International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

Network Intrusion Detection System using Machine Learning Techniques

K. Neeraj Reddy[#1], S. Harish Reddy[#2], P. Vamshi[#3], A.Ashwini [#4]

 # Department of CSE, Vardhaman College of Engineering, Telangana, India.

Abstract - *The constant progress of technology has brought forth a lot of advantages and major difficulties for several businesses, chief among being cybersecurity. Robust intrusion detection systems (IDS) are necessary to protect against hostile activities due to the increase in cyber threats. In this study, we identify potential intrusions using machine learning techniques, namely the Support Vector Machine (SVM) algorithm, using the CICIDS2017 dataset. Our testing produced encouraging results, with SVM recognizing pilot port incursions with an accuracy of 97.80%. We also investigate the effectiveness of other algorithms, including Random Forest, Convolutional Neural Networks (CNN), and Artificial Neural Networks (ANN), which shown different accuracy levels between 63.52% and 99.93%. Our proposed system encompasses data collection, preprocessing, training, and testing modelling, culminating in the development of an attack detection model. The system holds immense potential in fortifying networks against malicious attacks, removing or securing malicious content, and ensuring the confidentiality of sensitive information.*

***Key Words*: Machine Learning, KDD, Cyber Security, Network, SVM, Random Forest.**

1.INTRODUCTION

Recent years have seen significant changes in many areas of connected technology, including smart grids, the Internet of Things (IoT), long-term development and 5G communications. The number of devices connected to an IP network is expected to exceed ten times the world's population by 2022, generating 4.8 ZB of IP traffic per year. Because a lot of sensitive information is transported across the unreliable "Internet" utilizing dated and inconsistent technology and communication protocols, this fast expansion raises severe security problems. To ensure the safety and security of the Internet, higher security measures and potential analysis should becarried out in the early stages of deployment.

Attacks must be prevented, detected, and responded to by the security mechanisms that have been put in place. intrusion. detection systems (IDS) are extensively employed to detect and identify any malicious activities and internal and external network attacks, along with any irregularities that may point to an incursion. Tools and methods for keeping an eye on network traffic and computer systems are combined into an intrusion detection system (IDS). It is an IDS that may be used as a hybrid, exception-based, or signature system. An intrusion detection system (IDS) that uses signatures compares proposed properties with pre- established intrusion patterns to find intrusions. In contrast, anomaly-based intrusion detection systems concentrate on proficiently interpreting actions to detect deviations. A variety of methods (e.g., statistics-based, data- based, and artificial intelligence-based, including recent research on deep learning) are used to detect implausibility.

The field of malpractice in computers has evolved and continues to move beyond trivial behaviour, such as monitoring access to information, primarily to uncover larger threats. Data protection from unauthorized use, disclosure, alteration, destruction, and corruption is known as information security. The words "data security," "computer security," and "data protection" are sometimes used synonymously. These interactive lists are made to guarantee information availability, confidentiality, and integrity. Empirical studies indicate that gathering and identifying system information is the initial stage of an attack. Checking open ports in the system is important information for attackers and requires the use of various tools such as antivirus software and IDS. Recently, an IDS model for port testing that captures knowledge and tactics has been developed using machine learning and support vector machine (SVM) methods.

2. RELATED WORK

This article provides an overview of current developments in the subject, with a particular emphasis on studies on the use of NSL-KDD data. Therefore, in order to promote improved comparability of research in the literature, all aspects stated subsequently should be handled in accordance with NSL-KDD. The majority of the research employed training and testing data, which is a significant restriction. Lastly, several search algorithms based on deep learning for comparable problems are explored. One of the earliest studies was the development of an artificial neural network (ANN) to identify intrusiondetection systems (IDS) with enhanced recovery capabilities.

This approach uses training data only for training (70%), validation (15%), and testing (15%); this results in decreased performance when using unnamed for testing.

Follow-up research employed 22 out of 41 feature sets for 10-fold cross-validation of the training data using the J48 decision tree classification. A similar study evaluated different tree managers, identifying the random tree model as the most effective in terms of accuracy and minimizing false positives. Several two-level classification techniques have been proposed, such as the use of nominal for second-level pairwise filtering with 10-fold crime and discriminant polynomial naive Bayes (DMNB) as the basis classifier [9]. In order to increase detection and lower false positives, this work has been expanded to include the usage of Random Forest at the second level and Balanced Nested Binary (END) clusters at the first level. The other two-stage method uses principal component analysis (PCA) for minimization and support vector machine (SVM) (using the radial basis function) for final classification. When detection accuracy was achieved using training data and a total of 41 features, reducing the features to 23 increased accuracies for some attacks, although overall performance decreased.

The authors improved their method by using incremental data as a selection for optimal performance and behaviour, reducing the feature set to 20, and using knowledge training to improve reporting accuracy. The next group considers training and testing data. The first experiment in this category used a genetic algorithm for fuzzy classification with over 80% detection accuracy and low alarm rate. Another important study using unsupervised methods showed that performance decreased when both training and testing data were used. Similar performance was achieved using the k-point algorithm for both data sets, resulting in slightly better detection accuracy and lower false alarms. Using test and preparation datasets, Optimal Path Forest (OPF), a less well-known technique, uses graph partitioning for classification and offers a third more accurate detection than the SVM RBF method. It also reduces the false positive rate and somewhat improves recognition accuracy. It was demonstrated that the less well-known OPF (optimal way woodlands) technique, which uses chart apportioning for include classification, had a good identification accuracy within 33% of the time when compared to the SVM RBF approach.

3. PROPOSED SYSTEM

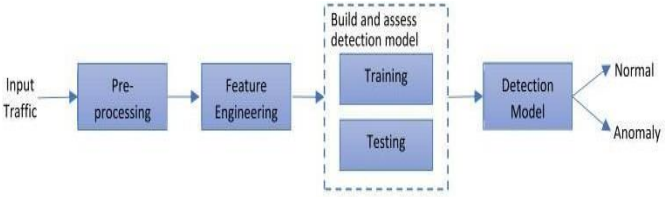

Fig 1: Proposed System

The process illustrated in the image represents a typical machine learning pipeline for anomaly detection. It starts with 'Input Traffic,' where raw data is collected. This is followed by 'Pre-processing,' where the data is cleaned and formatted. Next, 'Feature Engineering' is applied to extract relevant characteristics from the data. The 'Training' stage involves developing a model using the engineered features. 'Testing' then assesses the model's performance. The 'Detection Model' stage involves the model's actual deployment, where it classifies new data as 'Normal' or 'Anomaly.' Finally, an action is taken based on the detected anomalies.

Module implementation:

- **Data Collection:** Compile appropriate software structures and data structures.
- **Data Preprocessing:** To improve speed, data augmentation techniques will be applied.
- **Train and Test Modelling:** The training and test sets of data are separated. After the model has been trained using the training set, its performance is evaluated using the test set.
- **Attack Detection Model:** The model-trained algorithms will identify the anomalous changes.

This algorithm includes several important steps, described below and shown in Figure 1. First, all datasets are standardized. Then the data set is divided into two: test and training set. Then, an intrusion detection system (IDS) was developed using RF, ANN, CNN and SVM algorithms. Finally, the performance of each model is carefully evaluated.

The following are this system's benefits:

- Protect your network from malicious attacks.
- Remove and/or secure malicious content on existing networks.
- Prevent users from accessing the network without permission.
- Refusing to accept services from potentially infectious programs.
- Protect confidential information

4. EXPERIMENTAL RESULTS

A. Datasets Description

The DARPA Identity Assessment Project of 1998 managed and created by MIT Lincoln Laboratory focused on studying identity and conducting research. It documented a series of actions that disrupted the military setting and led to harm to civilians. The data, from the 1999 KDD Intrusion Detection Challenge is a version of this

information, influenced by the DARPA initiative. The DARPA ID evaluation team gathered network-based IDS data by simulating a LAN at an Air Force base with over 1,000 UNIX nodes and monitoring users at Lincoln Laboratory for nine weeks. To eliminate TCP data the process requires training and testing for either 7 or 14 weeks. The MIT laboratory, backed significantly by DARPA and AFRL funding employs Windows and UNIX nodes to manage traffic across LANs as opposed to systems. A total of 300 attacks were simulated in this dataset comprising 7 scenarios with 32 attack types. Since its introduction the KDD 99 dataset has been widely used for assessing IDSs. Includes information from, around 4.9 million individuals associated with 41 unique identifiers.

Detailed instructions were given for the simulated attacks as described below;

Denial of Service Attack (DoS); Intrusion involves an attempt to disrupt a host's function by temporarily or sometimes permanently interfering with services.

To achieve this the goal is to flood the intended computer with a number of requests leading to an overload of the system. A User-to-Root-Attack (U2R) is a frequently used attack method in which the attacker tries to obtain root power by trying to acquire a user's previous access and taking advantage of security holes. An attacker using a Remote-to- Local-Attack (R2L) can transmit data packets to the target, but they do not have a user account on the computer. Their goal is to utilize a vulnerability to obtain local access while posing as the user in use.

The purpose of a probing attack is for the attacker tolearn as much as possible about the machines in the company in order to get past the firewall and gain root access.

DARPA's Identity Assessment Team dedicated itself to gathering enterprise information on IDS by creating a terminal LAN and deploying more than 1,000 UNIX hubs over the course of nine weeks. Currently, hundreds of users have joined

Lincoln Laboratory consists of two components; a 7 day training session and a 14 day testing phase. This approach is aimed at examining TCP data dumps. With support, from DARPA and AFRL the MIT lab stands out from facilities by utilizing Windows and UNIX platforms to cover a wide range of access points across remote LANs.

The data generation process involved executing 7 scenarios and carrying out 32 attacks resulting in a total of 300 simulated attacks.

The KDD '99' dataset has become widely recognized as one of the datasets for evaluating different Intrusion Detection Systems (IDSs). It contains information on than 4,900,000 individuals with 41 attributes.

This article delves into the characteristics of attack attempts, such as Denial of Service (DoS) attacks, well as R2L/U2R attacks. Generally speaking, hacking attempts and DoS attacks can take place with damage patterns resembling those seen in R2L and U2R attacks. This distinction could be attributed to the fact that R2L and U2R attacks typically involve packets with connections while testing and DoS assaults require connections to various sites, within a short timeframe. Detecting attacks relies heavily on identifying features known as signatures that help pinpoint complex behaviors associated with these malicious activities.

B. Results:

Four different algorithms. Support Vector Machine (SVM) Artificial Neural Network (ANN) Random Forest (RF) and Convolutional Neural Network (CNN). Were utilized in the study. The primary aim of this investigation was to ascertain the algorithms for detecting cyberattacks and enhancing event prediction accuracy.

Based on the data sourced from CICIDS2017 various experiments were carried out using support vector machines, random forests, artificial neural networks, convolutional neural networks and learning models. The results indicated that deep learning models outperformed support vector machines, neural networks, random forestsand convolutional neural networks.

A bar graph has been included below to illustrate the distribution of protocol types within a given dataset. Notably the dominant orange bar signifies a prevalence of one protocol type compared to the frequencies represented by blue and green bars. This graphical representation serves as a summary of protocol usage trends that mirror network traffic patterns or cybersecurity data, within the observed dataset.

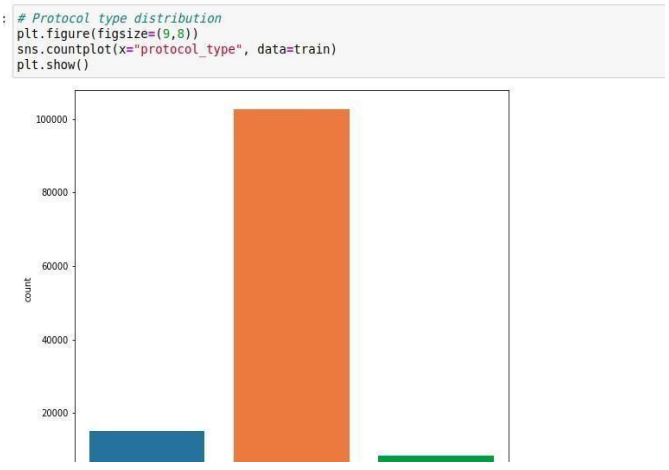

Fig: 2 Protocol Type Distribution

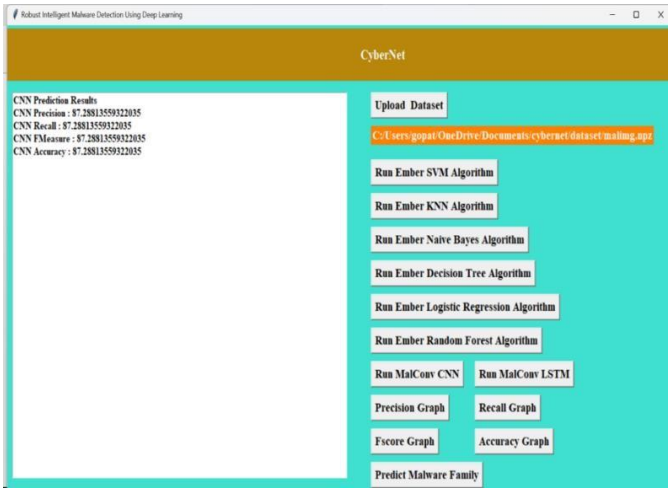

Fig: 3 Type of Attack Prediction

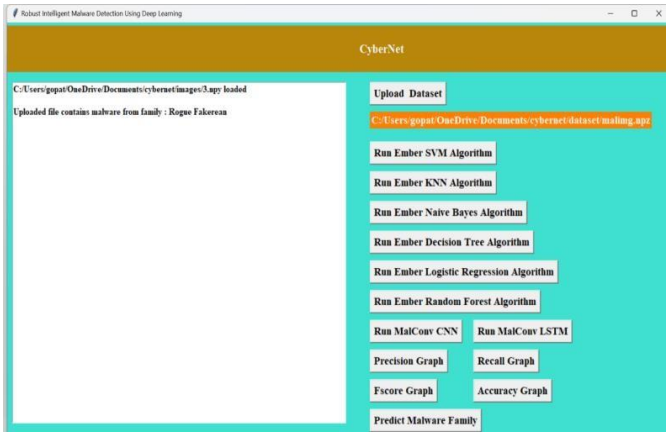

Fig: 3 File Dialogue Window

The two pictures display the user interface of 'CyberNet' a cybersecurity software. In the image labeled as Fig; 3 there are predictions of types of attacks shown on the CyberNet interface probably related to uploading datasets. The second image, in Fig; 3 exhibits a file dialog box and a confirmation message on CyberNet confirming a file upload. Each illustration represents a step in the process of inputting data into the system for analysis offering options to execute algorithms and visualize outcomes. This layout simplifies user engagement, for data handling and security assessment.

The accuracy of several machine learning methods in predicting the kind of attack is demonstrated by Predicting the type of Attack. Random Forest has the highest accuracy (99.93%), followed by CNN (99.11%), ANN (93.29%), and SVM (97.80%). This information can be used to select the best algorithm for the intrusion detection system.

5. CONCLUSIONS

In summary our research highlights the importance of utilizing machine learning methods to enhance network security in the face of growing cybersecurity threats. By conducting experiments and analyses we have showcased how SVM and various other algorithms can effectively detect and counter intrusions. These results emphasize the significance of embracing technologies to strengthen network defenses and prevent activities. Moving forward we intend to explore the integration of learning and artificial intelligence, with Spark and Apache Hadoop technologies to enhance the resilience of network security frameworks. Our goal is to leverage algorithms and predictive analytics for identification and mitigation of cyberattacks safeguarding critical assets and ensuring the integrity of network infrastructures.

We aim to utilize Apache Hadoop and Spark technologies for tasks like port scanning and defending against types of attacks alongside employing intelligence and deep learning approaches. These data insights are crucial for detecting cyber threats on networks. Reflecting on incidents reveals a history of attacks allowing us to document their characteristics in a database, for analysis. Leveraging this knowledge enables us to anticipate cyber threats effectively. The objective of this article is to determine the algorithm for precise identification and prediction of cyber-attacks.

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˘gi sistemlerinde kullanilan arac¸larin incelenmesi,” in 1st InternationalSymposium 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. 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] 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 910, September/October 2020, pp. 4055-4060(6) https://doi.org/10.1166/jctn.2020.9019.

[8] 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.

[9] 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

[10] 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.

[11] Farooq, Muhammad Junaid, et al. "A Survey of Machine Learning Techniques for Cybersecurity." IEEE Access 8 (2020): 145897-145924. doi:10.1109/ACCESS.2020.3011805.

[12] Reddy, Karan, et al. "Feature Selection Methods in Intrusion Detection Systems: A Review." 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2020. doi:10.1109/ICOEI48617.2020.9142727.