International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

EXPLAINABLE AI IN INTRUSION DETECTION SYSTEMS

Nikita Nerkar[1], Atharva Nile[2], Amit Kulkarni[3], Onkar Kadlag[4], Prashant Gadakh[5]

Abstract: As the use of internet in increasing day by day and the chances of system get compromised due to various types of attacks has increased. Intruders are finding new techniques to compromise the system. The concern about the cyber security is growing and for the user most of the model is perceived as a black box. There is need of finding the attack correctly and then proper reports should be generated to show how the system got compromised. So we are proposing a system where Intrusion Detection System (IDS) can detect the attack and Explainable artificial intelligence tell us about what type of attack is being performed on the system. Intrusion Detection System keeps track of the malicious packets entering in the system. Explainable Artificial Intelligence will show the report on which type of attack took place. In the proposed system we have use the NSL-KDD dataset for classification of attack detected by our proposed Intrusion Detection System.

Keywords: Intrusion detection System, Explainable artificial intelligence, NSL-KDD, classification.

I. INTRODUCTION

Today internet is been used in vast number of areas like organizations, businesses, entertainment industry, personal day to day activities etc. One of the most important issues nowadays is security. When an intrusion takes place the security of the system is compromised. The assumption of the behaviour of the intrusion is different from the legal user in the system. To deal with intrusions in the network is the main aim of the IDS.

Explainable Artificial Intelligence presents the results of the solution that can be understood by the system administrators effectively. Our system consists of four classes of the intrusion like User to Root (U2R), Denial of Service (DoS), and Remote to User (R2U), Probing attacks. An IDS alone is only able to detect that attack has taken place and alarms admins but it is not able to detect type of attack. Using Explainable Artificial Intelligence when can detect which type of attack has took place from the four classes and generate the reports for the same. For Classification purpose we have used NSL-KDD dataset for training a model which classifies the attack. Pre-processing over the model is done by using one hot encoding, label encoding and standard scalar techniques.

II. LITERATURE SURVEY

We have referred “Intrusion Detection System Using Data Mining Technique: Support Vector Machine” [4] In which they have Classified the attack done on the system using support vector machine(SVM) method and using the

[1, 2, 3, 4 S] *Researcher Department of Computer Engineering, International Institute of Information Technology, Hinjewadi, Pune,India.*
[2] *Professor, Department of Computer Engineering, International Institute of Information Technology, Hinjewadi, Pune,India.*

Received: 22 Dec 2019 | Revised: 13 Jan 2020 | Accepted: 05 Feb 2020

NSL-KDD Cup'99 dataset. They have also reduced the time which is required to build the SVM model. They have mentioned the types of attacked in there research like Denial Of Service, User to Route, Probe and Remote to Local. They have used Gaussian RBF kernel of SVM along with the 10 fold cross validation .77.07 seconds is there accuracy to build the model. 94.187 % is the accuracy to detect the attack. They changed their method to 10 fold cross validation to increase the accuracy which is 98.5749%.

Then we referred "An Adversarial Approach for Explainable AI in Intrusion Detection System" [1].In this paper they have given Human understandable clear explanation of misclassification of attacks by Intrusion detection System. They have tried to correctly classify the attack by doing some minor chances. They have given the visual representation of the feature that misclassified the attack. Adversarial approach of Explainable AI is been used for the modification purpose.

Another paper referred is "Importance of Intrusion Detection System (IDS)" [3].In there research they have mentioned the importance of IDS and what IDS is? They said that IDS is one of the effective methods to protect our system. Misuse detection and anomaly detection are two types of Intrusion detection System they are used to detect the attack and generate the alert. Intrusion detection System is only use to detect the attack to prevent the attack. It is a detective system not preventive. IDS become accurate when it generates very less false positive alarms.

We referred "Intrusion Detection System Using Data Mining Technique: Support Vector Machine" in which they have outlier Detections which is one of the new approach to detect the intrusion. They have tested their system with the real word data as well.

III. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable AI (XAI) is artificial intelligence methodology that average humans can easy understand basically use for analysis purpose and decision making.

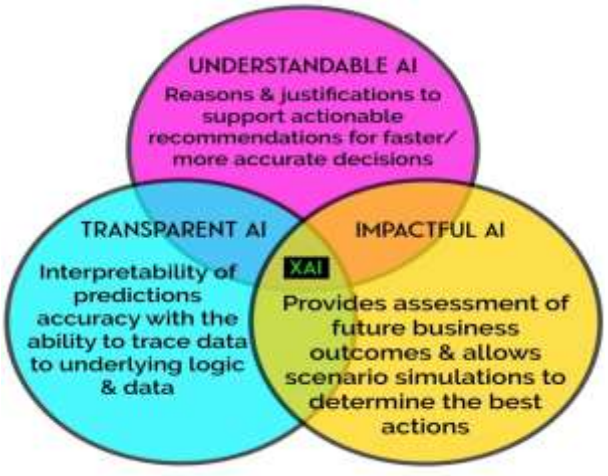

Fig 1: Explainable Artificial Intelligence
Source: https://www.kdnuggets.com/2019/01/explainable-ai.html

The main aim of Explainable AI (XAI) program is that:

- To produce more explainable models, it will give clear explanation about the attack on the system.
- High level of learning performance is maintained.
- Human understandable explanation is given.

XAI brings transparency to the forefront of business deciding, unlocking the ability of AI and machine learning, so delivering unjust insights that unlock truth business worth and support people, so they can make better, faster, more accurate decisions.

Advantages of Explainable AI (XAI):

- Improved explainability and transparency.
- Faster adoption.
- Improved debugging.
- Enabling auditing for regulatory requirements.

IV. DETECTION SYSTEMS (IDS) & ATTACK CATEGORIES

Intrusion detection system is the guard for our system or network who keeps continues watch's on the system to check if any anemology is taking place by the intruder and generate the alarm if any intrusion happens.

Advantages of IDS:-

- System is continually monitored for any attack.
- Effective detection of any injury to the system or network.

Disadvantage of IDS:-

- IDS cannot prevent the attack; it can only detect the attack.

There are four categories [4], of attacks such as Dos (Denial of Service), Probing, R2L (Remote to Local) and U2R (User to Route).The other fifth one is Normal which is not an attack.

Attack class	Attack type
Dos	back, land, neptune, pod, smurf, teardrop
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
U2R	buffer_overflow, loadmodule, perl, rootkit
Probe	ipsweep, nmap, portsweep, satan

Fig 2: Attack types

Source: https://www.researchgate.net/figure/Attack-types-of-DoS-R2L-U2R-Probe-categories_tbl1_327110465

Dos basically include back, land, smurf, Teardrop, Udpstorm etc. Probing include attacks such as Lpsweep, Mscan, namp, Portsweep, saint, satan. Remote to local includes Xsnoop, Multihop, Snmpgetattack, Warezmaster, Worm etc. User to route includes Loadmodule, Sqlattack, Buffer_overflow, Rootkit etc.

Types of attacks:

Denial-of-service (DoS):- Legitimate users are unable to access or over load the server by giving false requests.

Probe: Probing is web application based attack where attacker scans the complete web site infrastructure as well as web sites structure.

Remote to Local (R2L): Through unauthorised it enters the victim machine is remote 2 local attacks.

User to Route (L2R): when legally accessing a local machine, for illegally obtaining the root's privileges.

V. DATA SET COLLECTION AND PREPROCESSING

To train our model we have used NSL KDD dataset which has been collected from official github repository of defcom. KDD '99 dataset,[4] is considered as a standard benchmark dataset in intrusion detection field. NSL KDD dataset is upgraded version of this KDD '99 dataset that has eliminated redundant and duplicate records. Training dataset has 1, 25,973 instances while testing dataset has 22,544 instances.

Dataset Pre-Processing -:

Protocol type, service and flag are three columns need to be pre-processed in our approach. One-hot encoding and label encoding techniques were used to perform the process.

ABLE I. " NSD- KDD CUP'99 Dataset Features"

Sr.no	Feature Name
1	_Duration_
2.	_Protocol__type_
3.	_Service_
4.	_Flag_
5.	_Src_bytes_
6.	_Dst_bytes_
7.	_Land_
8.	_Hot_
9.	_Urgent_
10.	_Num_Failed_logins_
11.	_Logged_in_
12.	_Num_compromised_
13.	_Root_shells_
14.	_Wrong_Fragement_
15.	_Su_attempted_
16.	_Num_root_
17.	_Num_file_Creations_

Received: 22 Dec 2019 | Revised: 13 Jan 2020 | Accepted: 05 Feb 2020

18.	_Num_Shells_
19.	__Num_access_files
20.	_Num_outbound_cmds_
21.	_Is_host_Login_
22.	_Is_guest_Login_
23.	_Count_
24.	_Srv_count_
25.	_Serror_rate_
26.	_Srv_Serror_Rate_
27.	_Rerror_rate_
28.	_Srv_rerror_rate_
29.	_Same_srv_rate_
30.	_Diff_srv_rate_
31.	_Srv_diff_host_rate_
32.	_Dst_Host_Count_
33.	__Dst_Host_Srv_Count_
34.	_Dst_Host_same_srv_rate_
35.	_Dst_Host_diff_srv_rate_
36.	_Dst_Host_same_srv_port_rate_
37.	_Dst_Host_diff_srv_host_rate_
38.	Dst_Host_serror_rate_
39.	Dst_Host_srv_serror_rate_
40.	Dst_Host_rerror_rate_
41.	Dst_Host_srv_rerror_rate_
42.	Attack type

Source: Intrusion Detection System Using Data Mining Technique: Support Vector Machine [4].

VI. SQLIA MODEL

To get rid of the SQL infection attacks a defense model has been proposed which helps in preventing the SQL injection attacks. The model is a multi-tasking model. This model is used to validate input and also specify information based on web address, which is relatively important too, especially for detecting the sensible characters. IP address reliability gets verified by the server side. User gets rejected from logging in if the input values are found unreal or illegal. After IP address verification input values are tested by the server side using various parameters such as length, format, range, type. If the input string matches with that of SQL rules then only the user is allowed to access the web page. After that the server side verifies the privilege of the user. If access permissions gets exceed for a particular user, results in blocking of user and message is been send to the system administrator by the system. When all the verifications performed by the server side are incorrect then at such situation the injection attack is recorded by the

server [8].

VII. ALGORTIHAM

- Start.
- Read dataset.
- Give column names to datasets.
- Find unique categories which will be used further for one-hot encoding.
- We have to make dummies for all the categories that have datatype as object (viz. protocol_type, service and flag).
- Test set and gather set has fewer categories than training set.
- For label encoding, insert categorical features in 2D numpy array.
- Make column names for dummies of training set, testing set and gathered set and add them up to get total dummy columns for every set.
- Perform label encoding.
- Perform one-hot encoding over label encoded sets.
- Add new categories in test set and gather set since they had fewer categories than training set.
- Join categorical data and non-categorical data i.e. join encoded dataframes to original dataset.
- Rename every attack by classifying it in the category of normal, Denial of Service (DoS), Probing, User to Root (U2R), Remote to Local (R2L).
- Put these new labels in label column of training dataset.
- Split training dataframe in four parts viz. DoSdataframe, Probing dataframe, Remote to Local dataframe and User to Root dataframe.
- Split every dataframe as dataframe related to features anddataframe related to labels i.e. outcone variable.
- Perform standard scaling.
- Feature selection using f-test was performed but it didn't increase accuracy so this step is omitted.
- Four different models of Decision tree were built.
- Learning with noisy labels was performed using cleanlab class to predict labels of gathered dataset.
- Print result of prediction made using cleanable class.
- Cross validation is performed.
- Print confusion matrix for all four models.
- Print accuracy of all four models.

Received: 22 Dec 2019 | Revised: 13 Jan 2020 | Accepted: 05 Feb 2020

➢ End.

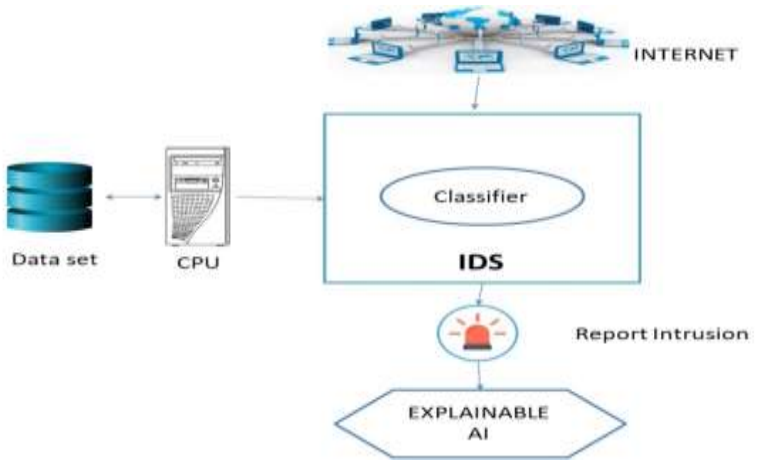

Fig 3: Overview of Proposed System

The system is based on using the XAI approach in IDS:

- With the help of Artificial Intelligence, Intrusion detection system will become smart.
- The security experts can identify the type of attacks or intrusions on the system with the help of XAI the part of AI would be able to detect the type of attack based on the results of Intrusion detection system.
- The results from the IDS would be analyzed by highly trained AI model, and then a analysis report will be made based on that.

The report will be studied by the security experts to analyze the efficiency of the system and to improve it.

VIII. Attacker Injecting SQL Queries

Prediction of attack results -:

- DoS -:

[1 1 0 0 0 0 0 0 0 0 0 0]

- Probe -:

[1 1 0 1 1 0 1 1 1 1 1 1]

- R2L -:

[0 0 0 0 0 0 0 0 0 0 0 0]

- U2R -:

[0 0 0 0 0 0 0 0 0 0 0 0]

CONFUSION MATRIX -:

- DOS -:

Predicted attacks Actual attacks	0	1
0	9496	215
1	2418	5042

Accuracy Score: 0.99639 (+/- 0.00341)

- PROBE

Predicted attacks Actual attacks	0	1
0	1838	7873
1	165	2256

Accuracy Score: 0.99571 (+/- 0.00328)

- R2L

Predicted attacks Actual attacks	0	1
0	9703	8
1	2704	181

Accuracy Score: 0.97920 (+/- 0.01053)

- U2R

Predicted attacks Actual attacks	0	1
0	9705	6
1	49	18

Accuracy Score: 0.99652 (+/- 0.00228)

VIII. CONCLUSION

The security of the system is major concern for many organizations. There is great demand to develop a system that can detect the intrusion and maintain security and privacy of the system. We have developed a system with 10 fold cross validation using decision tress. Here the Build model classifies the attack from the attack category using one hot encoding and label encoding.

10 fold cross validation on NSL KDD dataset is carried out for experiment. Accuracy score for detecting the

attacks as a PROBE is 99.57%.

Accuracy score for detecting the attacks as a R2L is 97.92%. Accuracy score for detecting the attacks as a U2R is 99.65% and DOS is 99.63%.Heighest accuracy score of attacks detect is of DoS which is 99.65%.

REFERENCE:

[1] Daniel L. Marino, ChathurikaS.Wickramasinghe, Milos Manic, “An Adversarial Approach for Explainable AI in Intrusion Detection System” Department of Computer Science Virginia Commonwealth University Richmond, USA.

[2] JABEZJa, Dr.B.MUTHUKUMAR, “Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach” International Conference on Intelligent Computing, Communication & Convergence.

[3] Asmaa Shaker Ashoor, Prof. Sharad Gore, “Importance of Intrusion Detection System (IDS)” International Journal of Scientific Engineering Research.

[4] Yogita B. Bhavsar1 ,Kalyani C.Waghmare2, “Intrusion Detection System Using Data Mining Technique: Support Vector Machine,” International Journal of Emerging Technology and Advanced Engineering, March 2013.

[5] GulshanKumar ,Krishan Kumar ,Monika Sachdeva" The use of artificial intelligence based techniques for intrusion detection: a review" 4 September 2010.

[6] UdayaSampath K. PereraMiriyaThanthrige, JagathSamarabandu, Xianbin Wang, “Machine Learning Techniques for Intrusion Detection on Public Dataset,” 2016.

[7] James P. Anderson, “Computer Security Threat Monitoring and Surveillance,” Technical report, James P. Anderson Co., Fort Washington, Pennsylvania. April 1980.

[8]Tomas Abraham, “IDDM: INTRUSION Detection using Data Mining Techniques” Technical report DTSO electronics and surveillance research laboratory, Salisbury.

[9] Wenke Lee and Salvatore J. Stolfo, "A Framework for constructing features and models for intrusion detection systems” ACM transactions on Information and system security (TISSEC), vol.3, Issue 4, Nov 2000.