Virus Detect URL Web Application Firewall

P.Gayathri[1], K. Vijayalakshmi[2], Mr.G.Gurudeep[3]
[1, 2] Dept of Computer Science Engineering
[3]Professor, Dept of Computer Science Engineering
[1, 2, 3] P.S.R.Rengasamy College of Engineering For Women, Sivakasi, Tamil Nadu.

Abstract- *Online apps can be maliciously abused by malicious HTTP requests. Usually, aWeb Application Firewall defend web application against established threats using a pattern matching system..However, the implementation of WAF is typically costly according to the case. In addition, the device cannot block on unknown malicious order. In this abstract, come up with an effective, because it involves the specification of pattern matching learning solution to solve this problem. Our method uses Character -Level –Convolution--Neural Network (CLCNN)with a very wide global max pool to extract the HTTP request function and to classify it in a regular or malicious request we tested our method on the NSL-KDD dataset and achieved 98.8%accuracy under 10-fold cross-validation and on average processing time per request was 2.35ms. New distinguishing and plaintext recovery attacks against all versions of TLS and DTLS and in almost all implementation of the two protocols. Over attacks are based on timing-based side channel and exploit TLS and DTLS design and implementation decision. We describe how to connect a full plaintext recovery attacks against implementation that follow the standard, and the partial plaintext recovery attack against implementation that do not. We discuss a number of counter measure for the attacks, and describe their practicality and effectiveness. We conclude the thesis by discussing the wider implication of our work on the design and implementation of our work on the design and implementation of secure network protocols.*

Keywords- Web Application Firewall, Character – Level-Convolution-Neural-Network, Network Security Laboratory – Knowledge Discovery Database, Hypertext Protocol.

I. INTRODUCTION

The evolution of secure network protocols has been largely driven by the discovery and the successful exploitation of weaknesses in either the design or the implementation of these protocols. The Domain Name System (DNS) and Transport Layer Security (TLS) provide good examples that demonstrate this broken reactive model of evolution. Maintaining the security of DNS has been a continuous challenge with high- profile and high- impact attacks frequently emerging (for example, Kaminsky's cache poisoning attack against DNS), which are usually followed by the development of ad hoc security protocols that try to protect DNS from these attack . In most cases, attack against DNS have exploited trivial , and on occasion known, weakness. Most of these attack would have been prevented if the basics DNS security mechanisms had been deployed. TLS, on the other hand, is by far the most widely deployed secure network protocols today, and which best show-cases the failure of this ad hoc approach of designing and implementing secure network protocols. Attacks of different severity and practicality levels have been published against TLS(and its predecessor, Secure Socket Layer), triggering ad hoc and non-coordinated responses from the Internet Engineering Task Force (IETF) who maintain the protocol specification, and the TLS open and closed source code development community. Alarmingly, the number of attacks against TLS has recently being on the rise including , for example, BEAST ,CRIME, Lucky 13 and BREACH .Our work takes advantage of previously unknown weakness introduced by this ad hoc approach to develop attacks that exploit the above mentioned protocols using basic , but novel , techniques.

II. LITERATURE REVIEW

2.1 Data-driven cyber security incident prediction:A survey

Sun, N.; Zhang, J.; Rimba, P.; Gao, S.; Zhang, L.Y.; Xiang. Y

Driven by the increasing scale and high-profile cyber security incidents related public data, recent years we have witnessed a paradigm shift in understanding and defending against the evolving cyber threats, from primarily reactive detection toward proactive prediction. Meanwhile, governments, businesses, and individual Internet users show the growing public appetite to improve cyber resilience that refers to their ability to prepare for, combat and recover from cyber threats and incidents. Undoubtedly, predicting cyber security incidents is deemed to have excellent potential for proactively advancing cyber resilience.

Research communities and industries have begun proposing cyber security incident prediction schemes by utilizing different types of data sources, including organization's reports and datasets, network data, synthetic data, data crawled from webpages, and data retrieved from

social media. With a focus on the dataset, this survey paper investigates the emerging research by reviewing recent representative works appeared in the dominant period. We also extract and summarize the data-driven research methodology commonly adopted in this fast-growing area. In consonance with the phases of the methodology, each work that predicts cybersecurity incident is comprehensively studied. Challenges and future directions in this field are also discussed.

2.2 A Survey on the Development of Self-Organizing Maps for Unsupervised Intrusion Detection

Qu, X.; Yang, L.; Guo, K.; Ma, L.; Sun, M.; Ke, M.; Li, M

This paper describes a focused literature survey of self-organizing maps (SOM) in support of intrusion detection. Specifically, the SOM architecture can be divided into two categories, i.e., static-layered architectures and dynamic-layered architectures. The former one, Hierarchical Self-Organizing Maps (HSOM), can effectively reduce the computational overheads and efficiently represent the hierarchy of data. The latter one, Growing Hierarchical Self-Organizing Maps (GHSOM), is quite effective for online intrusion detection with low computing latency, dynamic self-adaptability, and self-learning. The ultimate goal of SOM architecture is to accurately represent the topological relationship of data to identify any anomalous attack. The overall goal of this survey is to comprehensively compare the primitive components and properties of SOM-based intrusion detection. By comparing with the two SOM-based intrusion detection systems, we can clearly understand the existing challenges of SOM-based intrusion detection systems and indicate the future research directions.

2.3 Cyber intrusion detection bycombined feature selection algorithm

Mohammadi, S.; Mirvaziri, H.; Ghazizadeh-Ahsaee, M.; Karimipour, H

Due to the widespread diffusion of network connectivity, the demand for network security and protection against cyber-attacks is ever increasing. Intrusion detection systems (IDS) perform an essential role in today's network security. This paper proposes an IDS based on feature selection and clustering algorithm using filter and wrapper methods. Filter and wrapper methods are named feature grouping based on linear correlation coefficient (FGLCC) algorithm and cuttlefish algorithm (CFA), respectively. Decision tree is used as the classifier in the proposed method. For performance verification, the proposed method was applied on KDD Cup 99 large data sets. The results verified a high accuracy (95.03%) and detection rate (95.23%) with a low false positive rate (1.65%) compared to the existing methods in the literature.

2.4 Key-recovery attacks on KIDS, a keyed anomaly detection system

Tapiador, J.E.; Orfila, A.; Ribagorda, A.; Ramos, B

Most anomaly detection systems rely on machine learning algorithms to derive a model of normality that is later used to detect suspicious events. Some works conducted over the last years have pointed out that such algorithms are generally susceptible to deception, notably in the form of attacks carefully constructed to evade detection. Various learning schemes have been proposed to overcome this weakness. One such system is Keyed IDS (KIDS), introduced at DIMVA “10. KIDS” core idea is akin to the functioning of some cryptographic primitives, namely to introduce a secret element (the key) into the scheme so that some operations are infeasible without knowing it. In KIDS the learned model and the computation of the anomaly score are both key-dependent, a fact which presumably prevents an attacker from creating evasion attacks. In this work we show that recovering the key is extremely simple provided that the attacker can interact with KIDS and get feedback about probing requests. We present realistic attacks for two different adversarial settings and show that recovering the key requires only a small number of queries, which indicates that KIDS does not meet the claimed security properties. We finally revisit KIDS' central idea and provide heuristic arguments about its suitability and limitations.

2.5. A survey of data mining and machine learning methods for cyber security intrusion detection

Buczak, A.L.; Guven, E

This paper presents the results of a literature survey of machine learning (ML) and data mining (DM) methods for cyber security applications. The ML/DM methods are described, as well as several applications of each method to cyber intrusion detection problems. The complexity of different ML/DM algorithms is discussed, and the paper provides a set of comparison criteria for ML/DM methods and a set of recommendations on the best methods to use depending on the characteristics of the cyber problem to solve. Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. Cyber security systems are composed of network security

systems and computer (host) security systems. Each of these has, at a minimum, a firewall, antivirus software, and an intrusion detection system (IDS). IDSs help discover, determine, and identify unauthorized use, duplication, alteration, and destruction of information systems. The security breaches include external intrusions (attacks from outside the organization) and internal intrusions (attacks from within the organization).

There are three main types of cyber analytics in support of IDSs: misuse-based (sometimes also called signature-based), anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel (zero-day) attacks. Anomaly-based techniques model the normal network and system behavior, and identify anomalies as deviations from normal behavior. They are appealing because of their ability to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, thereby making it difficult for attackers to know which activities they can carry out undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors.

The main disadvantage of anomaly-based techniques is the potential for high false alarm rates (FARs) because previously unseen (yet legitimate) system behaviors may be categorized as anomalies.Hybrid techniques combine misuse and anomaly detection.

They are employed to raise detection rates of known intrusions and decrease the false positive (FP) rate for unknown attacks. An in-depth review of the literature did not discover many pure anomaly detection methods; most of the methods were really hybrid. Therefore, in the descriptions of ML and DM methods, the anomaly detection and hybrid methods are described together.

III. EXISTING SYSTEM

In Exiting system, virus propagation has been based on network structure and virus features. Of the many models proposed to simplify analysis of the virus propagation process, the most representative are the SIS model proposed by Kephart and White and the SIR model extended from the SIS model. To address the problems noted , several specific challenges must be tackled: The search engine is a complicated platform , spreading information across time and spatial domains. It act as a virtual virus pool if the indexed pages include malicious codes. The virtual virus pool changes over time ,gathering more and more viruses. Any access of users to the virtual virus pool can be considered black to be a virtual virus propagation path that changes the social network structure. In contrast to traditional propagation , the virtual propagation path can spread viruses among disconnected users or separated social networks.

IV. PROPOSED MODEL

The actual dedication methods thought within additional areas (example: database)that depend on an enormous amount of record information, this harnesses utilization OS- level info moves as well as adware and spyware actions to do safe dedication. Consequently , Secom imposes an inferior cost to do business upon host OS, while using the regular data logging technique might considerably decelerate the entire program. In proposed system , we quantitatively analysis the web intrusion and detection virus propagation affects and the stability of the virus propagation process in the presence of the browser in networks. First, although social networks have a community structure that impedes virus propagation, we find that the search engine generates a propagation wormhole. Second, we propose an epidemic feedback model and quantitatively analysis the propagation effects employing four metrics : infection density , the propagation wormhole effect , the epidemic threshold, and the basic reproduction number. Third, we verify our analysis four real world data sets and two simulated data sets . Moreover , we prove that the proposed model the property of partial stability. Evaluation results show that , comp black with to a case without a search engine present , virus propagation with the search engine has a higher infection density , shorter network diameter , greater propagation velocity , lower epidemic threshold and larger basic reproduction number.

V. MODULES

5.1 Modules

- Cyber Secure Node
- Fault Node Recovery
- Generation Modules
- Network Authentication

5.2 Modules Description

5.2.1 Cyber Secure Nodes

The WSN may fail due to a variety of causes, including the following: the routing path might experience a break; the WSN sensing area might experience a leak.

5.2.2 Fault Node Recovery

In the HORA algorithm, the number of nonfunctioning sensor nodes I calculated during the wireless sensor network operation, and the parameter the HORA algorithm creates the grade value, routing tables, a set of neighbor nodes, and payload values for each sensor node, using the grade diffusion algorithm

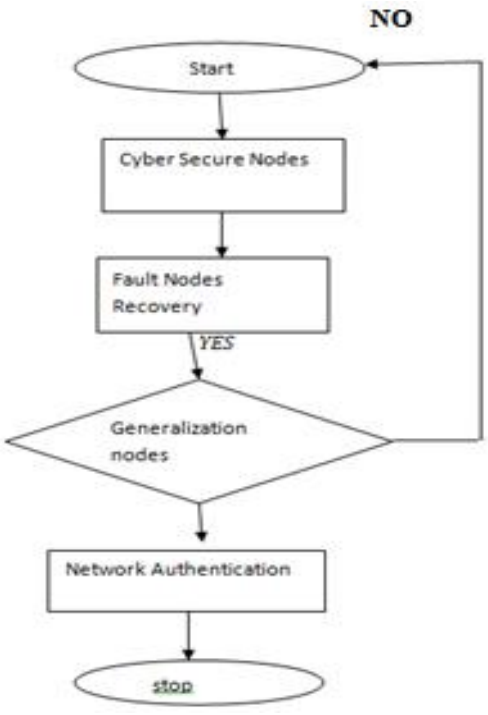
5.2.3 Generalization Module

The second protocol is aimed at generalization-based anonymous databases, and it relies on a secure set intersection protocol, such as the one found in, to support privacy-preserving updates on a generalization- based k-anonymous DB.

5.2.4 Network Authentication

If it presented in the same network it will allow you to use the application. Else it will not allow you to use the application. If you are in the authorized network means it will transfer to the next module.

Diagram



VI. ALGORITHM

6.1 Character Level Convolution Networks

In this section, We introduce the design of character-level Convolution Nets for text classification. The design is modular, Where the gradients are obtained by back-propagation to perform optimization.

6.1 **Key Modules**

The main components is the temporal convolution modules, Which simply computes a 1-D convolution. Suppose we have a discrete input function and a discrete kernel function.

Where c=k-d+1 is an offset constant. Just as in traditional convolution network in version, the modules is parameterized by a set of such kernel function and which we call weights, on a set of inputs and outputs. We call each input features, and feature size. The output is obtained by a sum over of the convolutions.

One Key modules that helped us to train deeper modules is temporal max-pooling. It is the version of the max-pooling modules used in computer vision.

The non-linearity used in our model is the rectifier or thresholding function. Which make our convolutional layers similar to rectified linear units. The algorithm used is stochastic gradient descent with a minibatch of size 128, using momentum and initial step size which is halved every 3 epoches for 10 times. Each epoch takes a fixed number random training samples uniformly sampled across classes. This number will later be detailed foe each dataset sparately. The implementation is done using torch.

6.2 **Character quantization**

Our modules accept a sequences of encoded characters as input. The encoding is done by prescribing an alphabet of size for the input language, and then quantize each character using encoding. Then, the sequences of character is transformed to a sequences of such sized vector with fixed length. Any character exceeding length is ignored, and any character that are not in the alphabet including blank character and quantized as all-zero vectors. The character quantization order is backward so that latest reading on character is always placed near the begin of the outputs, making it easy for fully connected layers to associate weights with the latest reading.

6.3 **Models Design**

We design two convolutions nets-one large and one small. They are both 9 layers deep with 6 convolutional layers and 3 fully-connected layers. The input numbers of features equals due to our character quantization method and the input feature length is 1014. It seems that 1014 character could already capture most of the texts of interest.

6.4 Data Augmentation using Thesaurus

Many researchers have found that appropriate data augmentation techniques are useful for controlling generalization error foe deep learning models. Theses techniques usually work well when we could find appropriate properties that the modules should possess. In terms of texts, it is not reasonable to augment the data using single transformation as done in images or speech recognition.

VII. CONCLUSION

An intrusion detection tree IntruDTree machine-learning-based security model. In our approach, We have first taken into account the ranking of security features according to their importance and then built a tree-based generalized intrusion detection modal based on the selected important features. We have done this to make he security modals effective in teams of prediction accuracy for unseen test cases, and efficient by reducing the computational cost with the processing of less number of features while generating the resultant tree-like model. Finally, the effectiveness of our IntruDTree model was examined by connecting a range of experiment on cyber security dataset. We have also compared the outcome result of IntruDTree with several traditional popular machine learning method to analyze the effectiveness of the resulting security model. Future work to assess the effectiveness of the IntruDTree model by collecting large dataset with more dimensions of security features in IoT security services, and measuring its effectiveness at the application levels in the domain cyber security.

REFERENCES

[1] H. Liu, C. Yang, M. Huang, and C. Yoo, ‘‘Soft sensor modeling ofindustrial process data using kernel latent variables-based relevance vectormachine,’’Appl. Soft Comput., vol. 90, May 2020, Art. no. 106149.

[2] H. Haimi, M. Mulas, F. Corona, and R. Vahala, ‘‘Data-derived soft-sensorsfor biological wastewater treatment plants: An overview,’’Environ. Model.Softw., vol. 47, pp. 88–107, Sep. 2013.

[3] X. Yuan, Y. Wang, C. Yang, Z. Ge, Z. Song, and W. Gui, ‘‘Weighted lineardynamic system for feature representation and soft sensor applicationin nonlinear dynamic industrial processes,’’IEEE Trans. Ind. Electron.,vol. 65, no. 2, pp. 1508–1517, Feb. 2018.

[4] C. Shang, F. Yang, D. Huang, and W. Lyu, ‘‘Data-driven soft sensor development based on deep learning technique,’’J. Process Control, vol. 24,no. 3, pp. 223–233, Mar. 2014.

[5] W. Yan, P. Guo, Y. Tian, and J. Gao, ‘‘A framework and modeling methodof data-driven soft sensors based on semisupervised Gaussian regression,’’Ind. Eng. Chem. Res., vol. 55, no. 27, pp. 7394–7401, Jul. 2016.

[6] H. Yu and F. Khan, ‘‘Improved latent variable models for nonlinear anddynamic process monitoring,’’Chem. Eng. Sci., vol. 168, pp. 325–338,Aug. 2017.

[7] S. Wold, N. Kettaneh-Wold, and B. Skagerberg, ‘‘Nonlinear PLS modeling,’’Chemometric Intell. Lab. Syst., vol. 7, nos. 1–2, pp. 53–65,Dec. 1989.

[8] Y. H. Bang, C. K. Yoo, and I.-B. Lee, ‘‘Nonlinear PLS modeling withfuzzy inference system,’’Chemometric Intell. Lab. Syst., vol. 64, no. 2,pp. 137–155, Nov. 2002.

[9] H. W. Lee, M. W. Lee, and J. M. Park, ‘‘Robust adaptive partial leastsquares modeling of a full-scale industrial wastewater treatment process,’’Ind. Eng. Chem. Res., vol. 46, no. 3, pp. 955–964, Jan. 2007.

[10] S. H. Woo, C. O. Jeon, Y.-S. Yun, H. Choi, C.-S. Lee, and D. S. Lee, ‘‘On-line estimation of key process variables based on kernel partial least squaresin an industrial cokes wastewater treatment plant,’’J. Hazardous Mater.,vol. 161, no. 1, pp. 538–544, Jan. 2009.

[11] Y. Lv, J. Liu, and T. Yang, ‘‘Nonlinear PLS integrated with error-basedLSSVM and its application to NOx modeling,’’Ind. Eng. Chem. Res.,vol. 51, no. 49, pp. 16092–16100, Dec. 2012.

[12] H. Liu, C. Yang, B. Carlsson, S. J. Qin, and C. Yoo, ‘‘Dynamic nonlinearpartial least squares modeling using Gaussian process regression,’’Ind.Eng. Chem. Res., vol. 58, no. 36, pp. 16676–16686, Sep. 2019.

[13] W. Ku, R. H. Storer, and C. Georgakis, ‘‘Disturbance detection and isolation by dynamic principal component analysis,’’Chemometric Intell. Lab.Syst., vol. 30, no. 1, pp. 179–196, Nov. 1995.

[14] Y. Dong and S. J. Qin, ‘‘Dynamic latent variable analytics for process operations and control,’’Comput. Chem. Eng., vol. 114, pp. 69–80, Jun. 2018.