

# Detection of Cyberattacks Using Machine Learning Techniques

*Neelima Santoshi K<sup>1</sup>, Yogesh M<sup>2</sup>, Jayaram majeti<sup>3</sup>, Balakrishna reddy D<sup>4</sup>, Chandra Sagar M<sup>5</sup>*

<sup>1</sup>Assistant Professor, Department of Computer Science Engineering, Gitam University, Visakhapatnam.

<sup>2,3,4,5</sup>Student, Department of Computer Science Engineering, Gitam University, Visakhapatnam.

**Guide - Mrs. Neelima Santoshi K<sup>1</sup>**

## Abstract:

Contrasted with the past, upgrades in PC and correspondence improvements have given extensive and propelled changes. The use of latest improvements give exceptional advantages to people, organizations, and governments, be that as it is, messes a few up against them. For instance, the safety of significant data, security of positioned away statistics stages, accessibility of statistics and so forth. Contingent upon those problems, virtual worry primarily based totally oppression is one of the maximum significant problems on this day and age. Digital worry, which made a first rate deal of problems people and establishments, has arrived at a stage that would undermine open and state safety with the aid of using extraordinary gatherings, for example, criminal association, proficient humans and virtual activists. Along those lines, Intrusion Detection Systems (IDS) has been created to preserve a strategic distance from virtual assaults. Right now, mastering the bolster support vector machine (SVM) calculations had been applied to understand port sweep endeavours depending on the new CICIDS2017 dataset with 97.80%, 69.79% precision rates had been achieved individually. Rather than SVM we are able to introduce a few different algorithms like the random forest, CNN, ANN in which those algorithms can accumulate accuracies.

## INTRODUCTION:

Nowadays machine learning is growing rapidly make people dependent on machine learning techniques and classifiers than ever before. And same time the number of security intrusions has growing rapidly. Therefore the security is

important. This says that the security and reliability of devices, as well as effective protection against various networks attacks that create vulnerabilities in installed security system. The intrusion detection system is considered one of the machine learning tools to monitor suspicious activities. In the modern world everyone is using their internet through smartphones and laptops so that the internet facility should be 24 × 7 without interruption. Before finding malicious attacks one should know about the basic nature of such attacks. The use of new innovations give incredible advantages to people organizations and governments be that as it may mess some up against them for instance the protection of significant data security of put away information stages accessibility of information and so forth contingent upon these issues digital fear based oppression is one of the most significant issues in this day and age digital fear which made a great deal of issues people and establishments has arrived at a level that could undermine open and national security by different gatherings, for example, criminal association proficient people and digital activists along these lines intrusion detection systems IDS has been created to maintain a strategic distance from digital assaults

## LITERATURE REVIEW:

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

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.

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.

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.

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.

## METHODOLOGY:

### Logistic Regression:

This algorithm gives perceivability into discrete arrangements of classes and uses the sigmoid capacity to recover the stamping worth of at least 2 classes. There are various sorts of this algorithm, like,

- Binary
- Multi
- Ordinal.

Binary Logistic Regression(BLR) is utilized in this paper. Sigmoid Function is utilized in this algorithm and this guides a worth to another esteem and these qualities scale from 0 to 1. The sigmoid capacity is given by:

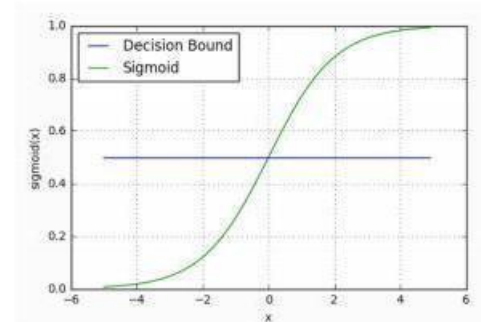
$$S(z) = \frac{1}{1 + e^{-z}}$$

**Fig-1**

Sigmoid Function

Here  $S(z)$  is the yield somewhere in the range of 0 and 1,  $z$  is the function's input. What's more,  $e$  is the regular log's base. An edge esteem called the

choice bound is chosen to map the likelihood score which the order work gets back to a discrete class.



**Fig-2**

Logistic Regression

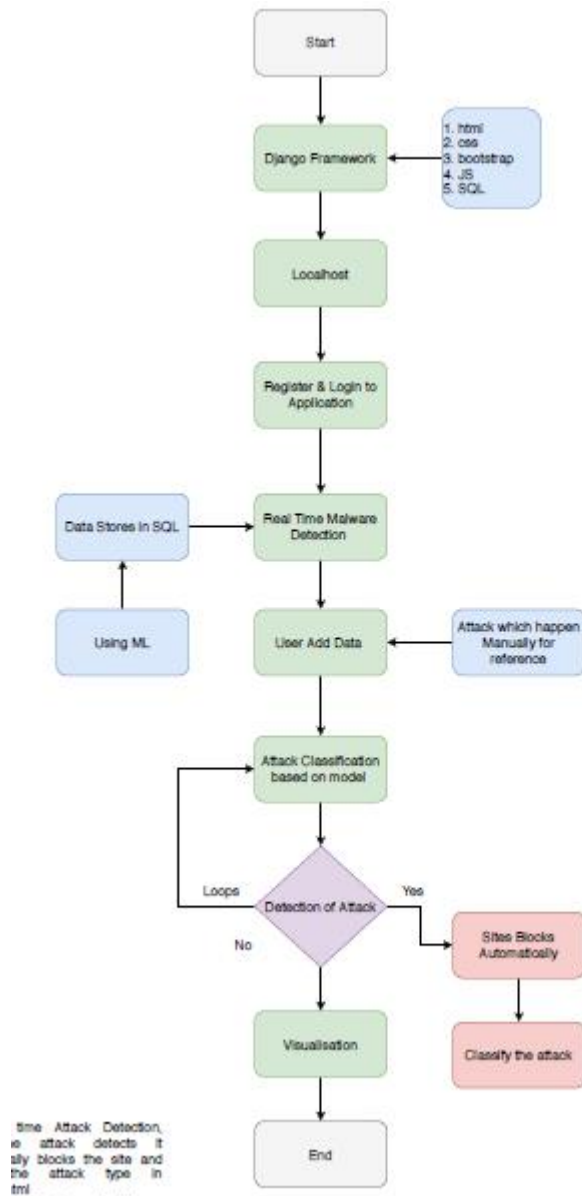
From these sigmoid capacities and choice limits, we can process the forecast result of the characterization by the Logistic Regression model. A resource separation utilizes the sigmoid capacity to change over the outcome into a number of chances; the point is to diminish work expenses to accomplish better openings. Cost work is determined as displayed in

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} \log(h_{\theta}(x)), & y = 1, \\ -\log(1 - h_{\theta}(x)), & y = 0. \end{cases}$$

**Fig-**

3 Cost work

This calculation was carried out by bringing in the library Logistic Regression from Scikitlearn in the way: from sklearn.linear\_model import LogisticRelapse. The classifier was then fit on the preparation elements and marks. The work predict\_probability was utilized to assess the likelihood. The capacity anticipate was utilized to make the genuine expectations for class names



### Validation:

It is the cycle to ensure that the item fulfill the referenced necessities toward the finish of the advancement stage. All in all, to ensure the item is worked as client necessities.

### Basics of software testing:

The two basics are:

### Black box Testing:

It is a trying procedure that keeps away from the interior instrument of framework and concentration on yield created against any information worth and execution of the framework. Practical exploration is one more name for it.

### White box Testing:

It is a trying procedure that takes into the inward instrument of a framework.

### Classification Accuracy:

It is the proportion between the number right expectations to the complete number of input tests in the dataset.

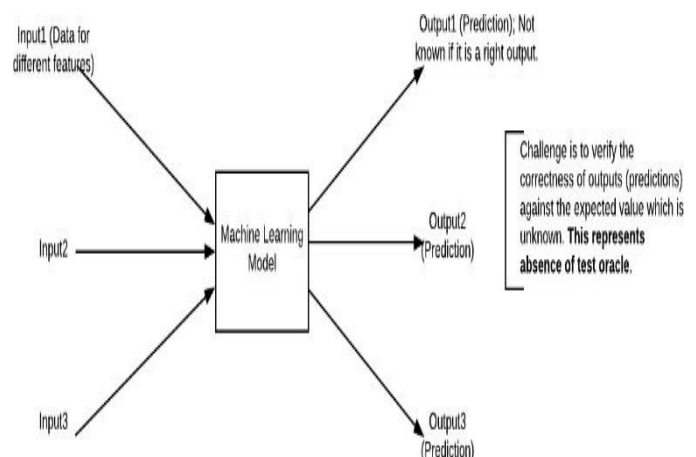
$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

### Software Testing:

It is the strategy for estimating a product object to recognize disparities between the provided input and the anticipated output. Testing tracks down the nature of the item. A cycle ought to be finished during the advancement exchange. In another words this is known to be verification and validation measure.

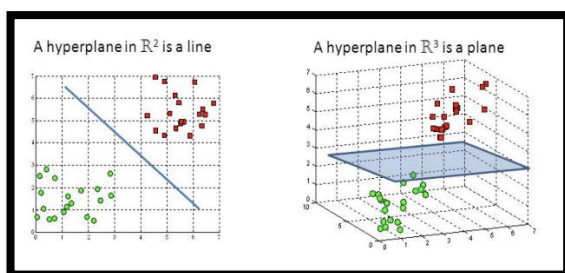
### Verification:

It is the interaction to ensure that the item fulfill the conditions and take advantage toward the beginning of advancement stage. In another words, it is to make sure the item acts as the result we need.



## SVM:

The goal of the support vector machine strategy is to find a hyper plane in a  $n$ -layered space  $n$  the quantity of elements that unmistakably groups the items to isolate the two classes of items there are numerous conceivable hyper planes that could be picked our goal is to observe a plane that has the greatest edge for example the most extreme distance between relevant informative elements of the two classes expanding the edge distance gives some support so future information focuses can be arranged with more certainty hyper planes are choice limits that assist with characterizing the elements information focuses falling on one or the other side of the hyper plane can be ascribed to various classes additionally the component of the hyper plane relies on the quantity of elements in the event that the quantity of info highlights is 2 the hyper plane is only a line on the off chance that the quantity of info highlights is 3 the hyper plane turns into a two-layered plane it becomes hard to envision when the quantity of highlights surpasses 3 support vectors are information focuses that are nearer to the hyper plane and impact the position and direction of the hyper plane utilizing these help vectors we boost the edge of the classifier erasing the help vectors will change the place of the hyper plane these are the focuses that assist us with building our svm.



## Random forest

This ML strategy that is utilized to tackle relapse and characterization issues. It uses troupe realizing, which is a procedure that consolidates numerous classifiers to give answers for complex issues.

An random forest calculation comprises of numerous choice trees. The 'forest' created by this strategy is prepared through sacking or bootstrap accumulating. Stowing is a gathering meta-calculation that works on the exactness of ML calculations.

The (random forest) calculation lays out the result in view of the expectations of the choice trees. It predicts by taking the normal or mean of the result from different trees. Expanding the quantity of trees builds the accuracy of the result.

An random forest kills the limits of a choice tree calculation. It lessens the over fitting of datasets and increments accuracy. It produces forecasts without requiring numerous designs in bundles (like Scikit-learn).

### Highlights of a Random Forest Algorithm:

It's more exact than the choice tree calculation.

It gives a successful approach to taking care of missing information.

It can create a sensible expectation without hyper-boundary tuning.

It addresses the issue of over fitting in choice trees.

In each irregular woods tree, a subset of elements is chosen haphazardly at the hub's parting point.

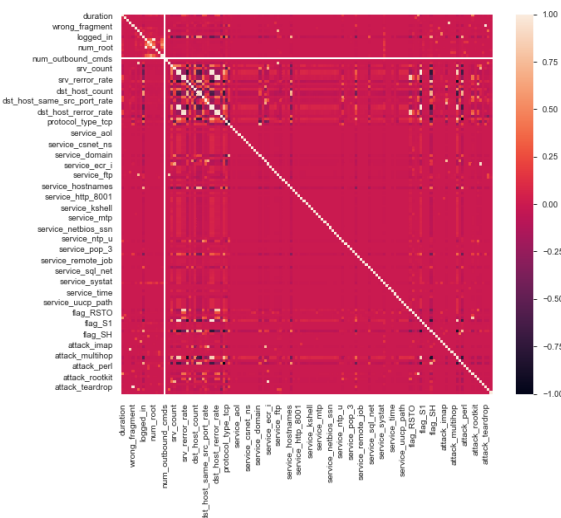
Decision trees are the structure squares of this calculation. A decision tree is a choice help method that frames a tree-like construction. An outline of choice trees will assist us with seeing how these calculations work.

## Decision tree

Decision tree is a managed ML strategy where information is consistently isolated at each column in view of specific guidelines until the ultimate result is produced lets take a model guess you open a shopping center and obviously you would need it to develop in business with time so besides you would require returning clients in addition to new clients in your shopping center for this you would plan different business and promoting procedures, for example, sending messages to potential clients make offers and arrangements focusing on new clients and so forth however how do we have any idea who are the possible clients all in all how

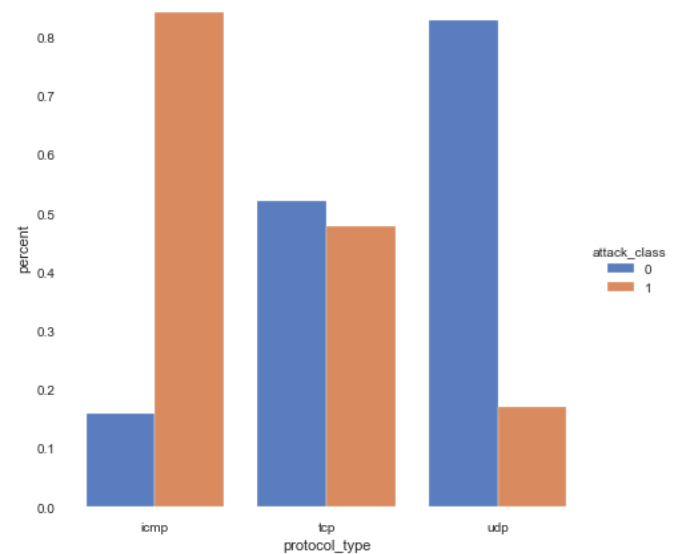
would we characterize the class of the clients like a few clients will visit once in a week and others might want to visit more than once in per month or some will visit in a quarter so choice trees are one such order calculation that will group the outcomes into bunches until no greater comparability is left

## Result:



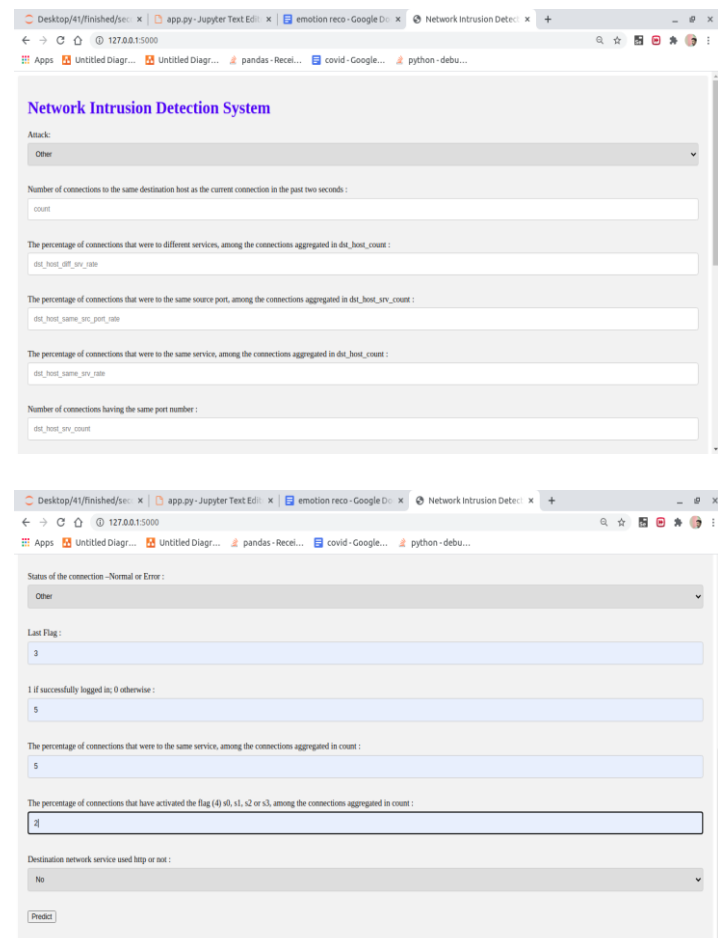
The figure is about the heat map plot output. Here we check whether correlation is there or not in the columns that are in the data set. so as we can see that diagonal line in the plot says about the there is correlation.

Since for each section Is connected to itself so that is the reason the correlation is high in the diagonal line. On the off chance that the shading dimness is high, the connection is low else if the relationship is high the haziness of plot will be low.



In this graph we are identifying relationships (between Y & numerical independent variables by comparing means)

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\_srv\_cnt

The percentage of connections that were to the same source port, among the connections aggregated in dst\_host\_srv\_cnt: dst\_host\_src\_port\_cnt

The percentage of connections that were to the same service, among the connections aggregated in dst\_host\_count: dst\_host\_srv\_cnt

Number of connections having the same port number: dst\_host\_srv\_count

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) s1, s2 or s3, among the connections aggregated in count: 2

Destination network service used http or not: No

Predict



Predict

Attack Class should be **DOS**

## Conclusion:

Right now, estimations of help vector machine, ANN, CNN, Random Forest and profound learning calculations dependent on modern CICIDS2017 dataset were introduced relatively. Results show that the profound learning calculation performed fundamentally preferable outcomes over SVM, ANN, RF and CNN. We are going to utilize port sweep endeavors as well as other assault types with AI and profound learning calculations, apache Hadoop and sparkle innovations together dependent on this dataset later on. All these calculation helps us to detect the cyber attack in network. It happens in the way that when we consider long back years there may be so many attacks happened so when these attacks are recognized then the features at which values these attacks are happening will be stored in some datasets. So by using these datasets we are going to predict whether cyber attack is done or not. These predictions can be done by four algorithms like SVM, ANN, RF, CNN this paper helps to identify which algorithm predicts the best accuracy rates which helps to predict best results to identify the cyber attacks happened or not.

## 8. 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 ğan, "Bilgi ğ uvenli ğ i sistemlerinde kullanilan arac,larin incelenmesi," 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 unswnb15 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.