Concept Analysis And Dynamic Projection Intrusion Detection Using Ensemble Learning

Mrs.Jeyalakshmi P[1], Ajay Arjunaa S[2], Cyril Jeyakumar G[3], Sai Kiran S[4]
[1]Assistant Professor, Dept of computer science and Engineering
[2, 3, 4]Dept of computer science and Engineering
[1, 2] SRM Institute of Science and Technology Chennai, India

Abstract- *Intrusion detection systems (IDS)plays an important role in network security by monitoring network traffic for malicious activity and detecting the exploitation of vulnerabilities against targeted applications and computers. A large number of redundant and irrelevant functions increases the dimensionality of the dataset, increases the computational load of the system and degrades performance. This article describes various filter-based feature selection techniques to improve system performance. It uses a selected well-performing subset of features, followed by an ensemble learning technique that combines multiple feature subsets to select the best feature subset.In conclusion, we examine feature selection combined with ensemble learning.*

In conclusion, we examine feature selection combined with ensemble learning.

Performance in algorithms implemented in existing studies in terms of accuracy, false alarm rate, and true positive rate, and observe their shortcomings.

Keywords- Intrusion Detection System (IDS), Machine learning (ML), wireless sensor networks (WSNs).

I. INTRODUCTION

Applications in business, research, healthcare and human lives are largely reliant on wired and wireless computer networks and their priceless data is sent across the internet every second. Unfortunately, present solutions like user authentication, hardware and software firewalls and data encryption techniques are unable to meet the challenge of rising demand and are unable to prevent the numerous cyber threats to computer networks. The development of efficient approaches and tools to stop service disruption, unauthorized access and the leaking of sensitive information is motivated by the volume and sophistication of network-based attacks.

An Intrusion Detection System (IDS) is a crucial defence tool against complex and more common network threats, but in order to be effective, especially Machine Learning (ML) based systems, need a sizable amount of valid and trustworthy network traffic datasets. Modern networks are becoming more diversified and even though the majority of recently available datasets cover a variety of network attack types, traffic patterns, and information about the attacking infrastructure, they frequently insufficient to create useful classification mechanisms.

SNIDS (Signature-based Intrusion Detection Systems) and ANIDS (Anomaly-based Intrusion Detection Systems) are two IDS techniques. The SNIDS technique is successful against known threats because it searches for particular patterns or signatures, such as known dangerous instructions sequences used by malware or bytes in network traffic. In contrast, the ANIDS methodology use ML Algorithms to analyse and track network data in order to spot any strange activity, making it a useful technique for spotting unidentified attacks.

A Class of Deep Reinforcement Learning (DRL) algorithms that are able to detect the latest and sophisticated types of network attacks has been developed as a result of the development of deep learning and its combination with Reinforcement Learning (RL). Software agents, also known as learning entities, can learn how to accomplish their objectives with the aid of DRL, which blends artificial neural networks with an RL framework. DRL combines target optimization and function approximation to map states and actions to the rewards they produce. Our learning agents can use this policy to guide their decisions in light of the present situation. DRL is used to train an agent to respond appropriately to a state represented as a collection of feature values for detecting network assaults.

Every network has distinctive characteristics in its patterns and methods change over time. Of course, vulnerabilities change as well. As the validity and reliability of the available datasets continuously deteriorate, IDS classification accuracy fails to improve. Furthermore, privacy issues frequently prevent the sharing of trustworthy data. Even the unknown vulnerabilities and assaults are not entirely represented in the publicly accessible databases for network attacks.

II. RELATED WORKS

In this section, we first describe cyber-security, which is defined as the systems or software that protects data, programmes and connections between computers from a variety of undesirable attacks such as unauthorized attacks, alteration, and fabrication. We concentrate on creating an intrusion detection system (IDS) to investigate and discover the system's security because traditional security methods are insufficient for detecting network security.

Unauthorized activity that harms an information system is referred to as intrusion. This means that any attack that might endanger the confidentiality, integrity, or availability of information is regarded as an intrusion. Presently, the primary defensive tools used to stop invasions and identify internal attacks are firewalls, access control, and cryptography. However, both internal and external threats are detected by intrusion detection systems. In this study, we want to focus on an anomaly-based intrusion detection model in spite of detecting known assaults on the signature-based IDS outlined above. A situation of divergence from customary action is referred to as an abnormality. Profiles are the general or desired action that are gleaned from monitoring user, network connection, and host activities over a predetermined period of time.

Computers are used in machine learning to make decisions. It is associated with computational statistics and a subset of artificial intelligence. Classification is supervised learning that uses training security data to predict the cyber-attack class labels of samples. In order to create a successful data driven IDS predictive model for delivering intelligent services of cyber security, we analyse a variety of widely used classification techniques, such as Bayesian approach, Tree-based model, and Artificial Neural Network based model.

With the dramatic rise in internet usage, network security has emerged as one of the issues that both internet users and service providers are most concerned about. A secure network is one that is protected against various invasions by both its hardware and software. By putting in place robust monitoring, analysis, and defence measures, a network can be made secure. Network intrusion detection systems (NIDS) are a class of systems that employ these defence techniques to protect networks from intrusions from both insiders and outsiders. These systems keep an eye on the network’s incoming and outgoing traffic, run periodic analyses, and alert users when an intrusion is found. Misuse detection (MD) and Anomaly Detection (AD) are two broad categories for NIDS. The signatures or patterns of current assaults are used by MD-based NIDS.

III. PROPOSEDMETHOD

The predicted decrease in entropy following the split is measured by information gain if the training data is divided based on the values of this feature. Thus, the ability to categories samples more accurately is a property of greater information gain features.

1)DATA PRE-PROCESSING

For various data sources and machine learning models, various methods of feature pre-processing are needed. There are some pre-processing techniques that apply to all data types. A technique for normalising the variety of independent variables or features in data is called feature scaling. It is frequently called "normalisation." Non-tree-based models are more affected by feature scaling than are tree-based models. Thus, normalising your numerical features is something you shouldthink about doing if you want to get decent results using a non-tree-based model.

2)MACHINE LEARNING MODEL TRAINING

In order to reduce the feature space to the smallest size possible depending on some criteria, feature selection is a technique that chooses a subset of the underlying features. The process of extracting new features that can be used separately or in combination is known as feature extraction. Additionally, it has the ability to find and select the data's considerably more advantageous qualities. It's a crucial step in the learning process since it helps to reduce the number of fitting issues, the effectiveness of adaptation on test data, the length of training, and the interpretability of the model. Filter-based feature selection, wrapper-based feature selection and embedded feature selection are the three basic categories of feature selection methods.
The embedded feature selection approach includes built-in feature selection, which aids in the development of models.

3)MACHINE LEARNING MODEL TRAINING

Finding the ideal set of parameters that we set our algorithm in order to get the best accuracy metrics is known as hyper parameter tuning. In general, grid search and random search are the two methods employed for this purpose. In order to identify the optimal values for the model, the grid search approach evaluates every potential list of values in every combination, and the random search method tests random combinations of parameters. In this study, we used a single decision tree experiment to determine the optimum hyper parameters.

IV. RESULTS AND EVALUATION

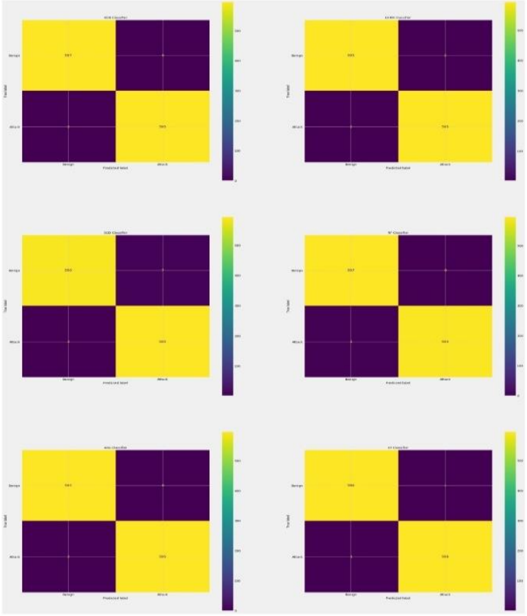
In this Section we will cover the outcome of the proposed and evaluate to test the accuracy of the system.

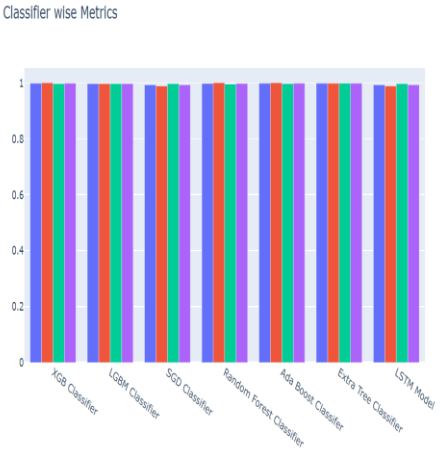
A. EVALUATION METRIC OF INTEREST

To assess the effectiveness of our DRL model and other ML algorithms, we utilized the accuracy and F1-score metrics, which combine precision and recall. This choice of performance metrics enables us to assess the percentage of samples that were mistakenly classified, whereas the accuracy score only evaluates the percentage of samples that are correctly classified. This is crucial for NIDS in particular since imbalanced datasets like network traffic data, which often contains much more regular traffic, cannot be evaluated using the accuracy performance indicator. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) data are used to calculate these performance indicators. This confusion matrix is shown and it is employed by our evaluation approach.

B. TRAINING MODEL ACCURACY

Out of all the predictions generated by the model, accuracy counts how many were accurate. Accuracy here refers to how well a model can distinguish between normal and malicious traffic records. Precision indicates the model's degree of accuracy in predicting attack records over the overall number of attacks predicted. In this scenario, precision measures the number of right positive predictions out of the total number of positive predictions.





V. CONCLUSION

The study analyses datasets created in the Intrusion Detection System (IDS) field. These datasets have been used to assess how well the EL and DM based IDS perform. The study showed that the underlying dataset has to be updated in order to better identify recent assaults in the field of IDS. This is because different methods and technologies are used by the attackers to carry out attacks. Additionally, the method of carrying out various attacks replicates the requirement for datasets with accurate network circumstances. CSE-CIC-IDS on AWS datasets have been introduced to meet the demand of generating an intrusion detection dataset with realistic network traffic and current network assaults. The properties of these datasets are reviewed in this work, along with some of their drawbacks.

REFERENCES

[1] P. Sinha, V. K. Jha, A. K. Rai, and B. Bhushan, “Security vulnerabilities attacks and countermeasures in wireless sensor networks at various layers of os+ reference model: A survey,” in Proc. IEEE ICSPC, 2017

[2] Y. Maleh, A. Ezzati, Y. Qasmaoui, and M. Mbida, “A global hybrid intrusion detection system for wireless sensor networks,” Procedia Comput. Sci., vol. 52, pp. 1047–1052, 2015.

[3] E. Kabir, J. Hu, H. Wang, and G. Zhuo, “A novel statistical technique for intrusion detection systems,” Future Gener. Comput. Syst., vol. 79, pp. 303–318, 2018.

[4] L. Fernandez Maimo et al., “A self-adaptive deep learning-based system for anomaly detection in 5G networks,” IEEE Access, vol. 6,pp. 7700–7712, 2018.

[5] M. Lopez-Martin, B. Carro, J. I. Arribas, and A. SanchezEsguevillas,“Network intrusion detection with a novel hierarchy of distances between embeddings of hash IP addresses,” Knowledge-based Syst., vol. 219, 2021

[6] Z. Chen, F. Han, L. Wu, J. Yu, S. Cheng, P. Lin, and H. Chen, “Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents,” Energy Convers. Manage., vol. 178, pp. 250–264,2020.

[7] F. Zhang, Y. Wang, S. Liu, and H. Wang, “Decisionbased evasion attacks on tree ensemble classifiers,” World Wide Web, vol. 23, no. 5,pp. 2957– 2977, 2020.

[8] T. M. C. J. W. H. Y. M. Yin, J. and Y. Lin, “Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning,” World Wide Web, pp. 1–23, 2021.

[9] S. Huang, Y. Lu, W. Wang, and K. Sun, “Multi-scale guided feature extraction and classification algorithm for hyperspectral images,” Scientific Reports, vol. 11, no. 1, 2021.

[10] R. . Chen, C. Dewi, S. . Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” J.Big Data, vol. 7, no. 1, 2020.