

# DETECTION OF CYBER ATTACK IN NETWORK USING MACHINE LEARNING TECHNIQUES

<sup>1</sup>Dr.D.Rathna Kishore, <sup>2</sup>Dr.Davuluri Suneetha

<sup>1</sup>Professor, Dept. of CSE, Andhra Layola College, Vijayawada, A.P, India

<sup>2</sup>Professor, Dept. of CSE, NRI Institute of Technology, Vijayawada, A.P, India

Abstract:

The evolution of personal computers and electronic mail has resulted in significant shifts when compared to the past. Even when using them maliciously, modern technologies provide enormous advantages to people, businesses, and governments. For instance, large data security, data stadium security, information accessibility, etc. One of the most pressing issues of our day may be digitally-enabled fear-based tyranny. In light of the many challenges posed by the rise of digital technologies, both individuals and institutions have begun to worry that criminal organizations, government agencies, and other non-state actors (also known as "hacktivists") could use these vulnerabilities to launch attacks on the United States. By design, IDSs are isolated from any potential harm caused by cyberattacks. Currently available intrusion detection systems (IDS) employ learning the calculation of the SVM (Bolster Support Vector Machine) to identify port sweep attempts that rely on a new data set of CICIDS 2017, with a combined success rate of 69.79%. Instead of SVM, we might use techniques like Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Random Forests to improve accuracy to levels like SVM (93.29), CNN (63.52), Random Forest (99.93), and ANN (99.11).

*Keywords: Machine Learning, KDD, Cyber Security, Network, SVM, RandomForest.*

## 1. INTRODUCTION:

Recently, the globe has seen a significant development in the many fields of connected innovations, such as dazzling matrices, the Internet of cars, long haul advancement, and 5G communication. According to Cisco [1,] it is expected that by the year 2022, the number of IP-connected devices would be several times more than the global population, generating 4.8 ZB of IP traffic annually. This prediction is based on current trends. This accelerated development raises overwhelming security concerns due to the trading of enormous amounts of sensitive data through asset-required devices and over the untrusted 'Internet' utilizing a variety of different technologies and communication conventions. This poses a significant threat to the security of

sensitive information. Before information is sent over the internet, more advanced security measures and flexibility testing need to be carried out so that the internet may continue to be accessible while still being safe.

The implemented security measures are responsible for preventing attacks, recognizing them when they occur, and responding appropriately to them. An interruption recognition system, also known as an IDS, is a commonly used method for identifying both internal and external interruptions that target a system, as well as irregularities that point to likely interruptions and questionable activities. This method is typically employed for the purpose of location. An intrusion detection system (IDS) is comprised of a number of different apparatuses and mechanisms. These are used for monitoring the computer system and the organization's traffic, as well as for dissecting activities with the intention of locating potential security breaches affecting the system. A combination IDS, a signature-based IDS, or an inconsistency-based IDS are all possible ways to implement an intrusion detection system. A signature-based IDS identifies interruptions by comparing observed behaviors and precharacterized interruption designs. An oddity-based IDS, on the other hand, focuses on understanding usual behavior in order in order to differentiate any divergence [2]. Identifying anomalies may be accomplished via the use of a number of approaches, including AI, information, and fact-based methods; more recently, research has been conducted on deep learning approaches.

Misdeeds committed with presentation computers have been steadily increasing. They are not merely restricted to irrelevant demonstrations, such as examining the login credentials of a structure, but in addition to this, they pose a greater danger overall. Information security refers to the process of preventing information from being used or accessed in an unauthorized manner, being exposed to unauthorized parties, being altered, destroyed, or damaged without authorization. The phrases "Information security," "PC security," and "information assurance" are often used interchangeably with one another. These domains are connected to one another and share destinations in order to provide availability, mystique, and authenticity in the information they provide. According to the findings of studies, the first step in the assault is the disclosure of information. The structure is observed in order to get current knowledge about it, hence observations are made. Finding a brief overview of open ports in a design provides an attacker with info that is very vital to their mission. Therefore, there are a large number of devices that may detect open ports [3, such as insect detection systems and underground insect diseases]. As of right now, learning and SVM AI calculations have been used to the making of

IDS models in order to observe port yield efforts. the models were provided with an explanation of the materials and methods that were employed.

## 2. LITERATURE SURVEY

This part highlights a variety of recent accomplishments in and around this area. It is important to note that we only evaluate the work of those researchers who have benchmarked their performance by comparing it to the NSL-KDD dataset. Consequently, from right now forward, every dataset that is even remotely mentioned should be regarded as NSL-KDD. This technique enables a more precise comparison of the work in question with that of other examples contained in the text. A further limitation is that the majority of the work must use the information intended for preparation in both the testing and the preparation stages. At long last, we investigate a handful of deep learning-based approaches that have been developed up to this point for work that is equivalent to what we've been doing.

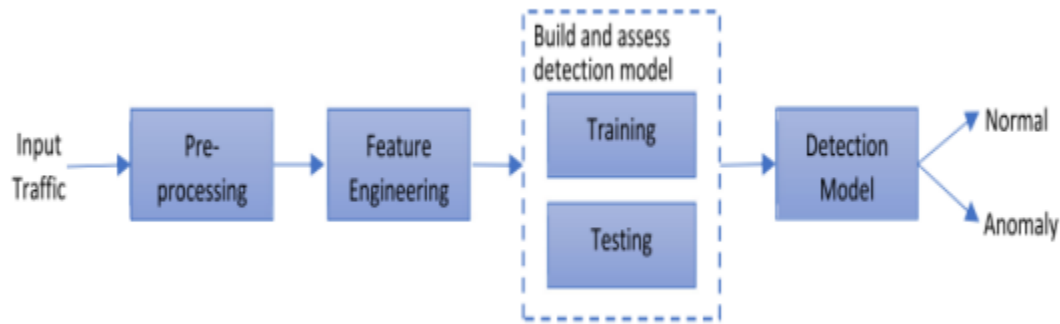
One of the most precise works ever discovered in written form made use of ANN with an enhanced robust back-spread for the purpose of planning such an IDS [6]. Only the preparation dataset was used throughout this project, and preparation accounted for 70%, approval for 15%, and testing for 15%. A decrease in execution time was achieved, as was to be predicted, by the employment of unlabeled information for testing. A subsequent piece of research used the J48 decision tree classifier with 10-overlay cross-approval for testing on the dataset used for the preparation [4]. Instead of the whole arrangement of 41 capabilities, this work made use of a reduced set of capabilities consisting of just 22 highlights. Similar research evaluated a variety of well-known regulated tree-based classifiers and found that the Random Tree model performed the best, with the highest degree of accuracy and the lowest percentage of false positives [5]. The researchers concluded that this model should be used.

There have also been several master presentations of different 2-level characterisation methodologies. In one such study, the Discriminative Multinomial Naive Bayes (DMNB) algorithm was used as the primary classifier, followed by Nominal-to-Binary directed separation at the second level, and 10-fold cross validation was used for testing [9]. This effort was an attempt to reach out and make use of Ensembles of Balanced Nested Dichotomies (END) at the primary level and Random Forest at the second level [10]. This update, as expected, led to an improvement in the location rate as well as a reduction in the rate of false positives. Another 2-level execution that used principal component analysis (PCA) for the list of capabilities

reduction and then support vector machines (using the Radial Basis Function) for final classification brought about a high recognition precision with just the training dataset and the full set of 41 highlights. A reduction in features from 23 to 23 brought in an improvement in location accuracy in several of the attack classes; nonetheless, the overall performance was worsened [11]. The authors made their job better by first using data gain to rank the highlights and then using a behavior-based element determination to narrow the list of capabilities down to 20. Because of this, there was an increase in the detailed accuracy achieved when employing the preparation dataset [12].

The next category to investigate made use of both the practice and the real-world datasets. An underlying effort in this classification made use of fuzzy characterisation in conjunction with hereditary computation, which resulted in a detection accuracy of 80%+ and a low proportion of false positives [13]. Another important piece of research utilized unaided grouping algorithms and discovered that the exhibition using just the preparation information was drastically reduced when test information was also utilized [6]. This finding was brought about by the fact that the two sets of data were compared to one another. Using both the training and the test datasets, a comparison execution of the k-point computation resulted to a slightly improved recognition accuracy and a decreased false positive rate [7]. OPF (optimal way woodlands), a less well-known approach that use chart apportioning for the purpose of include classification, was discovered to demonstrate a high identification accuracy [8] within 33% of the time when compared to the SVM RBF technique. This was the case when the comparison was made.

### **3. PROPOSED SYSTEM**



**Fig 1: Proposed System**

1. Data Collection: Collect adequate data samples as well as valid software samples.
2. The Data Preparation Process: In order to achieve higher levels of performance, augmented technologies will be used.
3. Train and Test Modelling: Separate the data into train data and test data Train data will be used for training the model, and Test data will be used to evaluate how well the model is doing.
4. Attack Detection Model: An algorithm that has been trained based on the model will determine whether or not a particular transaction is unusual.

The most important stages of the algorithm are outlined in the figure, which may be found below. 1) The normalization of each individual dataset. 2) Transform the dataset into one for testing and one for training. 3) Develop IDS models by using RF, ANN, CNN, and SVM algorithms as needed. 4) Evaluate every model's performances.

The following are some advantages that may be gained by using the suggested systems:

- Protection against harmful assaults made on your network.
- The elimination and/or assuring the presence of harmful components inside an already-established network.
- Prevents people from accessing the network in an unlawful manner.
- Programs access to certain resources that may contain an infection.
- Protecting private and sensitive information

#### **4. RESULTS**

Machine learning libraries such as numpy, pandas, and scikitlearn were used in order to carry out the research. The application is being built utilizing the Python programming language and the jupyter notebook integrated development environment (IDE).

Four different algorithms, including SVM, ANN, RF, and CNN, may be used to make predictions. This study helps to find which algorithm forecasts the greatest accuracy rates, which in turn helps to anticipate the best outcomes for determining whether or not cyber attacks have taken place.

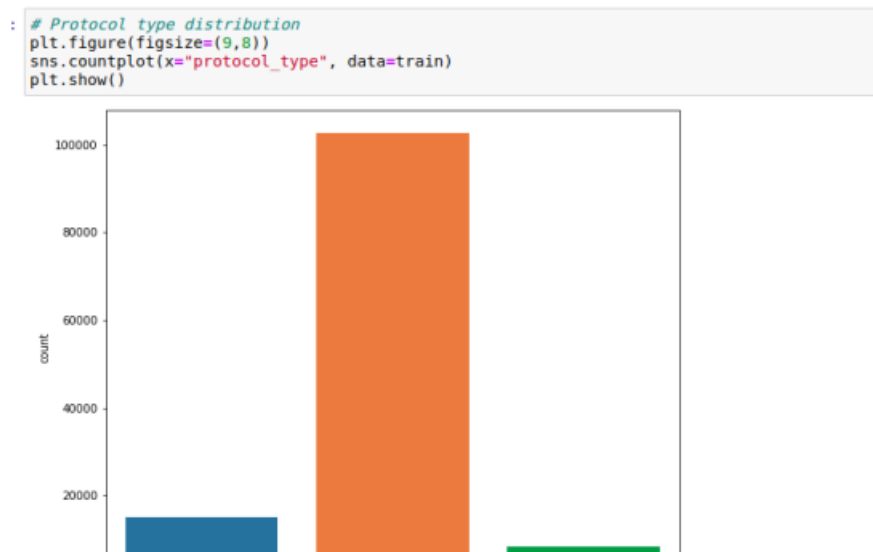


Fig: 2 Protocol Type Distribution

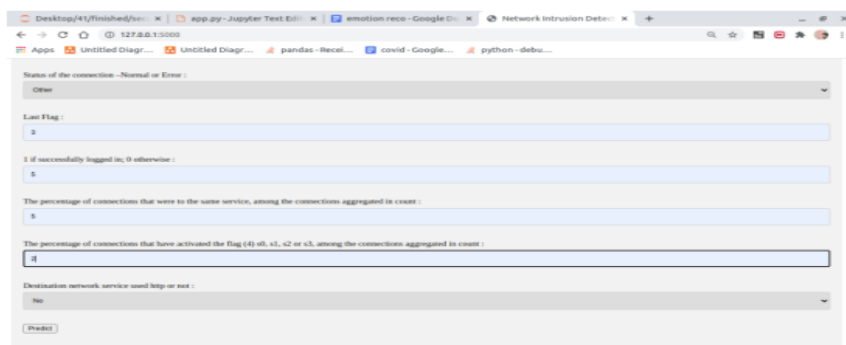
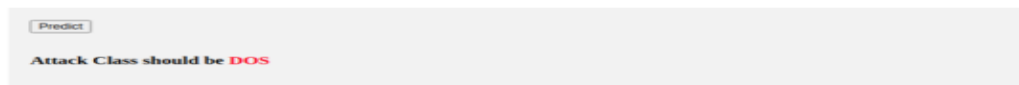


Fig: 3 Data Collection for Analysis



**Fig: 4 Predicting the type of Attack**

## 5. CONCLUSIONS:

Assessments of the assist vector machine, artificial neural networks, convolutional neural networks, random forests, and substantial learning estimates based on the current CICIDS2017 dataset were considerably provided at this time. The findings indicate that substantial learning estimation performed typically better than SVM, ANN, RF, and CNN when compared to the outcomes. In the future, we want to employ AI and major learning calculations, Apache Hadoop and technological developments together to ward against attacks using this dataset. Our defenses will include port scope efforts as well as other forms of attacks. Each and every one of these estimations helps us recognize the digital attack that was made on the network. It happens in the way that when we consider events that took place a very long time ago, there may have been a very large number of assaults that took place, and when these assaults are recognized, the information regarding the values at which these assaults are taking place will be saved in some datasets. Therefore, by analyzing these statistics, we will be able to forecast whether or not the digital attack has been completed. These projections ought to be attainable by the use of four different computations, namely SVM, ANN, RF, and CNN. This study helps identify which computation forecasts the greatest accuracy rates, which in turn helps with foreseeing the best results to detect whether or not the digital attacks happened.

## 6. 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 g ğuvenli ğgi sistemlerinde kullanılan arac,larin incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

- [4] Rashmi T V. “Predicting the System Failures Using Machine Learning Algorithms”. International Journal of Advanced Scientific Innovation, vol. 1, no. 1, Dec. 2020, doi:10.5281/zenodo.4641686.
- [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. Ibrahim 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] Girish L, Rao SKN (2020) “Quantifying sensitivity and performance degradation of virtual machines using machine learning.”, Journal of Computational and Theoretical Nanoscience, Volume 17, Numbers 9- 10, September/October 2020, pp. 4055- 4060(6) <https://doi.org/10.1166/jctn.2020.9019>.
- [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] Girish, L., & Deepthi ,T. K.(2018). Efficient Monitoring Of Time Series Data Using Dynamic Alerting. i-manager’s Journal on Computer Science, 6(2), 1-6. <https://doi.org/10.26634/jcom.6.2.14870>
- [11] Nayana, Y., Justin Gopinath, and L. Girish. "DDoS Mitigation using Software Defined Network." International Journal of Engineering Trends and Technology (IJETT) 24.5 (2015): 258-264.
- [12] Shambulingappa H S. “Crude Oil Price Forecasting Using Machine Learning”. International Journal of Advanced Scientific Innovation, vol. 1, no. 1, Mar. 2021, doi:10.5281/zenodo.4641697.
- [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.