NETWORK INTRUSION USING SUPERVISED MACHINE LEARNING TECHNIQUE WITH FEATURE SELECTION

Mr. N. Karthik, Assistant Professor, CSE(AI & ML), Vaagdevi College of Engineering (Autonomous),Bollikunta,Khila Warangal(Mandal),Warangal Urban-506005(T.S), India
Mrs. G. VijayaLaxmi, Assistant professor, CSE(AI&ML), Vaagdevi College of Engineering (Autonomous), Bollikunta, Khila Warangal (Mandal), Warangal Urban – 506005(T.S), India
Shashidhar. A (20641A66E4), UG Student, CSE(AI &ML), Vaagdevi College of Engineering (Autonomous), Bollikunta,Khila Warangal(Mandal),Warangal Urban-506005(T.S,)India
TamkanathNazeela (20641A66C9), UG Student, CSE(AI &ML), Vaagdevi College of Engineering (Autonomous), Bollikunta,Khila Warangal(Mandal),Warangal Urban-506005(T.S), India
R. Poorna Chander (21645A6617), UG Student, CSE(AI &ML), Vaagdevi College of Engineering (Autonomous), Bollikunta,Khila Warangal(Mandal),Warangal Urban-506005(T.S), India
N. Lokesh Kumar (21645A6614), UG Student, CSE(AI &ML), Vaagdevi College of Engineering (Autonomous), Bollikunta,Khila Warangal(Mandal),Warangal Urban-506005(T.S), India

ABSTRACT
A novel supervised machine learning system is developed to classify network traffic whether it is malicious or benign. To find the best model considering detection success rate, combination of supervised learning algorithm and feature selection method have been used. Through this study, it is found that Artificial Neural Network (ANN) based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic. To evaluate the performance, NSL-KDD dataset is used to classify network traffic using SVM and ANN supervised machine learning techniques. Comparative study shows that the proposed model is efficient than other existing models with respect to intrusion detection success rate.

1. INTRODUCTION
With the wide spreading usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate [1-2]. Intrusion detection is the first step to prevent security attack. Hence the security solutions such as Firewall, Intrusion Detection System (IDS), Unified Threat Modeling (UTM) and Intrusion Prevention System (IPS) are getting much attention in studies. IDS detects attacks from a variety of systems and network sources by collecting information and then analyzes the information for possible security breaches [3]. The network based IDS analyzes the data packets that travel over a network and this analysis are carried out in two ways. Till today anomaly based detection is far behind than the detection that works based on signature and hence anomaly based detection still remains a major area for research [4-5]. The challenges with anomaly based intrusion detection are that it needs to deal with novel attack for which there is no prior knowledge to identify the anomaly.

2. LITERATURE SURVEY
This study examines whether macro-level opportunity indicators affect cyber-theft victimization. Based on the arguments from criminal opportunity theory, exposure to risk is measured by state-level patterns of internet access (where users access the internet). Other structural characteristics of states were measured to determine if variation in social structure impacted cyber-victimization across states [6]. The current study found that structural conditions such as unemployment and non-urban population are associated with where users access the internet. Also, this study found that the proportion of users who access the internet only at home was positively associated with state-level counts of cyber-theft victimization. The theoretical implications of these findings are discussed [7].
With the proliferation of the internet and increased global access to online media, cybercrime is also occurring at an increasing rate. Currently, both personal users and companies are vulnerable to

cybercrime. A number of tools including firewalls and Intrusion Detection Systems (IDS) can be used as defense mechanisms. A firewall acts as a checkpoint which allows packets to pass through according to predetermined conditions. In extreme cases, it may even disconnect all network traffic [8]. An IDS, on the other hand, automates the monitoring process in computer networks. The streaming nature of data in computer networks poses a significant challenge in building IDS. In this paper, a method is proposed to overcome this problem by performing online classification on datasets. In doing so, an incremental naive Bayesian classifier is employed. Furthermore, active learning enables solving the problem using a small set of labeled data points which are often very expensive to acquire. The proposed method includes two groups of actions i.e. offline and online. The former involves data preprocessing while the latter introduces the NADAL online method. The proposed method is compared to the incremental naive Bayesian classifier using the NSL-KDD standard dataset. There are three advantages with the proposed method: (1) overcoming the streaming data challenge; (2) reducing the high cost associated with instance labeling; and (3) improved accuracy and Kappa compared to the incremental naive Bayesian approach. Thus, the method is well-suited to IDS applications [9].

Intrusions detection systems (IDSs) are systems that try to detect attacks as they occur or when they were over. Research in this area had two objectives: first, reducing the impact of attacks; and secondly the evaluation of the system IDS [10]. Indeed, in one hand the IDSs collect network traffic information from some sources present in the network or the computer system and then use these data to enhance the systems safety. In the other hand, the evaluation of IDS is a critical task. In fact, its important to note the difference between evaluating the effectiveness of an entire system and evaluating the characteristics of the system components. In this paper, we present an approach for IDS evaluating based on measuring the performance of its components. First of all, in order to implement the IDS SNORT components safely we have proposed a hardware platform based on embedded systems [11]. Then we have tested it by using a generator of traffics and attacks based on Linux KALI (Backtrack) and Metasploite 3 Framework. The obtained results show that the IDS performance is closely related to the characteristics of these components [12].

3. PROBLEM STATEMENT

IDS will be trained with all possible attacks signatures with machine learning algorithms and then generate train model [13], whenever new request signatures arrived then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating performance of two machine learning algorithms such as SVM and ANN and through experiment we conclude that ANN outperform existing SVM in terms of accuracy.

4. PROPOSED SYSTEM

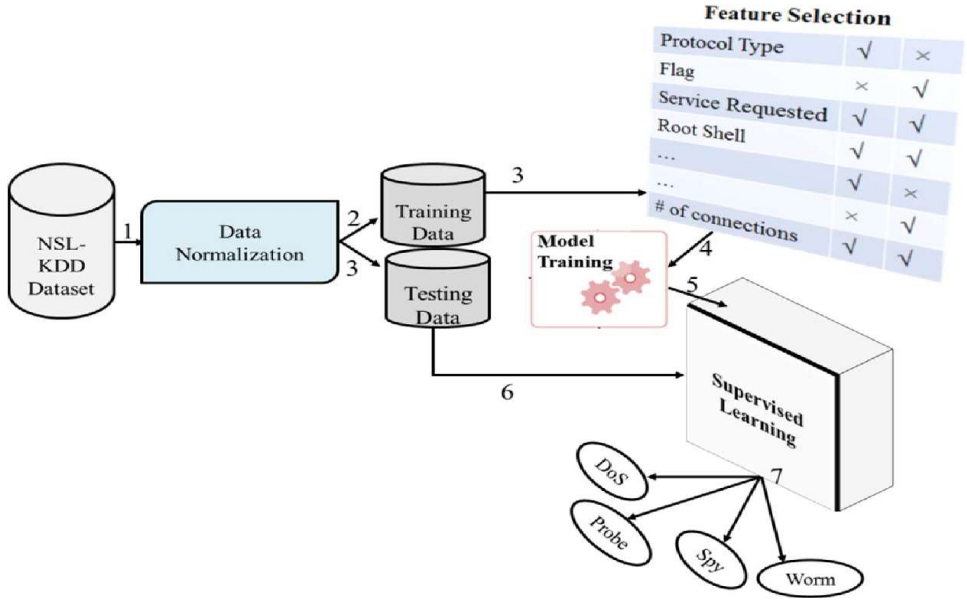
In this paper author is evaluating performance of SVM and ANN.

In this algorithms author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size, this feature selection algorithms removed irrelevant data from dataset and then used model with important features, due to this features selection algorithms dataset size will reduce and accuracy of prediction will increase [14].

To conduct experiment author has used NSL KDD Dataset and below is some example records of that dataset which contains request signatures. I have also used same dataset and this dataset is available inside ‘dataset’ folder [15].

ISSN: 2278-4632
Vol-14, Issue-5, May: 2024

5. SYSTEM ARCHITECTURE



6. IMPLEMENTATION

1.Feature Selection

Feature selection is an important part in machine learning to reduce data dimensionality and extensive research carried out for a reliable feature selection method. For feature selection filter method and wrapper method have been used. In filter method, features are selected on the basis of their scores in various statistical tests that measure the relevance of features by their correlation with dependent variable or outcome 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. In wrapper method a subset evaluator uses all possible subsets and then uses a classification algorithm to convince classifiers from the features in each subset. The classifier consider the subset of feature with which the classification algorithm performs the best. To find the subset, the evaluator uses different search techniques like depth first search, random search, breadth first search or hybrid search. The filter method uses an attribute evaluator along with a ranker to rank all the features in the dataset. Here one feature is omitted at a time that has lower ranks and then sees the predictive accuracy of the classification algorithm. Weights or rank put by the ranker algorithms are different than those by the classification algorithm. Wrapper method is useful for machine learning test whereas filter method is suitable for data mining test because data mining has thousands of millions of features.

2. Building Machine Intelligence

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.

3.Support Vector Machine (SVM)

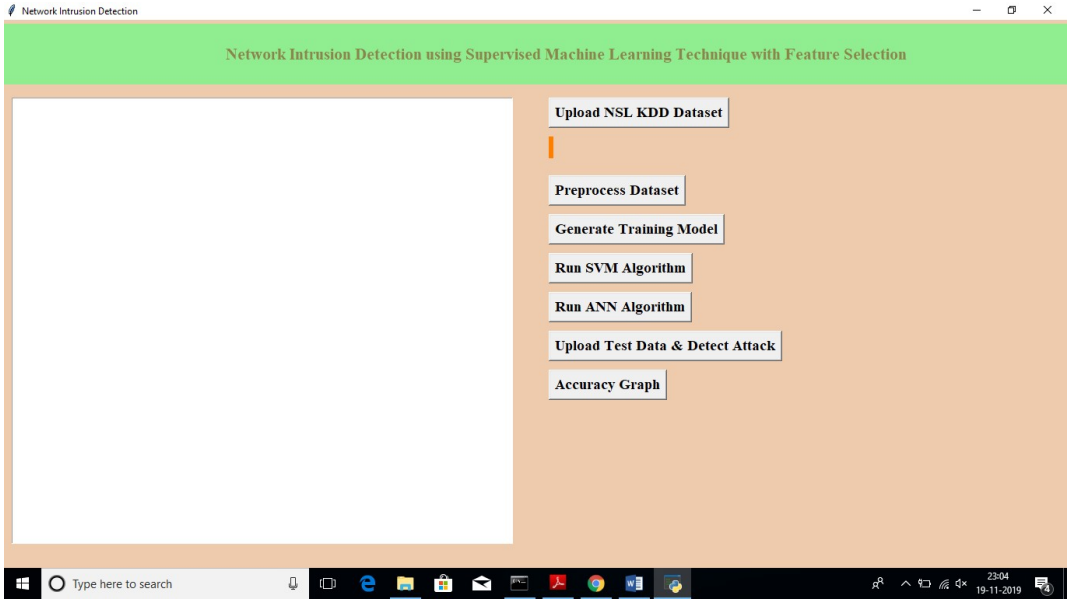
In SVM a separating hyper plane defines the classifier depending on the type of problem and available datasets. In case where dataset is one dimensional, the hyper plane is a point, for two dimensional data it is a separating line as shown in Fig 2, for three dimensional dataset, it is a plane and if the data dimension is higher it is a hyper plane. For a linearly separable dataset, the classifier or the decision function will have the form.

ISSN: 2278-4632
Vol-14, Issue-5, May: 2024

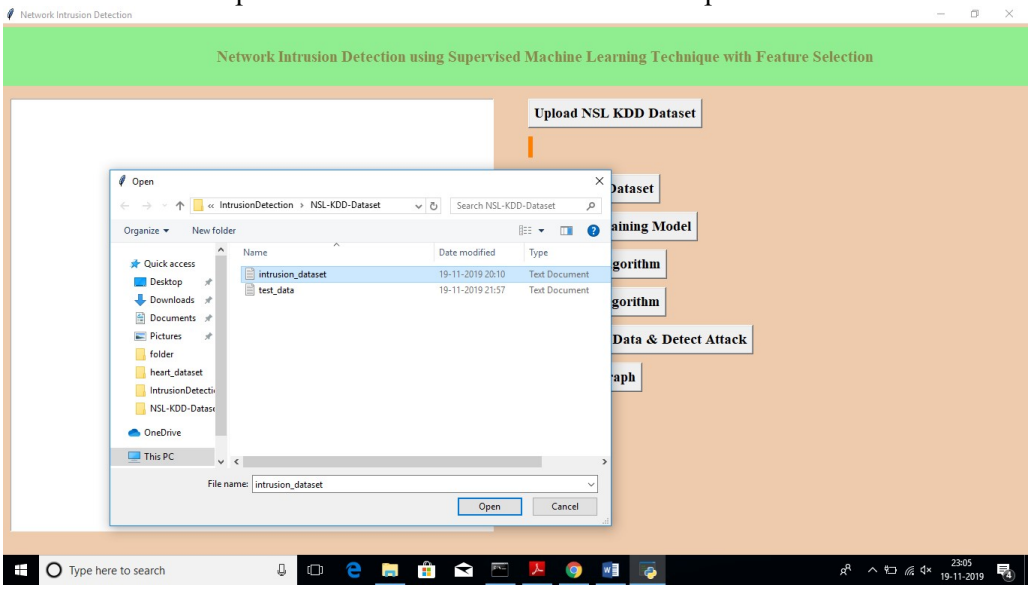
4. Artificial Neural Network (ANN)

Artificial Neural Network is another tool used in machine learning. As it name suggests, ANN is a system inspired by human brain system and replicate the learning system of human brain. It consists of input and output layers with one or more hidden layers in most cases as shown in Fig 3. The ANN uses a technique called back propagation to adjust the outcome with the expected result or class.

7. OUTPUT RESULTS



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



In above screen I am uploading 'intrusion_dataset.txt' file, after uploading dataset will get below screen

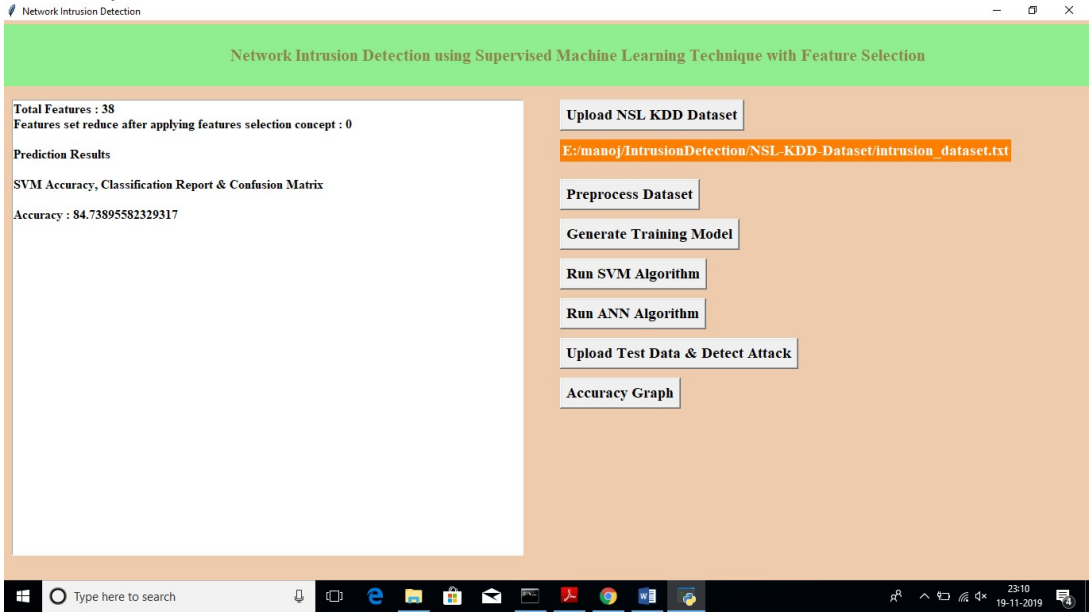
ISSN: 2278-4632
Vol-14, Issue-5, May: 2024

Now click on 'Pre-process Dataset' button to clean dataset to remove string values from dataset and to convert attack names to numeric values

After pre-processing all string values removed and convert string attack names to numeric values such as normal signature contains id 0 and anomaly attack contains signature id 1.
Now click on 'Generate Training Model' to split train and test data to generate model for prediction using SVM and ANN

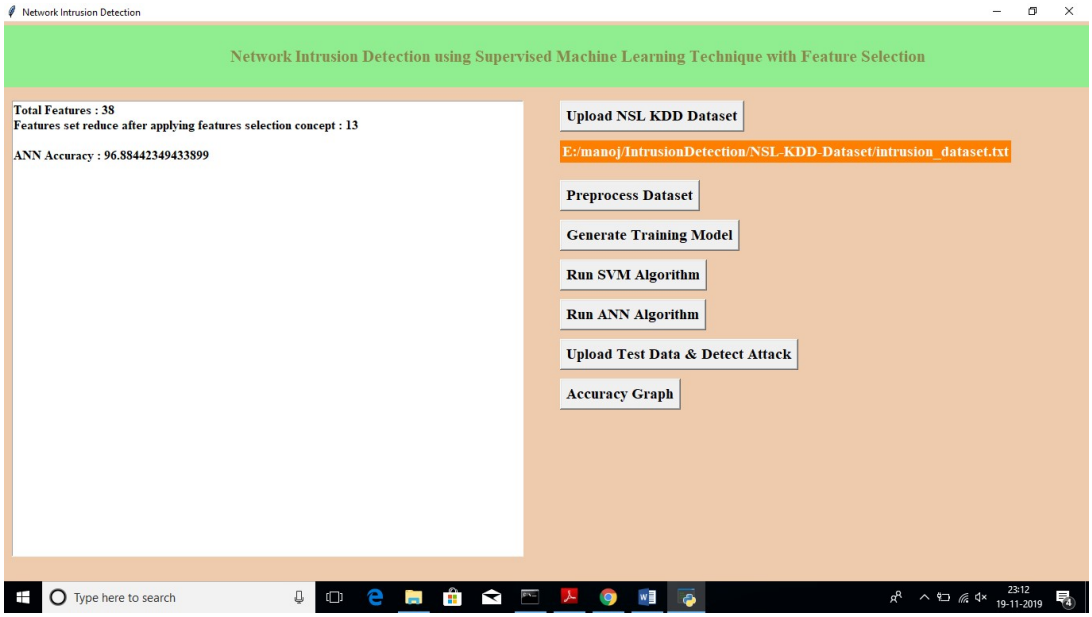
ISSN: 2278-4632
Vol-14, Issue-5, May: 2024

In above screen we can see dataset contains total 1244 records and 995 used for training and 249 used for testing. Now click on 'Run SVM Algorithm' to generate SVM model and calculate its model accuracy



In above screen we can see with SVM we got 84.73% accuracy, now click on 'Run ANN Algorithm' to calculate ANN accuracy

ISSN: 2278-4632
Vol-14, Issue-5, May: 2024



8. CONCLUSION

In this paper, we have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%. We believe that these findings will contribute to research further in the domain of building a detection 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

9. FUTURESCOPE

To avoid all attacks IDS systems has developed which process each incoming request to detect such attacks and if request is coming from genuine users then only it will forward to server for processing, if request contains attack signatures then IDS will drop that request and log such request data into dataset for future detection purpose.

10. 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. M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, “Modeling 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," arXiv preprint arXiv:1312.2177, 2013.
7.N. Chakraborty, "Intrusion detection system and intrusion prevention system: A comparative study," International Journal of Computing and Business Research (IJCBR) ISSN (Online), pp. 2229–6166, 2013.
8. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security, vol. 28, no. 1–2, pp. 18–28, 2009.
9. M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," Procedia Computer Science, vol. 89, pp. 117–123, 2016.
10. J. Zheng, F. Shen, H. Fan, and J. Zhao, "An online incremental learning support vector machine for large-scale data," Neural Computing and Applications, vol. 22, no. 5, pp. 1023–1035, 2013.
11. F. Gharibian and A. A. Ghorbani, "Comparative study of supervised machine learning techniques for intrusion detection," in Communication Networks and Services Research, 2007. CNSR'07. Fifth Annual Conference on, 2007, pp. 350–358.
12. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85–117, 2015.
13. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Military Communications and Information Systems Conference (MilCIS), 2015, 2015, pp. 1–6.
14. T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," in Industrial Electronics (ISIE), 2017 IEEE 26th International Symposium on, 2017, pp. 1881–1886.
15. L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446–452, 2015.

