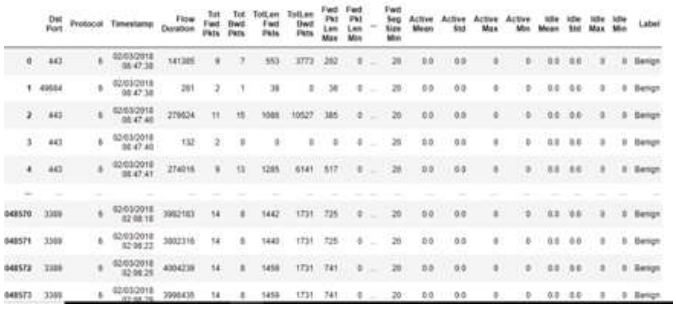
Decision Tree: The if-then rules of a decision tree are used to organise data in a hierarchical structure. The dataset is partitioned according to attribute values, and a tree-like structure is created from which judgements can be made at each node. Prediction, Module 5 Predictions can be made on previously unseen instances of network traffic data once the machine learning models have been trained and prepared. The new instance's features are fed into the learned models during prediction, and a prediction or classification is then obtained. In the context of intrusion detection, models examine the input data's patterns and properties to determine if the data is benign or malicious. To detect and respond to potential threats in real time, real-time intrusion detection relies heavily on the prediction phase.

ALGORITHEM:

Output: Results as R

1. **Start**
2. **Input dataset(S)**
3. **Pre-processing (S)**
4. **Extract features from training set()**
5. **For each model m in M**
6. **Train the model m**
7. **End For**
8. **For each model m in M**
9. **Use model for testing**
10. **Evaluate**
11. **Display results**
12. **End For**
13. **End**
14. **Return R**

Result analysis:



	Dst Port	Protocol	Timestamp	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	...	Fwd Seg Size Min	Active Mean	Active Std	Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min	Label
0	443	6	02/03/2018 08:47:38	141385	9	7	553	3773	262	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
1	49684	6	02/03/2018 08:47:38	261	2	1	38	0	38	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
2	443	6	02/03/2018 08:47:40	279624	11	15	1088	10527	385	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
3	443	6	02/03/2018 08:47:40	132	2	0	0	0	0	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
4	443	6	02/03/2018 08:47:41	274016	9	13	1285	6141	517	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
948570	3389	6	02/03/2018 02:08:18	3952183	14	8	1442	1731	725	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
948571	3389	6	02/03/2018 02:08:22	3802316	14	8	1440	1731	725	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
948572	3389	6	02/03/2018 02:08:25	4004239	14	8	1458	1731	741	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign
948573	3389	6	02/03/2018 02:08:29	3996435	14	8	1458	1731	741	0	...	20	0.0	0.0	0	0	0.0	0.0	0	0	Benign

Figure 3:Dataset Visualation

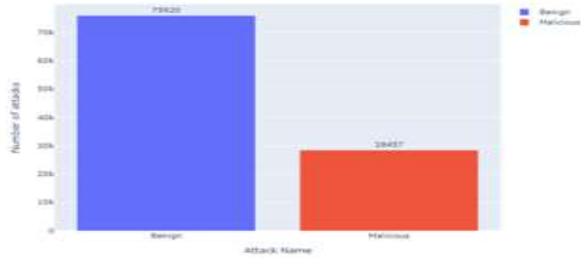


Figure 4: Visualation

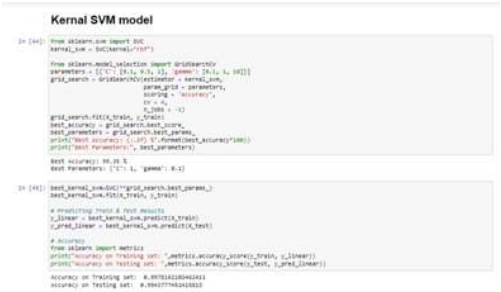


Figure 5: Model Building

Conclusion

In order to overcome the shortcomings of current Intrusion Detection Systems (IDS), this study suggests one that makes use of machine learning. Our proposed IDS uses the most recent CSE-CIC-IDS2018 dataset in conjunction with state-of-the-art algorithms like K Nearest Neighbour, Random Forest, Support Vector Machine, and Decision Tree to improve efficiency and accuracy in detecting intrusions, especially for less common attack types. Incorporating numerous algorithms and using a realistic dataset both contribute to better defence against complex and ever-evolving cyber attacks. The experimental findings verify the efficacy of our method, demonstrating a notable rise in the detection rate for diverse invasions. In sum, this study strengthens IDS technology and equips networked computers with a better defence against evolving cyber threats. The results of the comparative algorithms are in.

REFERENCES

[1] Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. IEEE access, 8, 32150-32162..

[2] Mbow, M., Koide, H., & Sakurai, K. (2021, November). An intrusion detection system for imbalanced dataset based on deep learning. In 2021 Ninth International Symposium on Computing and Networking (CANDAR) (pp. 38- 47). IEEE.

[3] Abdulhammed, R., Faezipour, M., Abuzneid, A., & AbuMallouh, A. (2018). Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic. IEEE sensors letters, 3(1), 1-4..

[4] Liu, L., Wang, P., Lin, J., & Liu, L. (2020). Intrusion detection of imbalanced network traffic based on machine learning and deep learning. Ieee Access, 9, 7550-7563.

[5] Ayachi, Y., Mellah, Y., Berrich, J., & Bouchentouf, T. (2020, November). Increasing the Performance of an IDS using ANN model on the realistic cyber dataset CSE-CIC-IDS2018. In

2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT) (pp. 1-4). IEEE.
[6] Al, S., & Dener, M. (2021). STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment. Computers & Security, 110, 102435.
[7] Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., & Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. Electronics, 8(3), 322. 46
[8] Lin, Y. D., Liu, Z. Q., Hwang, R. H., Nguyen, V. L., Lin, P. C., & Lai, Y. C. (2022). Machine learning with variational AutoEncoder for imbalanced datasets in intrusion detection. IEEE Access, 10, 15247-15260.
[9] Lin, Y. D., Liu, Z. Q., Hwang, R. H., Nguyen, V. L., Lin, Dwibedi, S., Pujari, M., & Sun, W. (2020, November). A comparative study on contemporary intrusion detection datasets for machine learning research. In 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 1-6). IEEE.
[10] Kasim, Ö. (2020). An efficient and robust deep learning based network anomaly detection against distributed denial of service attacks. Computer Networks, 180, 107390.
[11] Akshay Kumaar, M., Samiayya, D., Vincent, P. M., Srinivasan, K., Chang, C. Y., & Ganesh, H. (2022). A Hybrid Framework for Intrusion Detection in Healthcare Systems Using Deep Learning. Frontiers in Public Health, 9, 824898.
[12] Tran, N., Chen, H., Jiang, J., Bhuyan, J., & Ding, J. (2021). Effect of Class Imbalance on the Performance of Machine Learning-based Network Intrusion Detection. International Journal of Performability Engineering, 17(9).
[13] Khan, M. A. (2021). HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system. Processes, 9(5), 834.
[14] Guarino, I., Bovenzi, G., Di Monda, D., Aceto, G., Ciuonzo, D., & Pescapé, A. (2022, July). On the use of machine learning approaches for the early classification in network intrusion detection. In 2022 IEEE International Symposium on Measurements & Networking (M&N) (pp. 1-6). IEEE.
[15] 47 Comparative Analysis of Entropy Weight Method and C5 Classifier for Predicting Employee Churn. M Chaudhary, L Gaur… - 2022 3rd International …, 2022 - ieeexplore.ieee.org
[16] Alsoufi, M. A., Razak, S., Siraj, M. M., Nafea, I., Ghaleb, F. A., Saeed, F., & Nasser, M. (2021). Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review. Applied sciences, 11(18), 8383.
[17] Leevy, J. L., & Khoshgoftaar, T. M. (2020). A survey and analysis of intrusion detection models based on cse-cic-ids2018 big data. Journal of Big Data, 7(1), 1-19.
[18] Kumaar, M. A., Samiayya, D., Vincent, P. D. R., Srinivasan, K., Chang, C. Y., & Ganesh, H. (2021). A hybrid framework for intrusion detection in healthcare systems using deep learning. Frontiers in Public Health,