IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Increasing The Performance of Machine Learning-Based Models on an Imbalanced and Up-To-Date Dataset

Ankit Rathod[1], Omkar Thorat[2], Rahul Sanap[3], Prof. Sagar Dhanake[4]

Student, Computer Engineering[1, 2, 3]

Assistant Professor, Computer Engineering[4]

D Y Patil Institute of Engineering and Technology, Ambi, Pune, Maharashtra, India

Abstract: In growing times, the use of internet is spreading at a lightning speed and which as a result N/Wed computer has been increasing in our daily lives. This expanding chain of N/Wed computer weakens the servers which enable hackers to intrude on computer by using various means which may be know as well as unknown and makes them even harder to detect. So as a protection to the computers the Intrusion Detection System (MODEL) is introduced which is trained with some MACHINE LEARNING techniques by making use of previous available data. . The used datasets were collected during a limited period in some specific N/W and generally don't contain up-to-date data. In this paper, we propose six machine-learning-based MODELs by using Random Forest, Gradient Boosting, Ada boost, Decision Tree, and Linear Discriminant Analysis algorithm. To implement a more realistic MODEL, an up-to-date security dataset, CSE-CIC-MODEL2018, is used instead of older and mostly worked datasets. Therefore, to increase the efficiency of the system depending on attack types and to decrease missed intrusions and false alarms, the imbalance ratio is reduced by using a synthetic data generation model called Synthetic Minority Oversampling Technique. Experimental results demonstrated that the proposed approach considerably increases the detection rate for rarely encountered intrusions.

Keywords: Model, Intrision Detection, Smote, Machine Learning, Cse-Cic- Model 2018, Im Balanced Dataset

I. INTRODUCTION

Due to technological developments, most of the real-world transactions have been available in the cyber world. Thus, many operations, such as banking, shopping, online examinations, electronic commerce, and communication are used extensively within this new environment. With the widespread use of smartphones, people can connect to this global N/W and perform transactions at any time and from anywhere. Although this digitalization facilitates the daily work of human beings, due to the weakness of the servers and the newly emerged intrusion techniques, N/Ws are often attacked by the intruders who take advantage of the anonymous nature of the Internet not only to steal some information or money but also to slow down the operation of N/W services. Security administrators traditionally prefer password protection mechanisms, encryption techniques, and access controls in addition to firewalls as a means of protecting the N/W. However, these techniques are not sufficient for protecting the system

Motivation:

N/W security plays an essential role in secure communication and avo model financial loss and crippled services due to N/W intrusions. Intruders generally exploit the flaws of popular software to mount a variety of attacks against N/W

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

computer systems. The damage caused in the N/W attacks may vary from a little disruption in service to on developing financial loss. Recently, intrusion detection systems (MODELs) comprising MACHINE LEARNING techniques have emerged for handling unauthorized usage and access to N/W resources. With the passage of time, a wide variety of MACHINE LEARNING techniques have been designed and integrated with MODELs. Still, most of the MODELs reported poor intrusion detection results using false positive rate and detection rate. For solving these issues, we have proposed this system to increase the performance of intrusion detection through various MACHINE LEARNING algo. to give higher accuracy in intrusion detection than already proposed systems.

II. PROBLEM STATEMENT

Problem Definition

Our aim is to develop a system that Increases the Performance of MACHINE LEARNING-Based MODELs on a Imbalanced and Up-to-Date Dataset.

Goals and Objectives:

- The main goal of this system is to increase the performance of the IDS.
- To achieve this we trained our system with machine learning algorithms.
- We have used the data set CSE-CIC-IDS 2018 to increase the accuracy of the system.
- Machine Learning algorithms are used to train the machine are as follows:-
  1. Ada boost
  2. Decision Tree
  3. Random Forest
  4. KNN
  5. Gradient Boosting
  6. Linear Discriminant Analysis

III. LITERATURE REVIEW

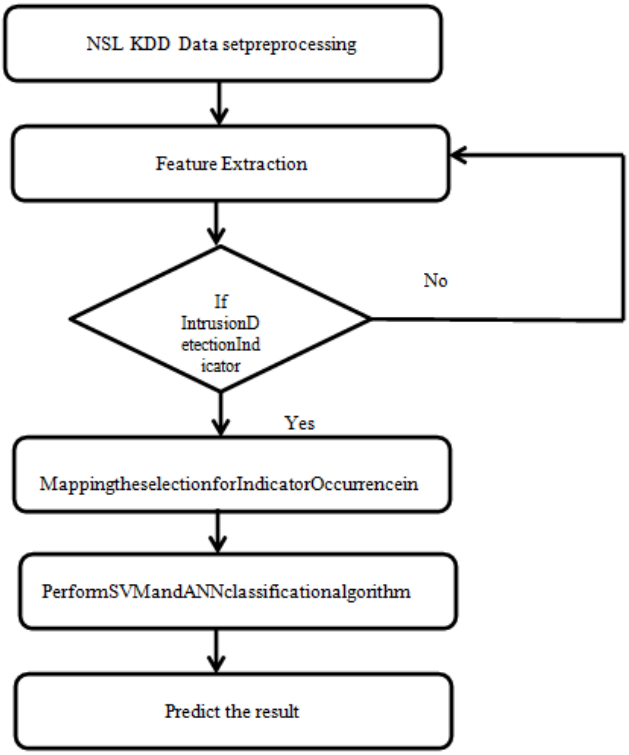
Sr. No.	Name of Paper	Authors	Publication Name	Published On	Elements
1.	Increasing the Performance of MACHINE LEARNING-Based MODELs on an Imbalanced and up-to-date dataset.	GOZDE KARATAS, ONDER DEMIR, OZGUR KORAY SAHINGOZ	IEEE	2020	Compared with other algo. of the same kind, the effect of the algo. is obviously improved, and it has a great practical value.
2.	detailed investigation and analysis of using MACHINE LEARNING techniques for intrusion detection.	P. Mishra. V.Varadharajan, U. Tupakula E. S. Pilli.	IEEE	2019	Even if an optimal feature set is sufficient for analyzing the behavior of an attack, it is not good for analyzing the behavior of other attacks. Hence, there is a need to define the optimal feature subset and a suitable technique for each type of attack.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

3.	Using MACHINE LEARNING to detect DoS attacks in wireless sensor N/Ws.	A. I. Al-issa, M. Al-Akhras, M. S. Alsahli, and M. Alawairdhi	IEEE	2019	The desicion trees technique achieved better(higher) true positive rate and better(lower) false positive rate than support vector machine.
4.	An adaptive ensemble MACHINE LEARNING model for intrusion detection.	X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu.	IEEE	2019	Ensemble MACHINE LEARNING has a. good generalization effect, which is Worthy of continuous promotion and optimization in the field of N/W security research and application

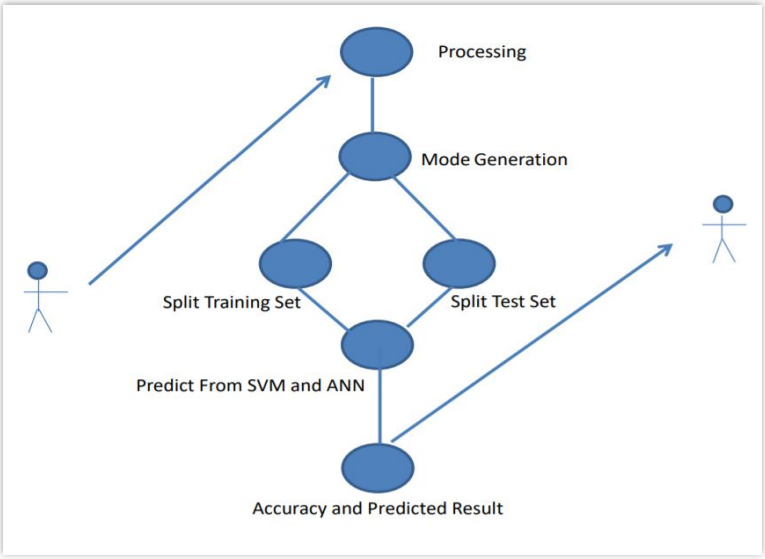
IV. SYSTEM DESIGN AND FLOW

Proposed systems were implemented in Keras/Tensorflow using the Python programming language, and Scikit learn libraries. To measure the performance metrics, experiments are executed on a workstation that has the properties. Proposed systems were executed on the Multicore structure of the NVIDIA Ge Forcer GTX 1080 Ti Graphic card, whose specifications are detailed in Table 10. To calculate the performance measure of the proposed systems; Accuracy, Precision, Recall, F1-Score and Error Rate values are used



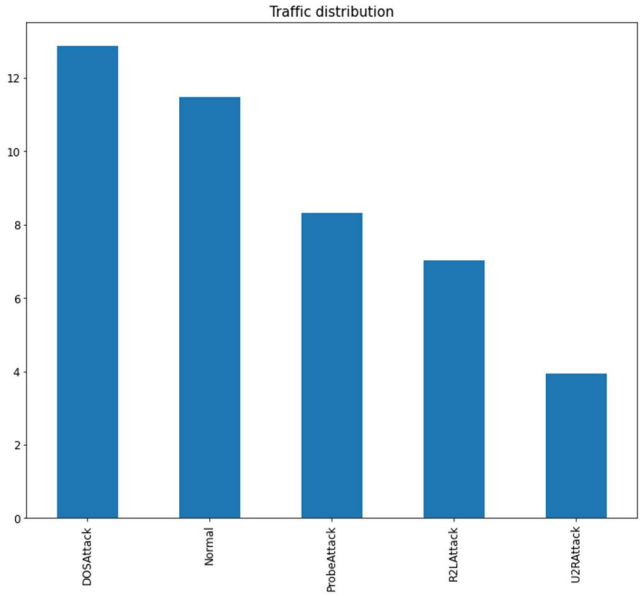
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Use Case Diagram



V. PROJECT IMPLEMENTATION

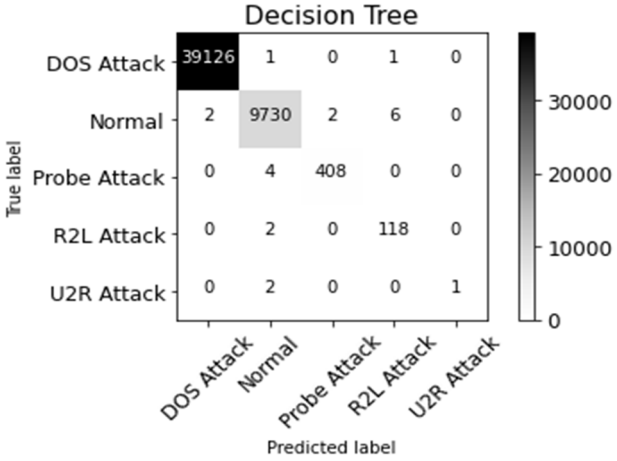
The selected dataset is also imbalanced. Therefore, to increase the efficiency of the system depending on attack types and to decrease missed intrusions and false alarms, the imbalance ratio is reduced by using a synthetic data generation model called Synthetic Minority Oversampling Technique. Data generation is performed for minor classes, and their numbers are increased to the average data size via this technique. Experimental results demonstrated that the proposed approach considerably increases the detection rate for rarely encountered intrusions. An online oversampling Principal Component Analysis designed to address the anomaly detection problem is proposed in.



Dig:- Traffic in Networks

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Their approach focuses on using online platforms for large-scale problems. By oversampling the minority class of the target instance, their proposed algorithm allows them to determine the anomaly of the target instance. However, if the dataset is imbalanced and a specific category composes the most significant part of the dataset, then the use of accuracy as a single metric is not much acceptable. If there is a large gap between the data size within the majority and minority categories, sophisticated attackers can focus on minority attack types to increase their efficiency.



Dig:- predicted label

VI. ADVANTAGES AND DISADVANTAGE

Advantages:

- It takes advantage of its simple structure to process large amounts of data quickly. In some cases, more complex trees have to deal with the classification of datasets.
- In such cases, decision trees become more complex, and it becomes more difficult to reach any of the goals. Over fitting is another problem in decision tree algorithms. Some of the leaf nodes are pruned out of the decision tree to solve this problem.
- The advantages of AIDSs, most current IDSs either directly use an AIDS or benefit from it within a hybrid approach. These IDSs need to be trained via machine learning model by processing the dataset. Most of the works on this topic adopted old datasets, which contain redundant information and imbalanced volumes of data types.

Disadvantages:

- This method is highly resistant to simple and noisy training data. As such, it has the disadvantage of requiring a lot of memory space because it stores all the cases in distance calculations.
- Over fitting is another problem in decision tree algorithms. Some of the leaf nodes are pruned out of the decision tree to solve this problem. Entropy and information gain should be calculated for decision trees.
- It is seen that hyper parameters of machine learning algorithms were set as default in the literature. Therefore, these values were left as default values in the study so that comparisons with other studies can be made.

VII. CONCLUSION

However, these minority classes are generally positive classes. Therefore, the imbalance ratio should be decreased to increase the efficiency of the system and to decrease its average accuracy. In this paper, six different MACHINE LEARNING models (Decision Tree, Random Forest, K Nearest Neighbor, Adaboost, Gradient Boosting, and Linear

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Discriminant Analysis) were implemented using a recent dataset (CSE-CIC-MODEL2018). O decrease the imbalance-ratio, a data sampling model was used by increasing the data size of the minority groups. The experimental results showed that the implemented models have a very good accuracy level when compared with recent literature. The use of a sampled dataset caused the average accuracy of the models to increase between 4.01% and 30.59%.

ACKNOWLEDGMENT

This paper is supported by many people, some of whom have a direct role and some of them have an indirect way by publishing their research online which helped to understand this concept easily. We express our deepest gratitude, sincere thanks and deep feeling of appreciation to our Project Guide Professor Sagar Dhanake, his presence at any time throughout the Semester, important guidance, opinion, comment, critics, encouragement, and support greatly improved this project work. We thank the college administration for providing the necessary infrastructure and technical support. Finally, we extend our heartfelt thanks to our friends and family members.

REFERENCES

[1] J. M. Johnson and T. M. Khoshgoftaar, ``Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, p. 27, 2019.
[2] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, ``An adaptive ensemble MACHINE LEARNING model for intrusion detection," IEEE Access, vol. 7, pp. 82512_82521, 2019.
[3] R. Abdulhammed, M. Faezipour, Abumallouh, ``Deep and MACHINE LEARNING approaches for anomaly-based intrusion detection of imbalanced N/W traffic," IEEE Sens. Lett., vol. 3, no. 1, pp. 1_4, Jan. 2019.
[4].Mohammed Yasin Jisan, and M. M. Rahman, ``N/W intrusion detection using supervised MACHINE LEARNING technique with feature selection," in Proc. Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST), Jan. 2019, pp. 643_646.
[5] A. I. Al-issa, M. Al-Akhras, M. S. Alsahli, and M. Alawairdhi, ``Using MACHINE LEARNING to detect DoS attacks in wireless sensor N/Ws," in Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT), Apr. 2019, pp. 107_112.
[6] E. Kurniawan, F. Nhita, A. Aditsania, and D. Saepudin, ``C5.0 algo. and synthetic minority oversampling technique for rainfall forecasting in Bandung regency," in Proc. 7th Int. Conf. Inf. Commun. Technol. (ICoICT), Jul. 2019, pp. 1_5.