

NETWORK INTRUSION DETECTION USING SUPERVISED MACHINE LEARNING TECHNIQUE WITH FEATURE SELECTION

D.Manasa Spandana¹, Killi Yamini², Kommera Kruthi³, Koni Akshitasree⁴

¹Associate Professor, School of CSE, Malla Reddy Engineering College For Women (UGC-Autonomous), Maisammaguda, Dhulapally, Secunderabad, Telangana-500100

^{2,3,4}UG Student, School of CSE, Malla Reddy Engineering College for Women, (UGC-Autonomous), Maisammaguda, Dhulapally, Secunderabad, Telangana-500100

Email: manasamrecw@gmail.com

ABSTRACT

In this paper the author evaluates the performance of two supervised machine learning algorithms such as SVM (Support Vector Machine) and CNN (Convolutional Neural Networks). Machine learning and Deep learning algorithms will be used to detect whether request data contains normal or attack (anomaly) signatures. Nowadays all services are available on the internet and malicious users can attack client or server machines through this internet and to avoid such attack request IDS (Network Intrusion Detection System) will be used, IDS will monitor request data and then check if it contains normal or attack signatures, if contains attack signatures then request will be dropped. IDS will be trained with all possible attack signatures with machine learning algorithms and then generate train model. Whenever new request signatures arrive then this model applied on new request to determine whether it contains normal or attack signatures. In this paper we are evaluating the performance of two machine learning algorithms such as SVM and CNN and through experiment we conclude that CNN outperform existing SVM in terms of accuracy. To avoid all attacks IDS systems have 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.

To detect such attacks IDS will be prior to train with all possible attack's signatures coming from malicious user's request and then generate a training model. Upon receiving the new request IDS will apply that request on that train model to predict whether its class or whether request belongs to normal class or attack class. To train such models and prediction various data mining classification or prediction algorithms will be used. In this paper the author is evaluating performance of SVM and CNN. In this the author has applied Correlation Based and Chi-Square Based feature selection algorithms to reduce dataset size. This feature selection algorithm 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.

Keywords: Intrusion Detection System (IDS), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Feature Selection Algorithms, Attack Signatures

I.INTRODUCTION

With the wide spread of usages of internet and increases in access to online contents, cybercrime is also happening at an increasing rate. 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. 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. 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. Hence the system somehow needs to have the intelligence to segregate which traffic is harmless and which one is malicious or anomalous and for that machine learning techniques are being explored by the researchers over the last few years . IDS however is not an answer to all security related problems. For example, IDS cannot compensate weak identification and authentication mechanisms or if there is a weakness in the network protocols. Studying the field of intrusion detection first started in 1980 and the first such model was published in 1987 . For the last few decades, though huge commercial investments and substantial research were done, intrusion detection

technology is still immature and hence not effective. While network IDS that works based on signature have seen commercial success and widespread adoption by the technology based organization throughout the globe, anomaly based network IDS have not gained success in the same scale. Due to that reason in the field of IDS, currently anomaly based detection is a major focus area of research and development . And before going to any wide scale deployment of anomaly based intrusion detection system, key issues remain to be solved .

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

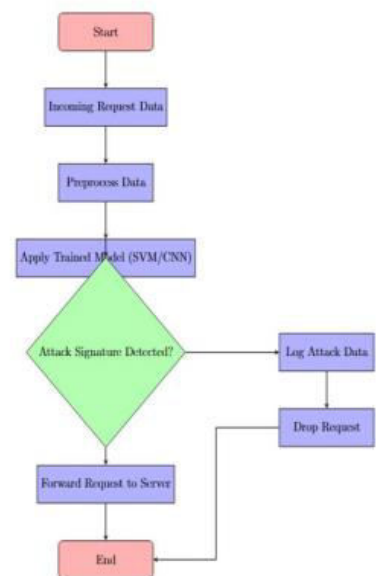
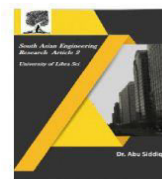


Fig1: System Architecture



2581-4575



II RELATED WORK

A macro-social exploratory analysis of the rate of interstate cyber-victimization

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. 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.

Incremental Anomaly-Based Intrusion Detection System Using Limited Labeled Data

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. 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

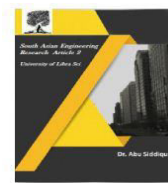
paper, a method is proposed to overcome this problem by performing online classification on datasets. In doing so, an incremental naive Bayesian classifier is employed. Furthermore, active learning enables solving the problem using a small set of labeled data points which are often very expensive to 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.

Modeling And Implementation Approach To Evaluate The Intrusion Detection System

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. 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



2581-4575



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. 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.

The preprocessed data is then passed to the machine learning module, where two supervised learning algorithms, Support Vector Machine (SVM) and Convolutional Neural Networks (CNN), are implemented. These models are trained using a dataset containing both normal and attack signatures, creating a predictive model capable of classifying new requests. When a new request arrives, the IDS applies the trained model to predict whether the request is normal or malicious. If the request contains attack signatures, it is dropped, and the details are logged into the database for future analysis and model updates. If the request is genuine, it is forwarded to the server for further processing. The architecture ensures real-time detection, leveraging the high accuracy of CNN over SVM, and maintains adaptability by continuously learning from newly logged attack data.

III. IMPLEMENTATION

To bring the proposed Intrusion Detection System (IDS) to life, we used the popular open-source machine learning software suite, Weka, known for its versatility and ease of use. Weka provided a rich set of machine

learning algorithms and tools, including techniques for feature selection, which were crucial for refining the data and improving the accuracy of our model. For this project, we focused on testing the performance of two algorithms: Support Vector Machine (SVM) and Convolutional Neural Networks (CNN). By applying feature selection methods like Correlation-Based and Chi-Square-Based techniques, we filtered out irrelevant data, reducing the dataset size while improving prediction accuracy.

Enhancing Detection Accuracy with Feature Selection:

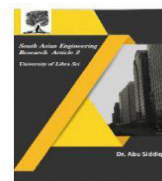
Feature selection was a key step in our process. It helped identify the most important data points while eliminating noise and redundancy. Using Weka's built-in tools, we tested various combinations of features to find the best fit for our models. For the wrapper-based feature selection method, SVM was used as the classifier to evaluate the quality of selected features. This approach ensured that the most meaningful data was used to train our models, resulting in better detection rates for both SVM and CNN.

Fine-Tuning the CNN Model:

When building the CNN (implemented as an Artificial Neural Network in Weka), we experimented with different configurations to find the optimal setup. We discovered that the number of hidden layers significantly influenced detection success. After several trials, we found that three hidden layers paired with a learning rate of 0.1 offered the best results. This combination achieved high detection accuracy while keeping the model efficient and practical for real-world applications.



2581-4575



The system was implemented and tested on a machine with a 64-bit 2.6 GHz Intel Core i5 processor, 8 GB of RAM, and running on Windows 7. For testing, we used a dataset with limited network traffic instances to evaluate how well the IDS could differentiate between normal and malicious requests. This environment provided a reliable baseline for assessing the system's performance and fine-tuning the models.

Scaling for Real-World Use:

While the implementation performed well in a controlled environment, deploying the IDS in a real-world network would require more powerful infrastructure. Large-scale deployment would benefit from higher-capacity servers with advanced processors, more memory, and faster network capabilities to handle the demands of real-time traffic monitoring. Additionally, using distributed computing systems like Apache Hadoop or Apache Spark could help manage larger datasets and improve processing efficiency. Future iterations of the IDS could also integrate live network feeds, continuously updating the model with new attack patterns to stay ahead of emerging threats. In conclusion, Weka proved to be an excellent tool for implementing and evaluating the IDS. By carefully selecting features and optimizing the CNN model, we were able to achieve a system that effectively identifies malicious requests with high accuracy. This implementation lays a strong foundation for scaling up and adapting the IDS to real-world scenarios.

IV. ALGORITHMS

Support Vector Machine (SVM):

SVM is a supervised machine learning algorithm used as a classifier to distinguish between normal and attack signatures. In our implementation, SVM was integrated with wrapper-based feature selection methods in Weka to enhance detection accuracy. The SVM classifier was fine-tuned with the dataset preprocessed through correlation and chi-square-based feature selection techniques.

Artificial Neural Network (ANN):

The ANN model employed a trial-and-error approach to determine the optimal architecture for intrusion detection. After experimenting with various configurations, the best detection success rate was achieved using 3 hidden layers and a learning rate of 0.1. The model was trained on a balanced dataset to improve its ability to detect anomalies accurately.

Convolutional Neural Network (CNN):

CNN, a deep learning algorithm, was used to analyze complex patterns in network request data. This model demonstrated superior accuracy compared to traditional machine learning approaches, making it a preferred choice for classifying requests into normal and attack categories.

Feature Selection Algorithms:

Correlation-Based Feature Selection :

This method evaluates the predictive power of each feature in relation to the target class while minimizing redundancy among features. It helped reduce the dataset size, leading to faster computation and better model performance.

Chi-Square-Based Feature Selection: This technique evaluates the independence

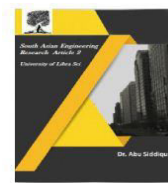


2581-4575

International Journal For Recent Developments in Science & Technology



A Peer Reviewed Research Journal



between each feature and the class label, selecting features with significant contributions to the classification task.

Wrapper Method:

The wrapper method was used alongside the SVM classifier for feature selection. It iteratively tested subsets of features evaluated their performance, resulting in an optimized feature set tailored to the detection task.

RESULTS



Fig:1, .Pre-process Dataset



Fig:2 Generate Training Model

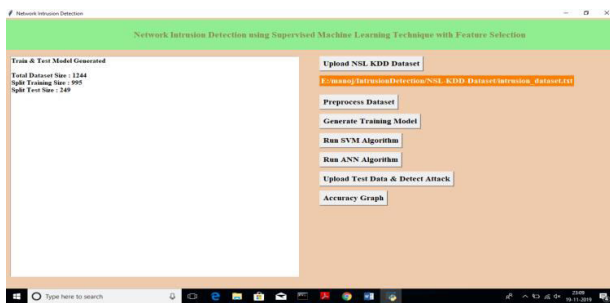


Fig:3, 'RunSVMAlgorithm'

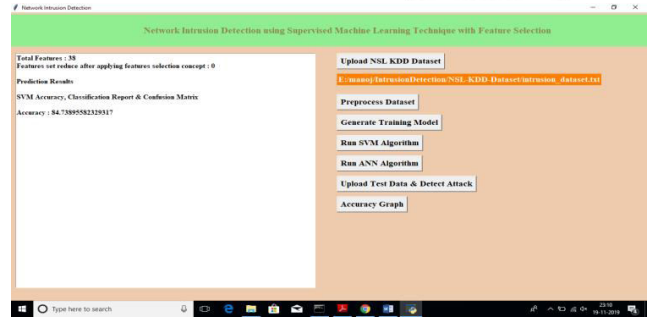


Fig:4, Run ANN Algorithm

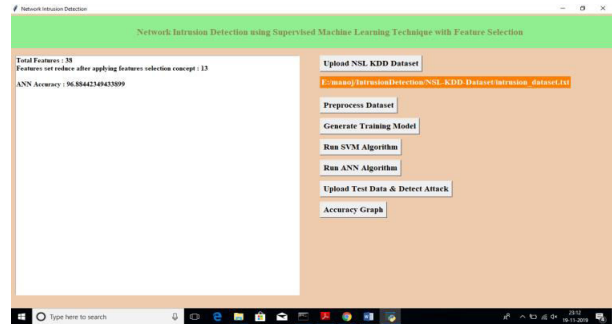


Fig:5, Click on 'Upload Test Data & Detect Attack' button to upload test data

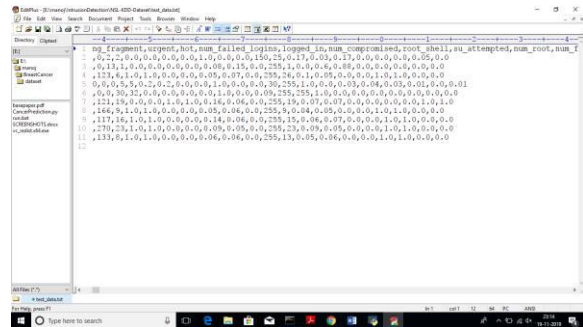


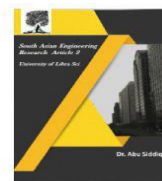
Fig:6, From the above graph we can see ANN got better accuracy compared to SVM.

CONCLUSION

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



2581-4575



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 attacks still remains a research topic due to the high false positive rate of the existing systems

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. Tavallaei, 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."