AN EFFICIENT FEATURE SELECTION MODEL FOR CYBER THREAT INTELLIGENCE USING DNN MODEL TO DETECT ABNORMAL NETWORK BEHAVIOR

J. Syambabu, **Mr. M M RAYUDU** M.Tech(Ph.D) Associate professor

Department of computer science engineering Prakasam Engineering College, Kandukur, Prakasam District, Andhra Pradesh

Abstract:

Network security and data security are the biggest concerns now a days. Every organization decides their future business process based on the past and day to day transactional data. This data may consist of consumer’s confidential data, which needs to be kept secure. Also, the network connections when established with the external communication devices or entities, a care should be taken to authenticate these and block the unwanted access. This consists of identification of the malicious connection nodes and identification of normal connection nodes. For that, we use a continuous monitoring of the network input traffic to recognize the malicious connection request called as intrusion and this type of monitoring system is called as an Intrusion detection system (IDS). IDS helps us to protect our network and data from insecure and malicious network connections. Many such systems exists in the real time scenario, but they have critical issues of performance like accuracy and efficiency. These issues are addressed as a part of this research work of IDS using machine learning techniques. The Proposed IDS is designed for increasing accuracy by using select k-best feature selection algorithm for best performance. In phase I of IDS, Support Vector Machine (SVM) and k Nearest Neighbor (kNN) are used and DNN deep learing model architecture.

Keywords: DNN,network,Knn,svm

1. Introduction

Critical infrastructures are highly complex systems that utilize cyber and physical components in their daily operations. The backbone of these facilities consists of an Industrial Control System..(ICS), which plays an important role in the monitoring and control of critical infrastructures such as smart power grids, oil and gas, aerospace, and transportation [1] [2]. Therefore, the safety and security of ICSs are paramount for national security. The inclusion of the Internet of Things (IoT) in ICSs opens up opportunities for cybercriminals to leverage the system vulnerabilities towards launching cyber-attacks [3] [4]. Awareness of the cyber-security vulnerability in ICSs has been growing since Stuxnet, the first cyber-attack that specifically targeted these technologies, revealed in 2010. Stuxnet intended to sabotage the system’s operation without disturbing Information Technology (IT) systems [5]. In 2015, another cyberattack by the name of Black-Energy was used to target Ukraine's power grids, causing a massive power outage that affected about 230,000 people [6]. In February 2020, three U.S. gas pipeline firms announced another cyber-attack alleging a shutdown of electronic communication systems for multiple days [7]. While some of these attacks may result in information leakage, others can damage the physical system or misrepresent the system state to the monitoring engineer. These examples emphasize the growing cyber threat on Operational Technology (OT), which runs much of the enabling computer technologies that ICS in critical infrastructure (i.e., power, gas, and water), now rely on [2] [8]. modern communications and the economic, agricultural, cultural, industrial, and political fields. However, legacy technology is not fit for purpose in the current technological landscape. Innovative technologies will be needed to meet the requirements of the new era in the coming years. The Internet of Things (IoT) is an example of new and innovative technology, whereby all objects may be integrated into smart systems utilizing wireless technologies, offering incredible efficiencies and convenience for users but entailing great risks in the event of malicious attacks. Due to the rapid spread of IoT technology, the expected number of devices in use will exceed 25 billion by the end of 2020 [1]. Internet users (i.e., the majority of people) increasingly produce, store, and use vast amounts of data and information, which is commensurately a highly valuable resource. Modern data storage and analysis techniques have enabled wholly new dimensions of intersection. between ICT fields and a vast array of areas of life not associated with the legacy internet landscape of the 20th century. The major challenge facing computer science at this juncture is to protect and secure data from risks, as well as to protect it (and its genuine users) from malicious attacks. In-depth defense, such as multifactor authentication and the building of secure systems, does achieve protection, but it cannot detect whether a company is experiencing a zero-day attack or not. Furthermore, as well as avoiding attacks, systems must protect and gather data as evidence, as well as generally employing existing solutions such as firewalls, encryption, and intrusion detection systems. However, there are two essential problems with the existing solutions: They cannot detect and deal with new attacks (such as zero-day attacks); They generate a high rate of false positive alerts, thus increasing the time and economic costs of checks needed to verify the validity of such alerts. Consequently, researchers have proposed using cyber threat intelligence to detect stealth-based attacks. Today, threat intelligence is considered a vital asset for many organizations because it helps protect information and guarantees privacy for users. For these reasons, many restrictive approaches have been suggested. However, most of them rely on detecting adversary behaviors and then taking retroactive action post facto. One of the best solutions to detect adversary behavior is using IDS, which can potentially identify attacks before they violate user data privacy and wreak damage on user systems.

2. Literature Survey:

Traditionally, ICSs were in an isolated environment with the focus on safety, where each system is safeguarded to stop the process if something goes wrong. However, the introduction of Internet protocols, IoT devices, and wireless technologies within ICSs has resulted in significantly less isolation from the outside world. Consequently, safety mechanisms, which were not designed to deal with malicious attacks, face more vulnerabilities than ever before. The majority of current existing techniques on cyberattack detection in ICSs are based on traditional IDSs, which are mainly designed for IT security analysis [5] [17]. IDSs can be categorized as signature-based and learning-based techniques. Signature-based approaches use databases and fixed signatures to detect known attacks, rendering them inefficient in detecting unknown or new attacks [19]. On the other hand, learning-based systems aim to identify process trends or behaviors that increase the efficiency to manage unexpected intrusions [20]. [21] used a common-path mining method for anomaly detection in smart cyber-physical grids. An attack detection technique based on the Pearson correlation between two sensor parameters was used in [22]. Authors in [23] utilized an IDS based on the Gaussian process to the attack strategy for anomaly detection. While these approaches are effective in detecting unusual activates, they are not reliable due to frequent upgrades in the network, resulting in different IDS topologies

In contrast, learning-based IDSs are designed based on a moving target to continually evolve and learn new vulnerabilities [24] [25]. These methods try to generate the normal behavior of the system using existing datasets, then identify the irregular pattern as abnormalities. The authors of [26] proposed an anomaly detection technique based on reinforcement learning and convolutional autoencoders for ICS. Alternatively, [27] addresses the detection of Denial of Service (DoS) attacks using Support Vector Machine (SVM) and RF. [28] suggested an unsupervised technique for the effective detection of privacy attacks based on observations of eavesdropping attacks. [29] uses a variety of DNN methods, including different variants of convolutional and recurrent networks for cyber-attack detection in water treatment facilities. An ICS anomaly detection method using Long Short-term Memory (LSTM) networks is proposed in [30]. The authors of [31] proposed an attack detection techniques based on Hierarchical Neural Network. Similarly, [32] proposed a deep learning-based IDS through utilizing Recurrent Neural Networks (RNNs). In another study [33], the authors applied a stacked Nonsymmetric Deep Autoencoder (NDAE) to develop their IDS. [34] proposed an unauthorized intrusion detection technique and conducted backdoor attacks on a SCADA Industrial Internet of Things (IIoT) testbed. [35] proposed a graphical model-based approach for detecting abnormal behavior in an ICS using Bayesian networks to map the relationship between sensors and actuators. [36] implemented a toolchain with multiple state-of-the-art Anomaly Detection (AD) techniques used for detecting attacks that appear as anomalies. Their findings suggest that detection rates can change dramatically when considering different detection modes, thereby necessitating a reliable and real-time AD technique to maintain resilience in critical infrastructures. [37] proposes a genetic algorithm (GA) to find the best NN architecture for a given dataset, using the NAB metric to determine the consistency and quality of different architectures. [38] evaluates the application of unsupervised machine learning algorithms, including DNN and SVM, to detect anomalies in the Cyber-Physical System (CPS) using data from a Secure Water Treatment (SWaT)...testbed. Results indicate that the DNN classifier results in less false positives when compared to the one-class SVM, while SVM can detect more anomalies. Although the above-mentioned works addressed some of the issues related to cyber-attack detection in ICSs, most of them are heavily reliant on feature engineering. These methods are quite complicated and require sophisticated learning techniques, which can potentially increase their computational burden. Furthermore, the majority…of current proposed…techniques are. evaluated using balanced datasets, which lack the standard representation of imbalanced data in the ICS environment. Thus, it is hard to deploy such algorithms as they cannot extract various discriminative information from real-world imbalanced datasets. As such, in this paper, we propose a deep learningbased attack detection technique, which extracts a new representation from raw imbalanced datasets, for reliable and accurate attack detection with a low false-positive rate in highly imbalanced datasets from ICS environments.

3. Implementation:

overcome the above limitations, ensemble machine learning models are being used by researchers which can work in multiple domains. This is another subset of machine learning which can overcome the limitation of the base level machine learning models as well as deep learning models. This model can work deeply to identify the anomalies. In this paper, we propose a voting ensemble machine learning model to detect intrusion in modern networks. The model we proposed is a combination of traditional machine learning modes that are capable of identifying anomalies in a huge range of network traffic. We evaluate our model using scikit learn by analysing CSE-CICIDS2018dataset [4]. This dataset is widely used as benchmarks in the similar works which allows us to make a direct comparison. This paper offers the following contributions:

1. We introduce the feature selection methods by calculating the p-value.

2. We apply the voting methods by combining different types of base level machine learning methodologies.

3. We improve the classification detection rate compared with SVM, RF, KNN and DNN.

Data Extraction:

The dataset which we have used in our research work was extracted from Canadian network traffic dataset named NSLKDD. This dataset is an improved version by solving various lacking of KDD CUP 99 dataset. CSE-CICIDS2018dataset includes four files: KDDTest+, KDDTrain+, KDDTest-21 and KDDTest-20. There are 125,973 network traffic samples in the KDDTrain+ dataset, 22,554 traffic samples in the KDDTest+ dataset and in the KDDTest-21 dataset has 11,850 network traffic samples with 43 features. When the dataset is not balanced, it is difficult to classify using class label; therefore, they can be categorized into 5 network attacking groups: Normal, Probe, R2L, U2R and DoS. There are some attacks which is not present in the training set but exist in test set, which make it more realistic. We discuss DoS, Remote to Local (R2L), Probe, and User to Root (U2R) in the following. When a network machine affected by DoS, it cannot handle the legitimate request and access. Using Probe attacker collects potential information of the target system to attack more in future. Using R2L, the attacker tries to get unauthorized access of the target system. Using U2R, the attacker gets access to another system which is unauthorized for the attackers and they try to gain the root access using this technique.

Data Cleaning:

By using python language programming, the first step that been used in this research after extracting the dataset is data cleaning . By extracting the Canadian CSE-CICIDS2018dataset, we get 43 features. Among them, we remove two features which mainly contain some continuous value instead of categorical data. And, we also remove another feature which contains constant value. Finally, after cleaning the dataset, we obtained 40 useful features. Among 40 features, one feature is used as a target feature.

Data Labeling:

Among the 40 features, there are 3 attributes which are non-numeric. They are “class”, “flag” and “protocol_type”. For instance, the feature protocol type has tcp, udp, and icmp types of attributes and after labeling, it is turned into 0, 1, and 2 respectively. Similarly, the feature “class” has 5 types of attributes and “flag” have 11 types of attributes. In the same way, all non-numerical values are transformed into numerical values after labeling and finally, our prediction target is mapped into 5 categories of classification.

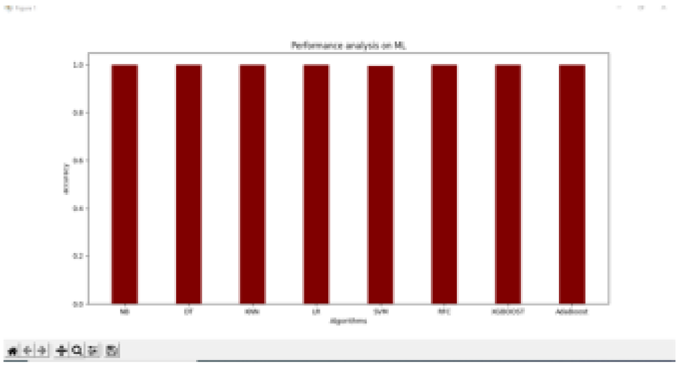
Feature Selection:

Among the feature selection techniques, the fastest approach is Filter method. There are different types of filter methods available. In our research, we consider the p-value of features using the scikit learn library . The p-value is used for selecting the important features that is significant statistically and under that statistical data which is able to identify the physical parameters or features that can classify normal and anomaly in the network traffic. If p-value is small, the parameters are said to be significant. the changes of detection rate according to the increment of number of features. For plotting the changes of DR, here the base level machine learning algorithm called logistic Regression is applied to identify the normal traffic in the network intrusion detection system. After feature selection we do scaling of the selected feature using StandardScaler function of scikit learn.

DNN Classifier:

proposed method where multiple methods can combine and provide better result. It can give more accurate result rather than the single base model. This method has been used in different sectors as a classification method to improve the performance. In this research work, we will apply this voting method to get more accurate result rather than a single model. The single model can be any machine learning algorithm, e.g.,SVM,KNN,RFC. When we use that single model as an input of voting ensemble methods, we call it the base model. DNN with select k-Best is used as a classification model. In this method, at first, we need to prepare multiple classification models using the training dataset. Here, the base algorithm can be prepared using various splits of the same algorithm and same dataset or using different algorithms with the same dataset.

4. Results:



5. Conclusion:

Critical infrastructures are complex cyber and physical systems that structure the lifeline of modern society, and their reliable and secure operations are essential to national security. In this paper, we proposed a generalized ensemble deep learning-based cyber-attack detection method specifically designed for ICS. The proposed technique includes a deep representation-learning model, which constructs new balanced representations from the raw imbalanced dataset. The new representations are then used in an ensemble deep learning algorithm based on DNN and DT classifiers to detect cyber-attacks. The performance of the proposed model is verified using two different ICS datasets obtained from real critical infrastructure facilities. Our proposed approach outperformed conventional classifiers with %10 higher f1-score in both datasets evaluated and produced higher accuracy, with %95.86 for the Gas Pipeline dataset and %99.67 for the Secure Water Treatment dataset. Results were compared with traditional classifiers, such as RF, DNN, and ADA, along with multiple peer proposed approaches in the current literature. The proposed approach outperformed other techniques in all four-evaluation metrics. Although our approach performed better than existing techniques, there is room for improvement when dealing with few samples, as illustrated in the GP dataset. Additionally, identifying the attack type and its location is also very important to prevent processing downtime and computation efficiency once an attack is detected. Therefore, our future work will focus on optimizing the accuracy of the proposed method and developing an additional model to identify different attack types and their locations. This will avoid critical system failure and improve the network security of ICSs against similar cyber-attacks.

REFERENCES

1. H. Karimipour and V. Dinavahi, "Extended Kalman filter-based parallel dynamic state estimation," 2016.

2. H. Karimipour and V. Dinavahi, "Parallel domain decomposition based distributed state estimation for large-scale power systems," IEEE Transactions on Industry Applications, vol. 2015, 2015.

3. J. Sakhnini et al., "Security aspects of Internet of Things aided smart grids: A bibliometric survey," Internet of Things, p. 100111, 16 9 2019.

4. H. HaddadPajouh et al., "A survey on internet of things security: Requirements, challenges, and solutions," Internet of Things, p. 100129, 9 11 2019.

5. F. Zhang et al., "Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data," IEEE Transactions on Industrial Informatics, vol. 15, no. 7, pp. 4362- 4369, 1 7 2019.

6. CISA, "Cyber-attack against Ukrainian critical infrastructure," 2016. [Online]. Available: https://www.us-cert.gov/ics/alerts/IR-ALERT-H16-056-01.

7. C. Buurma and A. Sebenius, "Ransomware shuts gas compressor for days in the latest attack," 2020. [Online]. Available: https://www.bloomberg.com/news/articles/2020-02-18/ransomwareshuts-u-s-gas-compressor-for-2-days-in-latest-attack.

8. H. Karimipour and V. Dinavahi, "Robust massively parallel dynamic state estimation of power systems against cyber-attack," IEEE Access, vol. 6, pp. 2984-2995, 21 12 2017.

9. E. M. Dovom et al., "Fuzzy pattern tree for edge malware detection and categorization in IoT," Journal of Systems Architecture, vol. 97, pp. 1-7, 1 8 2019.

10. I. N. Fovino et al., "Modbus/DNP3 state-based intrusion detection system," in Proceedings - International Conference on Advanced Information Networking and Applications, AINA, 2010.

11. B. Kang et al., "Towards a stateful analysis framework for smart grid network intrusion detection," in 4th International Symposium for ICS & SCADA Cyber Security Research 2016 (ICS-CSR), 2016.

12. S. Cheung et al., "Using model-based intrusion detection for SCADA networks," 2006.

13. I. Friedberg et al., "Combating advanced persistent threats: From network event correlation to incident detection," Computers and Security, vol. 48, pp. 35-57, 1 2 2015.

14. H. Karimipour and H. Leung, "Relaxation-based anomaly detection in cyber-physical systems using ensemble kalman filter," IET CyberPhysical Systems: Theory & Applications, 5 9 2019.

15. J. Yang et al., "Anomaly detection based on Zone Partition for security protection of industrial cyber-physical systems," IEEE Transactions on Industrial Electronics, vol. 65, no. 5, pp. 4257-4267, 1 5 2018.

16. A. Jahromi et al., "A deep unsupervised representation learning approach for effective cyber-physical attack detection and identification on highly imbalanced data," 2019.