NETWORK INTRUSION DETECTION SYSTEM USING SUPERVISED MACHINE LEARNING WITH FEATURE SELECTION

B.V.N. PRAVEENA Assistant Professor,
V.Lakshmi Chetana Associate Professor,
B. KAVYA, K. YAMINI, P. DEVI KUMARI, CH. BHARGAV UG Student,
Department of Computer Science and Engineering, DVR&Dr.HSMICCollege ofTechnology (Autonomous), India. Email:kavyabolla73@gmail.com,

ABSTRACT

Advances in network technology have enabled the communication sector to connect remote parts of the world, but they have also led to a rise in attacks on networking infrastructure from intruders or attackers. By using intrusion detection tools and systems, system administrators can try to stop such attacks. Machine Learning (ML) methods have become more and more common in intrusion detection systems in recent years (IDS). Due to their high generalizability and capacity to escape the dimensionality curse, Support Vector Machines (SVM) and Artificial Neural Network Networks (ANN) have emerged as the most widely used machine learning (ML) methods for intrusion detection. According to several experts, the number of dimensions still has an impact on how well SVM-based IDS and ANN function. Another problem brought up is how SVM evaluates each data characteristic equally. Many features in actual intrusion detection datasets

1.INTRODUCTION

The first line of defence against a security breach is intrusion detection. As a result, studies are paying a lot of attention to security solutions including firewalls, intrusion detection systems, unified threat models, and intrusion prevention systems (IPS). IDS gather data, analyse it for potential security breaches, and then identify assaults from a number of systems and network sources. The network-based IDS performs two types of analyses on the data packets that move through a network. A key field of research continues to be anomaly-based detection because it is still far behind signature-based detection in terms of efficiency [4-5]. Intrusion detection is the first step to prevent security attack. As a result, studies are paying a lot of attention to security solutions including firewalls, intrusion detection systems, unified threat models, and intrusion prevention systems (IPS). IDS gather data, analyse it for potential security breaches, and then identify assaults from a number of systems and network sources [3]. The network-based IDS performs two types of analyses on the data packets that move through a network. A key field of research continues to be anomaly-based detection because it is still far behind signature-based detection in terms of efficiency [4-5]. Anomaly-based intrusion detection faces difficulties since it must handle fresh attacks for which there is no prior knowledge to recognise the abnormality. Hence Machine learning approaches have been investigated by researchers over the past several years to help the system distinguish between traffic that is benign and traffic that is malicious or abnormal. IDS, however, is not a solution to every security-related issue. IDS, for instance, cannot make up for inadequate means for identification and authentication or for weaknesses in the network protocols.

The first intrusion detection model was published in 1987 after the field's first studies on it began in 1980 [7]. Although there have been significant commercial investments and research over the past few decades, intrusion detection technology is still in its infancy and is therefore ineffective [7]. The adoption of anomaly-based network IDS by technology-based organisations around the world has not been as widespread as that of signature-based network IDS. As a result, research and development in the field of IDS are currently focusing heavily on anomaly-based detection [8].

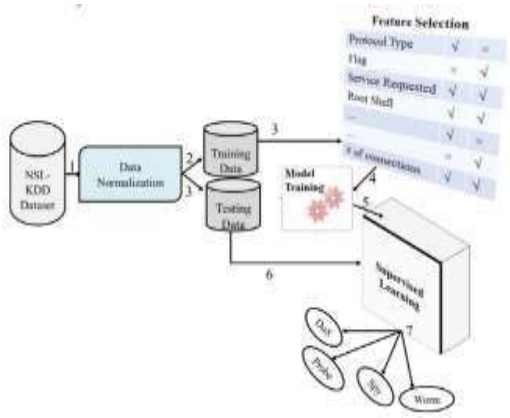
Moreover, significant problems need to be resolved before an anomaly-based intrusion detection system is widely deployed [8].But the literature today is limited when it comes to compare on how

 DOI: 10.36893.DRSR.2023.V13I04.001-007

intrusion detection performs when using supervised machine learning techniques [9]. To protect target systems andnetworks against malicious activities anomaly-based network IDS is a valuable technology. Despite the variety of anomaly-based network intrusion detection techniques described in the literature in recent years [8], anomaly detection functionalities enabled security tools are just beginning to appear, and some important problems remain to be solved. Several anomaly-based techniques have been proposed including Linear Regression, Support Vector Machines (SVM), Genetic Algorithm, Gaussian mixture model,nearestneighbour algorithm, Naive Bayes classifier, Decision Tree [3,5]. Among them the most widely used learning algorithm is SVM as it has already established itself on different types of problem [10]. One major issue on anomaly-based detection is though all these proposed techniques can detect novel attacks but they all suffer a high false alarm rate in general. The cause behind is the complexity of generating profiles of practical normal behaviour by learning from the training data sets [11]. Today Artificial Neural Network (ANN) are often trained by the back propagation algorithm, which had been around since 1970 as the reverse mode of automatic differentiation [12].

2.RELATED WORK

The major challenges in evaluating performance of network IDS is the unavailability of a comprehensive network-based data set [13]. Most of the proposed anomaly-based techniques found in the literature were evaluated using KDD CUP 99 dataset [14]. In this paper we used SVM and ANN –two machine learning techniques, on NSLKDD [15] which is a popular benchmark dataset for network intrusion.



3.LITERATURE SURVEY

1. Detecting Adversarial Examples Via Prediction Difference for Deep Neural Networks

To address this problem, we suggest a novel defence method termed transferability prediction difference (TPD), which considerably improves the adversarial robustness of DNNs while minimally degrading verified accuracy. The adversarial instances have more complicated decision boundaries, which leads to bigger prediction differences for various DNN models. These greater prediction differences can be used to detect adversarial cases via convergent decision bounds to a prediction difference threshold. We construct the transferability prediction difference threshold using benign data and the K-means clustering method to quickly and precisely identify adversarial samples. Furthermore, neither the target model alteration nor awareness of adversarial attacks are necessary for the TPD approach. We assess TPD models developed using four modern adversarial approaches (FGSM, BIM, JSMA, and C&W) to evaluate TPD models trained on MNIST and CIFAR-10 and the average detection accuracy is 96.74% and 86.61%. The results show that TPD model has high detection ratio on the demonstrably advanced white-box adversarial examples while keeping low false positive rate on benign examples.

 DOI: 10.36893.DRSR.2023.V13I04.001-007

2. A Data Mining Approach to Network Intrusion Detection

We proposed that a dimensionality reduction module and network traffic classification may be included in a machine learning technique. In order to boost learning rate and decrease false alarm rate, Osanaiye et al. emphasise the importance of feature selection in pre-processing by using information gain, gain ratio, and Chi-squared to pick essential characteristics. Onyekwelu et al. also discussed the significance of pre-processing employing several strategies for data transformation and data discretization. Garg promoted the use of the binarization approach on data; the best binarization algorithm resulted in varied performance on different datasets. Numerous intrusion detection tools, according to Ennert et al., are utilised in the IDS model to provide a set of chosen tests on the effectiveness of the IDS model that would expand the functionality of IDS. The results of this research have a poor detection rate because it is challenging to recognise unfamiliar signals. Due to the increased network traffic and low detection rate, machine learning techniques for intrusion detection are needed to address the problem of network invasions. Machine learning methods including Naive Bayes, Neural Networks, Fuzzy Logic, k Nearest Neighbour algorithms, Bagging, Random Forest, and many others are used to create an anomalous IDS. Despite its high rate of false alarms, an anomaly IDS is much more effective.

4.PROPOSED SYSTEM

Deep neural network algorithm is being used by the author to anticipate the System model in order to solve the issue with the current system. In order to compare the effectiveness of the various approaches, comparison research using the System model is presented. This study makes use of feature selection, machine intelligence development, and the mean square error criterion.

5.ALGORITHMS

In order to reduce the dimensionality of the data, feature selection is a crucial component of machine learning, and there has been substantial research into effective feature selection techniques. Filter method and wrapper method have both been utilised for feature selection. With the filter technique, features are chosen based on their results in several statistical tests that gauge their applicability by their correlation with the outcome or dependent variable.Wrapper method finds a subset of features by measuring the usefulness of a subset of feature with the dependent variable. Hence filter methods are independent of any machine learning algorithm whereas in wrapper method the best feature subset selected depends on the machine learning algorithm used to train the model. A subset evaluator employs all conceivable subsets in the wrapper technique before using a classification algorithm to persuade classifiers based on the features in each subset. The subset of features that the classification algorithm works best with are taken into account by the classifier. The evaluator uses a variety of search methods, including depth first search, random search, breadth first search, and hybrid search, to identify the subset. The filter technique ranks each feature in the dataset using an attribute evaluator and a ranker. Here, the classification algorithm's predicted accuracy is tested by removing one characteristic at a time from the dataset that has lower rankings. The weights or ranks inserted by ranker algorithms differ from those inserted by classification algorithms. The wrapper approach benefits machinelearning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

DOI: 10.36893.DRSR.2023.V13I04.001-007

Based on the best features found in the feature selection process, learning models are developed. To develop the learning model, machine learning algorithm is used. Training dataset is used to train the algorithm with the selected features. In supervised machine learning, each instance in the training dataset has the class it belongs to. The algorithm builds the learning model based on which machine learning algorithm is being used. Here we are using 2 different algorithms to find the accuracy of detection.
1.SVM (Support Vector Machine)
2.ANN (Artificial neural Network)

SVM (Support Vector Machine):
In SVM, a separating hyper plane specifies the classifier based on the kind of issue at hand and the datasets that are accessible. The hyper plane is a point for one-dimensional datasets, a separating line for two-dimensional datasets, as shown in Fig. 2, a plane for three-dimensional datasets, and a hyper plane if the data dimension is higher. The classifier or judgement function for a linearly separable dataset will take the form –

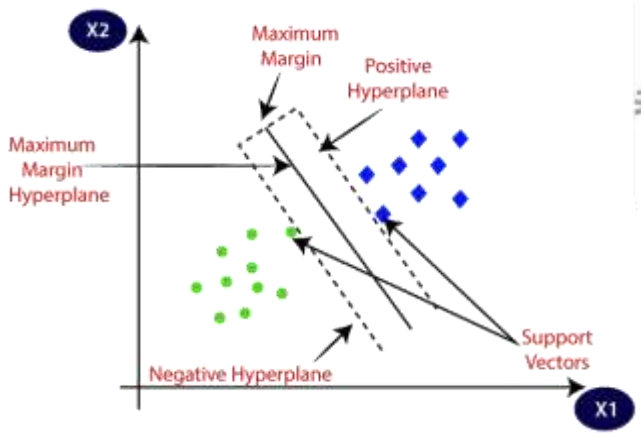


Fig 1 SVM

Biological Neural Network BNN	**Artificial Neural Network** ANN
Soma	Node
Dendrites	Input
Synapse	Weights or Interconnections
Axon	Output

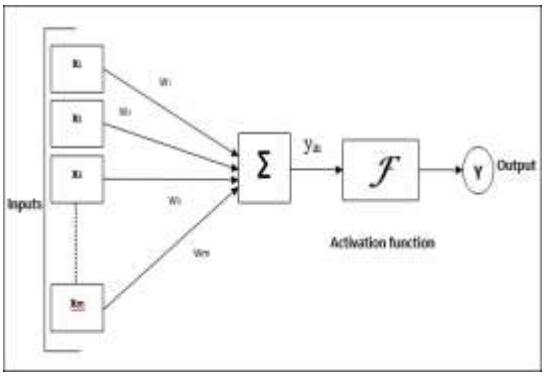
ANN (Artificial Neural Network):

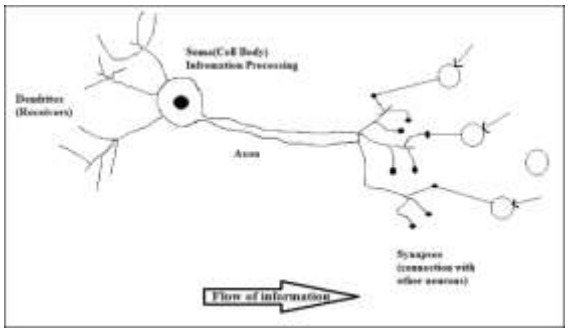
- ANN gathers information from a variety of sources, including system events, logs, and network traffic. The input layer of the ANN receives this pre-processed data. Based on the features that were collected from the data, the network then learns to distinguish between patterns of normal and pathological behaviour.
- Backpropagation, a well-liked technique for ANN training, is used to modify the weights of the connections between neurons in the hidden layers during training. Typical training material includes instances of both ordinary network traffic and different sorts of assaults.

- By analysing incoming network data in real-time once the network has been taught, it can be used to find fresh instances of network intrusions. The network's output layer generates a binary classification that indicates if the

 DOI: 10.36893.DRSR.2023.V13I04.001-007

Fig 2 ANN





For the above general model of artificial neural network, the net input can be calculated as follows −

$$yin=x1.w1+x2.w2+x3.w3\ldots xm.Wm$$

PREPROCESSING DATA

GENERATING TRAINING MODAL

i.e., net input yin=∑mixi. wi=∑\mixi. wi

The output can be calculated by applying the activation function over the net input.

$$Y=F(yin)$$

Output = function netinputcalculated

Technologies Used:

- Front-end: Python (GUI)
- Back-end: Python

Here we consider some accuracy from algorithms and by using an we give some keywords in dataset so based on that it compares and classifies the given into their respective domain.

 DOI: 10.36893.DRSR.2023.V13I04.001-007

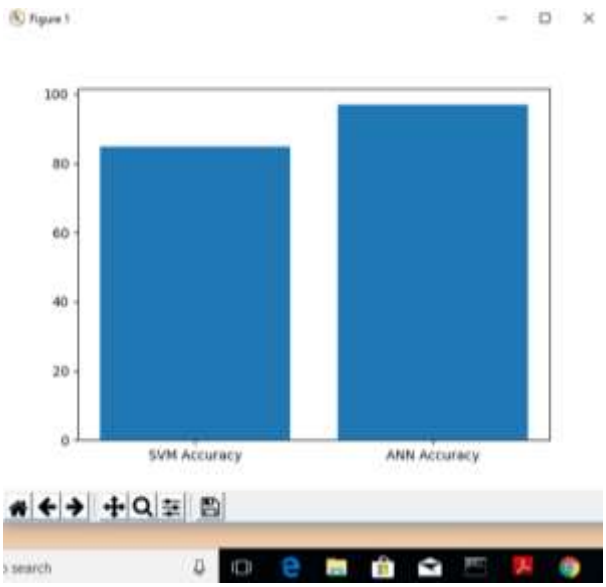
5.RESULTS AND DISCUSSION:

SAMPLE SCREENS

In above screen click on 'Upload NSL KDD Dataset' button and upload dataset

SVM ACCURACY 48.6%

ANN ACCURACY 98.6%



During pre-processing, all string values are eliminated, and attack names are converted to numeric values, such as "regular signature contains id 0" and "anomaly attack contains id 1".
Now select "Generate Training Model" to separate the train and test sets of data and create an SVM and ANN 6 prediction model.

 DOI: 10.36893.DRSR.2023.V13I04.001-007

CONCLUSION

In order to discover the optimum model, we have given various machine learning models in this work utilising various machine learning algorithms and feature selection techniques. With a detection rate of 94.02%, the model created utilising ANN and wrapper feature selection surpassed all other models in correctly identifying network traffic. We think that these discoveries will further scientific inquiry into the field of developing a detecting system that can detect known attacks as well as novel attacks. The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero-day attack still remains a research topic due to the high false positive rate of the existing systems.

7.REFERENCES

[1]H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization," American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.

[2]P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th International Conference on, 2017, pp. 178–184.

[3]Emharraf, and I. El Farissi, "Modelling and implementation approach to evaluate the intrusion detection system," in International Conference on Networked Systems, 2015, pp. 513–517.

[4]M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.

[5]A. S. Ashoor and S. Gore, "Importance of intrusion detection system (IDS)," International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.

[6]M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," arXivpreprint arXiv:1312.2177, 2013.

[7] V. S. Rao, V. Mounika, N. R. Sai and G. S. C. Kumar, "Usage of Saliency Prior Maps for Detection of Salient Object Features," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 819-825, doi: 10.1109/I-SMAC52330.2021.9640684

 DOI: 10.36893.DRSR.2023.V13I04.001-007