IJARST

International Journal For Advanced Research In Science & Technology

A peer reviewed international journal

www.ijarst.in

ISSN: 2457-0362

Detection of Cyber Attack in Network using Machine Learning Techniques

1. G. Lavanya, ASSISTANT PROFESSOR, DEPARTMENT of CSE, rapyaka.lavanya@gmail.com
2. B Divya, BTech, Department of CSE, (197R1A0565) 197r1A0565@gmail.com
3. B Arun Kumar, BTech, Department of CSE, (197R1A0567) 197R1A0567@cmrtc.ac.in
4. G. Akshay Kumar, BTech , Department of CSE, (197R1A0577) 197R1A0577@cmrtc.ac.in

Abstract : Computer and communication technology advancements, in contrast to the past, have brought about extensive and rapid changes. The implementation of novel concepts benefits individuals, organizations, and governments greatly; However, some people have prejudice against them. For instance, major data protection, the safety of stored information stages, accessibility to information, and so forth. Perhaps of the most major problem in this day and age is advanced apprehension based abuse, as per these concerns. Digital dread, which has raised a number of concerns for individuals and organizations, has reached the point where various groups, including criminal organizations, skilled individuals, and digital activists, may use it to harm open and national security. Considering this, Intrusion Detection Systems, or IDS, were made to avoid advanced assaults. In light of the new CICIDS2017 dataset, support vector machine (SVM) calculations were utilized to identify port range drives with isolated precision paces of 97.80% and 69.79%. We may introduce random forest, CNN, and ANN as alternatives to SVM, with accuracies of 93.29 for SVM, 63.52 for CNN, 99.93 for Random Forest, and 99.11 for ANN.

Index Terms : *SVM, CNN, ANN, Random forest, and the intrusion detection system (IDS).*

1. INTRODUCTION

The implementation of novel concepts benefits individuals, organizations, and governments greatly; However, some people have prejudice against them. For instance, major data protection, the safety of stored information stages, accessibility to information, and so forth. Perhaps of the most major problem in this day and age is computerized dread based mistreatment, as per these concerns. Digital dread, which has raised a number of concerns for individuals and organizations, has reached the point where various groups, including criminal organizations, skilled individuals, and digital activists, may use it to harm open and national security. Considering this, Intrusion Detection Systems, or IDS, were made to avoid advanced assaults.

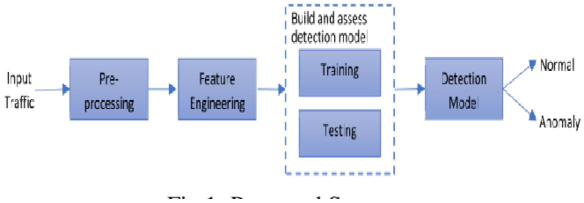


Fig 1 Example Figure

Political and monetary entertainers are progressively utilizing refined cyberwarfare to upset, obliterate, or stifle data content in PC organizations. It is important to give reliability against attacks by strong aggressors who could in fact control a subset of the members in the organization while creating network conventions. Both aloof (listening in, nonparticipation) and forceful (e.g., sticking, message dropping, debasement, and fashioning) assaults can be sent off by the controlled gatherings. The act of progressively checking occasions that happen in a PC framework or organization, examining them for indications of potential events, and much of the time forestalling unapproved access is known as interruption recognition. More often than not, this is finished via naturally gathering information from different frameworks and organization sources and afterward examining the information to search for potential security issues. With regards to appropriately shielding organizations and frameworks from progressively complex assaults like disavowal of administration, customary interruption discovery and counteraction arrangements like firewalls, access control systems, and encryption experience the ill effects of various critical inadequacies. Moreover, most of frameworks in light of such methodologies have high discovery rates for both bogus up-sides and misleading negatives, as well as an absence of progressing transformation to moving malevolence. Notwithstanding, throughout the beyond a decade, an extraordinary number of Machine Learning (ML) procedures have been applied to the issue of interruption identification with expectations of expanding flexibility and discovery rates. These strategies are much of the time used to stay up with the latest and far reaching. The protection against different cyberattacks and network safety have as of late become famous subjects. The fundamental legitimization behind this is the fast advancement of PC associations and the gigantic number of significant applications involved by people or relationship for individual or business purposes, particularly beginning from the introduction of the Internet of Things (IoT). In enormous organizations, cyberattacks inflict damage and monetary misfortunes.

2. LITEARTURE SURVEY

R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

Assailants every now and again utilize port checking to find benefits that they could take advantage of to get close enough to PCs. Administrations that pay attention to both notable and less notable ports are controlled by all frameworks that interface with a LAN or the Web by means of a modem. Port checking might give the assailant data about the frameworks being focused on: what administrations are running, who possesses them, whether mysterious logins are upheld, and whether some organization administrations require validation. Port checking can be achieved by making an impression on each port separately. The kind of reaction that is returned demonstrates whether the port is being used and can be examined for extra defects. Network security experts benefit from port scanners since they might uncover potential security blemishes on the designated framework. With the right apparatuses, port outputs can be run on your frameworks, however they can likewise be identified and how much data about open administrations can be controlled. There are open ports on each framework that can be gotten to by the general population. The goal is to deny approved clients admittance to shut ports while restricting admittance to open ports.

S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans,"

Portscanning is an important and common practice. Computer hackers frequently use it to identify networks or sites they suspect of hostile behavior. Framework heads and other organization defenders may thusly perceive portscans as potential signs of a more serious assault. Network protectors additionally every now and again use it to dissect and track down weaknesses in their own organizations. As a result of this, interlopers are anxious to determine whether the protectors of an organization are continually portscanning it. Nonetheless, while aggressors would need to hide their portscanning, safeguards are probably not going to do as such. All through the rest of this examination, we will allude to aggressors filtering the organization and protectors endeavoring to distinguish the output for the good of lucidity. On Web newsgroups and mailing records, portscanning is the subject of various legitimate and moral discussions. Whether or not port filtering of far off networks without the consent of their proprietors is a lawful and moral demonstration all by itself is one of the issues being examined. The majority of jurisdictions are currently unaware of this. However, our practical experience with following up on unsolicited remote portscans has demonstrated that almost all of them originate from compromised hosts and are extremely hostile. Therefore, we believe it is acceptable to report a portscan to the network administrators of the distant network from which it originated as potentially hostile. On the other hand, this work is focused on the technical aspects of detecting portscans, which are unrelated to how one views their significance or reacts to them. In addition, our primary focus is on the process of utilizing a network intrusion detection system (NIDS) to identify a portscan. While adhering to a realistic approach for use on busy networks, we make an effort to take into account some of the more obvious ways that an attacker might avoid detection. In the remainder of this section, we define portscanning, provide several examples, and describe the covert tactics that an attacker might use. The previous work on portscan detection is discussed in the following section. After that, we go over the approaches we intend to employ and provide some preliminary data to back up our strategy. Last but not least, we talk about potential additions to this work and potential applications. We assume that the reader is familiar with basic probability theory, information theory, and linear algebra, as well as Internet protocols, fundamental concepts of network intrusion detection and scanning. One of two reasons an attacker might conduct a portscan is: both the primary and secondary one. The fundamental objective is to gather data on the reachability and status of specific IP address and port blends (either TCP or UDP). (Although the principles can easily be applied to ICMP scans, we do not explicitly mention them in this study.) The other goal is to send a lot of notifications to intrusion detection systems to make it hard for network defenders to do their jobs or discourage them from doing so. Since it is simple to identify flood portscans, this work will focus on identifying information gathering portscans. However, one of the most important aspects of the design of our algorithm will be the possibility of being maliciously overwhelmed with information. The collection of port/IP combinations that the attacker intends to characterize is referred to as a scan footprint. Conceptually distinguishing the scan's footprint from its script, which describes the time sequence during which the attacker attempts to investigate the footprint, is helpful. Scan speed, randomness, and other script parameters have no effect on the footprint. The attacker creates a scan script based on the footprint's information gathering requirements for her scan and possibly additional non-information gathering goals (such as not being detected by an NIDS). A horizontal scan is currently the most common type of portscan footprint. This indicates that an attacker is looking for hosts that expose a particular service and has an exploit for that service. Consequently, she checks the port of interest on all IP addresses within a particular interest range. Currently, most of this is done sequentially on TCP port 53. DNS)

M. C. Raja and M. M. A. Rabbani, “Combined analysis of support vector machine and principle component analysis for ids,”

Security of networked systems is now a major global issue that affects individuals, businesses, and governments more than it was in the past. Assaults on arranged frameworks have expanded emphatically, and the techniques utilized by aggressors are continually evolving. For example, the security of information stockpiling frameworks, the protection of delicate data, the accessibility of ability, etc. Cyber terrorism is one of today's most pressing issues, according to these worries. Various parties, including criminal organizations, professionals, and cyber activists, have advanced cyber terror to the point where it may threaten public and national security. Cyber terror has caused numerous problems for individuals and institutions. Intrusion detection is one way to stop these attacks. For the improvement of intrusion detection systems (IDS), ML is a savvy and powerful methodology. The new CICIDS2017 dataset was utilized in this review to recognize port output endeavors utilizing profound learning and backing vector machine (SVM) methods. Presentation An organization IDS is either an equipment gadget or a product based program that distinguishes malignant organization action [1,2]. Interruption identification can be named irregularity based or signature-based, contingent upon the technique used to identify the interruption. The developers of IDS employ a variety of strategies for intrusion detection. Information security is the process of protecting information from being accessed, used, disclosed, destroyed, altered, or damaged in an illegal manner. There is a tendency to use the terms "information insurance," "computer security," and "information security" interchangeably. These spaces are interconnected and share the motivation behind guaranteeing data accessibility, privacy, and trustworthiness. As per research, the principal stage in an attack is disclosure [1]. At this point, information about the system is gathered through reconnaissance. A system's list of open ports could be very useful to an attacker. As a result, antivirus and intrusion detection systems are just two of the many tools available to discover open ports [2]. One of these plans is to use machine learning. Predicting and detecting threats before they cause serious security issues is possible with machine learning (ML) methods [3]. Binary classification is the process of dividing cases into two groups. Then again, the expression "multi-class characterization" alludes to grouping events into at least three classes. In this review, we utilize the two characterizations. Data security is the most common way of shielding data from being gotten to, utilized, uncovered, obliterated, modified, or harmed in an unlawful way. There is a propensity to utilize the expressions "data protection," "PC security," and "data security" conversely. The purpose of ensuring information availability, confidentiality, and integrity is shared by these domains, which are interconnected. As per research, the principal stage in an attack is revelation [1]. At this point, information about the system is gathered through reconnaissance. A framework's rundown of open ports could be exceptionally valuable to an aggressor. Hence, there are a couple of systems to find open ports [2], for instance, antivirus and interference revelation structures.

I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization."

It is turning out to be progressively clear that the potential mischief that can be brought about by sending off assaults is fundamentally expanding with the remarkable development of PC organizations and created applications. Meanwhile, interference area systems and interference aversion structures are essential security weapons against complex and consistently extending network risks. An absence of a reasonable dataset causes irregularity based procedures in interruption location frameworks to send, examine, and assess erroneously. Specialists have utilized an assortment of such datasets, including DARPA98, KDD99, ISC2012, and ADFA13, to assess the viability of their proposed interruption identification and interruption counteraction frameworks. Our examination of eleven openly accessible datasets from 1998 uncovered that various them are obsolete and unusable. A portion of these datasets need include set and metadata, others anonymize bundle data and payload, making it hard to address latest things, and others need traffic variety and volume. Still others don't cover many dangers. A dependable dataset that follows true norms and is made accessible to the general population contains seven successive assault network streams as well as harmless streams. To decide the best arrangement of highlights for recognizing explicit kinds of assaults, the article assesses the viability of an extensive variety of organization traffic qualities and ML strategies.

3. METHODOLOGY

On the KDD99 dataset, Almansob and Lomte [9] used Principal Component Analysis (PCA) and Faultless Bayes. For IDS, Chithik and Rabbani used PCA, SVM, and KDD99 likewise [10]. In Aljawarneh et alpaper, . The NSL-KDD dataset filled in as the reason for their IDS model's assessment and assessment [11]. Plan inspects reveal that the KDD99 dataset is dependably utilized for IDS [6-10]. There are 41 features in KDD99, which was made in 1999. Therefore, KDD99 is obsolete and needs data on state of the art new assault types like multi-day abuses and others. Subsequently, we utilized a fresh out of the plastic new and state of the art CICIDS2017 dataset [12] in our examination.

Drawbacks:

1) Stringent requirements
2) More difficult to use for non-technical users
3) Limited resources
4) Requires constant patching

Important algorithm phases that are constantly attacked are listed below. 1) Normalization is required for each dataset. 2) Make training and testing datasets from the dataset. 3) Utilize the RF, ANN, CNN, and SVM algorithms to construct IDS models. 4) Determine how each model performs.

Benefits:

1) Insurance against antagonistic organization attacks.
2) Getting rid of harmful components in an existing network and/or making sure they are safe.
3) Prevents users from illegally accessing the

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

network.

4) Prevent applications from gaining access to resources that could be contaminated.

5) Keeping confidential information safe

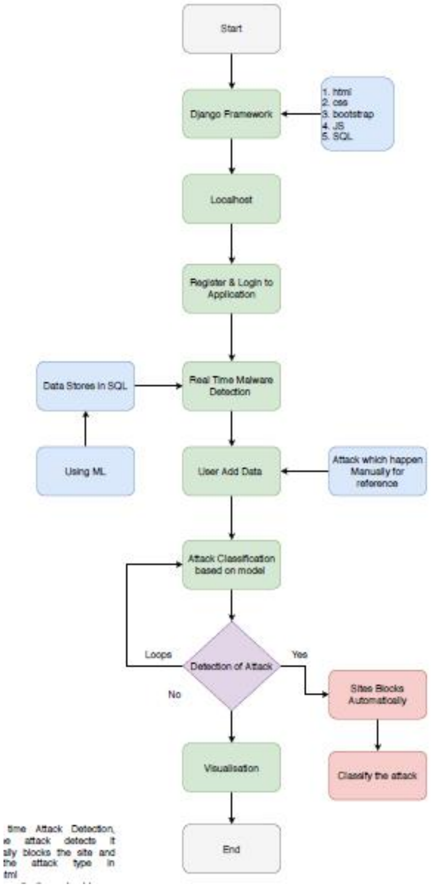


Fig 2 Proposed Flow Chart

Dengue data:

The UCI Machine Learning Repository serves as the source of the data set. There are two datasets in it: characteristics and labels of dengue.

There are 1869 records for dengue qualities and 1456 records for dengue marks in the assortment.

ndvi ne, ndvi nw, ndvi se, ndvi sw, ndvi se, ndvi s Sj, 2008, 18, 29-4-2008, - 0.0189, - 0.0189, o. 1027286, 0.0912, 78.6, 298.4928571, 298.55, 294.5271429, 25.37, 78.78142857, 26.52857143 The dataset segment names are shown previously

The dataset segment names are shown over all strong names, and the dataset values are shown underneath all qualities in the dengue lables dataset. Absolute cases, city, year, and seven day stretch of year Sj, 1990, 18, 4 Sj, 1990, 19, 5

4. IMPLEMENTATION

Algorithms:

ANN:

A artificial intelligence subfield that is designed according to the cerebrum is alluded to as an "artificial neural network." This subfield was roused by science. A PC network called a ANN depends on the natural brain networks that make up the design of the human mind. Neurons in counterfeit brain networks are associated with each other at different levels of the organization, very much like in the human cerebrum. Hubs are the names given to these neurons. Each part of the ANN is shrouded in this illustration. This illustration will discuss ANNs, versatile reverberation hypothesis, the Kohonen self-arranging map, building blocks, solo learning, and hereditary calculations, in addition to other things.

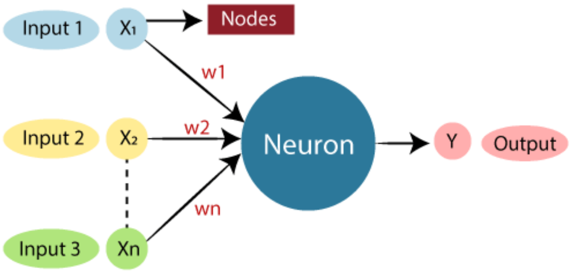


Fig 3 ANN

The architecture of an artificial neural network:

We should initially grasp what a brain network is before we can understand the idea of planning ANN. An enormous number of fake neurons, or units, are organized in a progression of layers to characterize a brain organization. We should investigate the different sorts of layers that a ANN could have.

There are three layers that make up ANN:

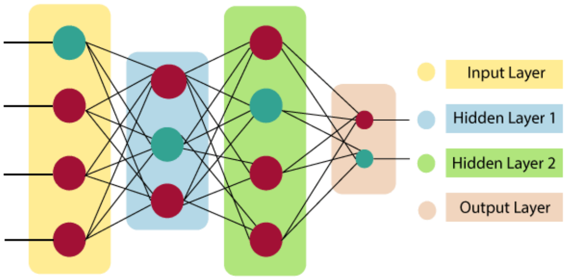


Fig 4 ANN Architecture

Layer of Input: It receives inputs in a variety of formats that the programmer specifies, as the name suggests.

In the middle of between the info and result layers is the secret layer. It does all the math to track down designs and secret attributes.

Layer of Result: The info goes through a progression of changes in the secret layer, bringing about yield given by this layer. The inclination and weighted absolute of the sources of info are processed by the fake brain network after it gets input. This computation is communicated utilizing an exchange capability.

$$\sum_{i=1}^{n} Wi * Xi + b$$

In the wake of sorting out the weighted aggregate, it takes care of that number into an actuation capability, which then creates the outcome. A hub's terminating state is chosen by initiation capabilities. Just the individuals who have been terminated come to the result layer. Contingent upon the sort of work we are doing, we can utilize any of various actuation capabilities.

CNN:

CNN is a deep learning model for handling information with a matrix design, similar to photos. It is intended to consequently and adaptively learn spatial orders of attributes, from low-level to undeniable level examples, propelled by the design of creature visual cortex.

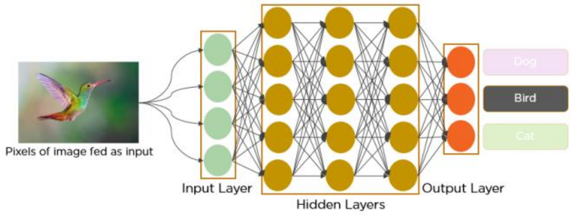


Fig 5 CNN

Random Forest:

A notable managed learning ML calculation is Random Forest. In ML, taking care of issues with characterization and regression can be utilized. It depends on the possibility of group learning, in which various classifiers are consolidated to take care of a confounded issue and work on the model's exhibition. As the name proposes, Random Forest is a classifier that utilizes the normal of various choice trees on different subsets of the dataset to work on the dataset's anticipated precision. The Random Forest gathers the gauges from each tree and predicts the last result in view of most of expectations, as opposed to depending on a solitary choice tree. The accuracy is better and the gamble of overfitting is bring down the more trees there are in a woodland.

The picture underneath portrays the Random Forest strategy:

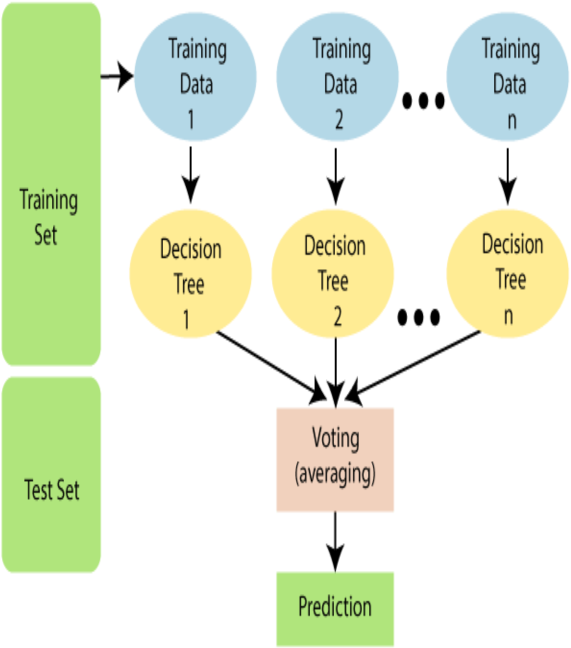


Fig 6 Random forest

There are two phases in how Random Forest functions: The initial step is to blend N choice trees to make the arbitrary timberland, and the subsequent step is to make expectations for each tree that is made in the primary stage. The functioning technique is portrayed in the means and realistic beneath:

Stage 1: Pick K irregular data of interest from the preparation set.

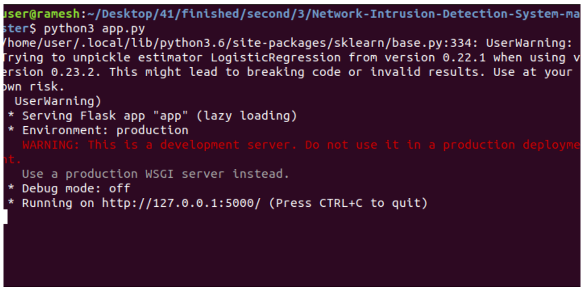
Stage 2: For the predetermined data of interest, make choice trees (subsets).

Stage 3: The number N addresses the quantity of choice trees you need to make.

Stage 4: Reiteration of stages 1 and 2

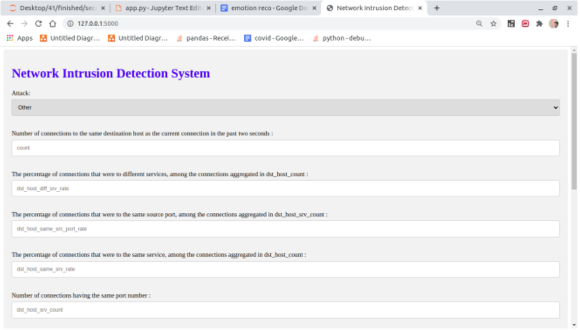
Stage 5: Dole out the shiny new information focuses to the class that got the most votes in the wake of finding the conjectures for every choice tree.

5. EXPERIMENTAL RESULTS



```
user@ramesh:~/Desktop/41/finished/second/3/Network-Intrusion-Detection-System-ma
ster$ python3 app.py
/home/user/.local/lib/python3.6/site-packages/sklearn/base.py:334: UserWarning:
Trying to unpickle estimator LogisticRegression from version 0.22.1 when using v
ersion 0.23.2. This might lead to breaking code or invalid results. Use at your
own risk.
  UserWarning)
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployme
nt.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig 7 Output Screen



Network Intrusion Detection System

Attack:

Other

Number of connections to the same destination host as the current connection in the past two seconds :

count

The percentage of connections that were to different services, among the connections aggregated in dst_host_count :

dst_host_diff_srv_rate

The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count :

dst_host_same_src_port_rate

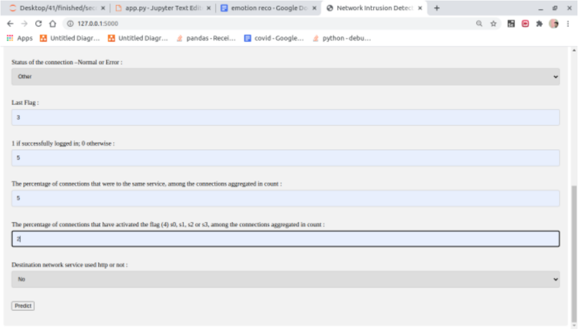
The percentage of connections that were to the same service, among the connections aggregated in dst_host_count :

dst_host_same_srv_rate

Number of connections having the same port number :

dst_host_srv_count

Fig 8 Main Page



Status of the connection –Normal or Error :

Other

Last Flag :

3

1 if successfully logged in; 0 otherwise :

5

The percentage of connections that were to the same service, among the connections aggregated in count :

5

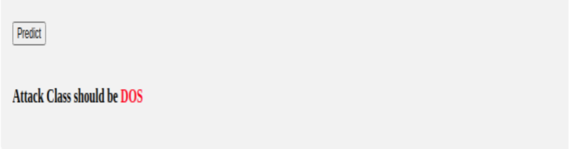
The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count :

2

Destination network service used http or not :

No

Fig 9 Upload Input Values



Attack Class should be DOS

Fig 10 Prediction Result

6. CONCLUSION

Assessments of assist vector with machining, ANN, CNN, random forest, and deep learning computations considering the continuous CICIDS2017 dataset have been presented lately. As indicated by the discoveries, the profound learning estimation performed better compared to SVM, ANN, RF, and CNN. In view of this dataset, we will involve port range drives as well as different sorts of assaults utilizing ML and deep learning computations, Apache Hadoop, and shimmer advancements later. These calculations help us in distinctive a computerized assault in an association. It occurs so that, assuming we recollect numerous years, there might have been such countless attacks that when these assaults are distinguished, the qualities of the places where they are occurring are kept specifically datasets. Thus, we will actually want to foresee whether a cyberattack will happen by using these insights. These expectations can be made by SVM, ANN, RF, and CNN, among other four calculations. This study recognizes the calculation with the most elevated exactness rates for anticipating the best results for deciding if cyberattacks occurred.

Later on, we will consolidate some ML calculations to improve accuracy.

REFERENCES

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, "Bilgi g ¨uvenli ˘gisistemlerindekullanilanarac¸larinincelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] S. Staniford, J. A. Hoagland, and J. M. McAlerney, "Practical automated detection of stealthy portscans," Journal of Computer Security, vol. 10, no. 1-2, pp. 105–136, 2002.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, "Surveillance detection in high bandwidth environments," in DARPA Information Survivability Conference and Exposition, 2003.Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[6] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in Wireless

Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

[7] N. Moustafa and J. Slay, “The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems,” in Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on. IEEE, 2015, pp. 25–31.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, “Detection and classification of malicious patterns in network traffic using benford’s law,” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864–872.

[9] S. M. Almansob and S. S. Lomte, “Addressing challenges for intrusion detection system using naive bayes and pca algorithm,” in Convergence in Technology (I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.

[10] M. C. Raja and M. M. A. Rabbani, “Combined analysis of support vector machine and principle component analysis for ids,” in IEEE International Conference on Communication and Electronics Systems, 2016, pp. 1–5.

[11] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization.” in ICISSP, 2018, pp. 108–116.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, “Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm,” in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.

[14] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, “Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark,” IEEE Access, 2018.

[15] P. A. A. Resende and A. C. Drummond, “Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling,” Security and Privacy, vol. 1, no. 4, p. e36, 2018.

[16] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[17] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, “Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct,” Bone marrow transplantation, vol. 49, no. 3, p. 332, 2014.