A Systematic Review – Intrusion Detection Algorithms Optimisation for Network Forensic Analysis and Investigation

Kanti Singh Sangher
School of IT
Centre for Development of Advanced Computing
Noida, India
kantisingh@cdac.in

Dr. Archana Singh
Amity School of Enginnering & Technology
Amity University
Noida, India
asingh27@amity.edu

***Abstract*—As the digital world is growing widespread crime in the cyberspace is also increasing. Knowledge sharing and utilization of services attracts to use the digital devices, but the concern here is malicious usage of the system. If the crime takes place over the network how to collect it, analyze it and investigate based on the evidences. So the role of forensic and incident response is crucial here. Digital forensics is categorized in Disk, Live, Network and Mobile Forensics. Anomaly or attack over the network comes under network forensic branch. In this paper extensive literature review is performed to compare the latest intrusion detection systems and based on the learning's a system is proposed which covers the peer to peer architecture system and utilization of web robots to trace the attack and log it in a form which will be a useful input for forensic investigation and analysis work.**

Keywords—Intrusion detection, Machine Learning, protocol analysis, network security, Anomaly Detection, social networks.

I. INTRODUCTION

In network forensics analysis of the activities performed during the particular time and based on that reconstruction of the events is needed by the investigators. To find out the suspected machine and user behind the crime, reconstructed crime events within a network are used in the analysis process. There are many techniques such as intrusion detection systems, Packet Capturing and NetFlow which can be used to extract the network data. Network forensic tools take captured network data as the input to perform the analysis and based on the reconstructed maps creates digital evidence to support the law enforcement agencies.

Intrusion detection systems (IDS) offers close check on all the network activities to provide the secure environment. If there is any suspicious traffic in the network, it alerts the administrator to take the necessary actions. So, the inputs given by the IDS plays a major role in the network forensic analysis, if there is any crime within the network happens.

Packet Capturing allows to record bit by bit information which travels inside the network. It ensures that data having particular connection characteristics such as from this to this specific system allowed to travel in the network. This process limits the data flow. As the packet capture data in the network creates large volume within the short duration, it is not feasible to retain it for longer period of time.

A NetFlow technique records network data on each connection passing through the devices used for monitoring. NetFlow data consists of source connection with destination alongwith volume of data passed. Here one important point is that while in case of packet flows it is not feasible to record and preserve the data for the long duration, NetFlow data as it stores summary of the data may be used to preserve for extended time. So commonly these three techniques are used in the network data collection and analysis to initiate the network forensic investigations and produce evidences to prove the crime.

The paper covers wired and wireless networks and presents the implementation applied to detect the various network attacks in each. Based on the results gaps within the work identified and as a solution a framework proposed. The remaining paper is organised as follows: Section 2 discusses the motivation of research and the objective of this review, Section 3 elucidates the sources for review literature and filtration methods for the specific papers chosen, Section 4 highlights the learning from the review, Section 5 enumerates results of the research and identified research gaps, and Section 6 concludes the review and highlights the areas of future exploration.

Steps performed in the Network Forensic analysis

Intrusion detection systems keep an eye on all the network related activities to record the policy violations if any or intrusion with malicious intentions. There are varieties of environment to utilize the combination of network forensics and intrusion detection mechanisms for example user's home system [1], here manually check of the intrusion is also possible. Most of the intrusion detection as well as prevention systems aim to record log details, possible incidents of intrusion and report of the intrusion instances if there is any.

As shown in the image network forensic analysis starts with inputs from the IDS through the logging information and based on the analysis outcome helps the system to recover from the similar type of intrusion.

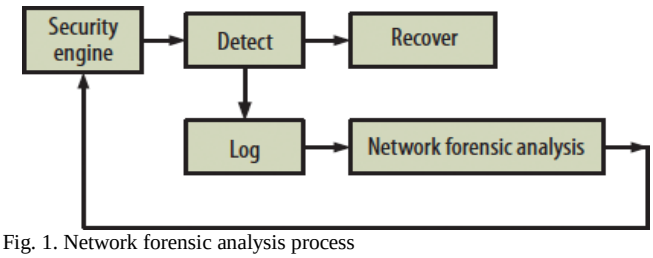

Fig. 1. Network forensic analysis process

Static system is not capable enough in case of today's digital world, so a realistic system to handle the automated blend of intrusion analysis with network forensic analysis is desired, to generate dynamic feedback helpful to update the access rules once the occurrence of real time attack detected.

II. Emerging Network Forensics Areas

A brief glance of existing network forensic areas emerging in today's era of technology will help to understand the process more. So, network forensics methods are amalgamation of different network such as social network, data mining techniques and digital forensics tools and technologies.

Social networks

Social Networking websites i.e. Google+, Facebook, YouTube, Twitter etc. have grown tremendously in recent time, due the interest they were able to generate with user community to socialize, to perform different household activities and also to explore the knowledge and technology. Due to the competition of getting more and more users these sites focus on applications and features with strong appearance to engage users, but at the same time lack on the integration of security of the system [2]. So, priority of the security is not at the place it deserves, and this attracts the attackers to steal the data as risks of the system not covered fully by the designers.

Data mining

Data Mining is a vast area but here it plays a very important role, analysis of the intrusion and feedback to work the system intelligently is the base of the usage. Data mining techniques helps approaches to find out the relevant patterns within the system data. For example, the different forensic profiles creation is possible with finding out the way user interacts within the application or system where large amount of data needs to be processed. There are several types of digital media in form of physical as well as logical data sources available for forensic analysis, huge amount of data processing and analysis is possible through the data mining techniques [3] to extract and analyse the information.

Digital forensics

Digital crime investigation evolved with numerous efforts given by the researchers to design and develop the state-of-the-art tools which assist to seize the various types of data and acquire them as digital evidences. But innovation in cybercrime techniques and development of various high-end technologies are making the digital investigation [4] tasks more challenging day by day. Amount of the data and complexity to analyse the data limits the digital investigation and analysis process difficult for the organisations. Data visualization techniques works in digital forensics whether in terms of network/disk/live or any form of data, as a doting tool and also needs to be explored to display significant data and represent them into dimensionality, size and at the complexity levels.

III. Factors to be Considered

In network forensics process as it is a natural extension of the computer forensics, several critical factors need to be considered based on the computer and digital data storage devices, file system, registry information and memory data. Whether it is IDS, NetFlow or packet capturing mechanism to be used for the implementation, it should be able to perform accurate monitoring to deal with capturing, logging and performing the analysis of the data within the network traffic with consideration of packets format. Mechanisms to detect the threshold of data capturing should be implemented with properly designed alerts. To defend the system from future attacks the patterns should be rigorously studied with factors such as how and where the attack took place, who was the wrongdoer, line of attack in peer-to-peer network and duration of the exploit, vulnerabilities of the system etc.

IV. Related Works

4.1. A Deep Learning Approach using Recurrent Neural Networks

Research work [3] and [4] are based on artificial intelligence and machine learning approaches in which utilization of algorithms such as SVM, random forest, recurrent neural network class implemented, and output gives visualization of data. For example, RNN implementation provides directed graph between the connected nodes along a series of events. Recurrent Neural Network [5] consists of input units, output units and most important part hidden units which actually performs major work in the research. To study the output and its relevance of RNN-IDS model have been designed. Binary classification (normal, anamoly) of the network traffic and 5-category (Normal, DoS, R2L, U2R and Prob) classification of attacks these two experiments test and show the output of RNN-IDS. Other machine learning methods performance compared with RNN based model with different experiments.

Research paper shows the performance comparison of Binary classification experiments with other popular deep learning techniques and machine learning algorithms such as ANN, random forests, naïve Bayesian, support vector machine, multi-layer perceptron etc. as mentioned in [6]. Similarly, RNN-IDS model [7] which is based on the NSL-KDD dataset analyses the multi-classification experiment also. Base of the RNN-IDS model is divided into two part i.e. Forward

Propagation and Back Propagation. In the model Forward Propagation gives the calculated output values and Back Propagation passes the residual values which are inputs for the weights updated in the connected nodes, which is basically training of the normal neural network to perform the analysis. This experiment is also compared with several machine learning algorithms such as J48, random forest and naïve Bayesian. The output of comparison shows that the performance gives higher detection rates and accuracy rate when condition is low false positive rates, majorly in case of multiclass classification which is based on the NSL-KDD dataset.

TABLE I. FEATURES OF NSL-KDD DATASET

No.	Features	Types	No.	Features	Types
1	duration	Continuous	22	is_guest_login	Symbolic
2	protocol_type	Symbolic	23	count	Continuous
3	service	Symbolic	24	srv_count	Continuous
4	flag	Symbolic	25	serror_rate	Continuous
5	src_bytes	Continuous	26	srv_serror_rate	Continuous
6	dst_bytes	Continuous	27	rerror_rate	Continuous
7	land	Symbolic	28	srv_rerror_rate	Continuous
8	wrong_fragment	Continuous	29	same_srv_rate	Continuous
9	urgent	Continuous	30	diff_srv_rate	Continuous
10	hot	Continuous	31	srv_diff_host_rate	Continuous
11	num_failed_logins	Continuous	32	dst_host_count	Continuous
12	logged_in	Symbolic	33	dst_host_srv_count	Continuous
13	num_compromised	Continuous	34	dst_host_same_srv_rate	Continuous
14	root_shell	Continuous	35	dst_host_diff_srv_rate	Continuous
15	su_attempted	Continuous	36	dst_host_same_src_port_ra	Continuous
16	num_root	Continuous	37	dst_host_srv_diff_host_rat	Continuous
17	num_file_creations	Continuous	38	dst_host_serror_rate	Continuous
18	num_shells	Continuous	39	dst_host_srv_serror_rate	Continuous
19	num_access_files	Continuous	40	dst_host_rerror_rate	Continuous
20	num_outbound_cmds	Continuous	41	dst_host_srv_rerror_rate	Continuous
21	is_host_login	Symbolic			

Features of the NSL-KDD dataset as shown in table 1 used in[1] represents that how different network attributes can be analyzed to get the intrusion detection.Compared with traditional classification methods, such as J48, naïve bayesian and random forest, the performance obtains a higher accuracy rate and detection rate with a low false positive rate, especially under the task of multiclass classification on the NSL-KDD dataset.

4.2. Wireless Anomaly Detection based on IEEE 802.11 Behavior Analysis

Network devices offer variety of services from wired based to wireless, in this research work [8], focus is on the pervasive wireless technology which demands high end security to cater the risks and vulnerabilities of the systems. In recent years security protocols for wireless networks however deal with the confidentiality aspects and privacy concerns, but availability and integrity (e.g. session hijacking, denial of service and MAC address spoofing attacks) needs to be addressed. This research paper explains how anomaly-based intrusion detection system can be implemented in the wireless networks such as IEEE 802.11 to perform the behaviour analysis [9] and detect change in behaviour from the normal due to the occurrence of the attacks in the wireless network. In the research presented anomaly behaviour of the IEEE 802.11 protocols uses the monitoring of n-consecutive transitions happened on the protocol state machine. To identify the n-transitions patterns from the wireless protocols sequential machine learning methods have been implemented and also checked the probabilities of being normal behaviour by the characterizing of the patterns. This research paper used supervised learning and prepares statistical metrics using fixed size sequential patterns (n-gram). Several experiments have been implemented to test the performance of the model and cross validation of the system with more than two diverse wireless channels gives result that a low false alarm rate (<0.1%) achieved. As the approach applied tests attack library of known wireless attacks 99% detection rate has been achieved. This method of intrusion detection is able to capture the footprints of known WIFI attacks such as Deauthentication attack, Association Flood, Disassociation Flood, Authentication Flood and Fake Authentication with the quite good false positive rates.

4.3. A security approach for social networks based on honeypots

Social network expanded itself in last few years in size as well as in terms of popularity. The social network platform mainly depends on the people who are primarily providing the content and due to this user of these systems are actually targeted by the attackers. So social network can utilize honeypots, to maintain the faith in the community and achieve the sustainability in the digital world. Social honeypots technology [10] finds malicious user profiles in the social platforms by using the intrusion detection methods.
Social honeypots are primarily based on two essential components: 1) Collection of information about malicious profiles with deployment of social honeypots. 2) Classifier creation of malicious user profiles and activities to filter them from rest of the profiles and also checking of new profiles, all these constitute analysis of the unique features of the profiles.

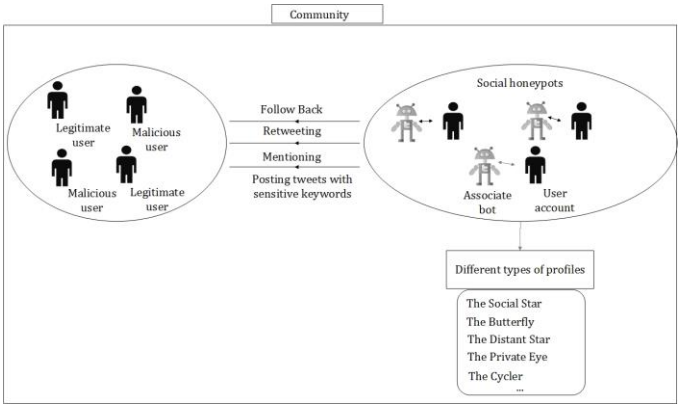

Fig. 2. Social honeypots-based intrusion detection mechanism

This research work [3] as shown in Fig 2, suggests the deployment of social honepots with agenda of targeting

malicious profiles. Social communities like twitter where number of characteristics of user can be captured and with help of the classifier creation method using machine learning algorithm within the deployed honeypots it can create malicious user profiles. In this research paper Decorate machine algorithm has been used and it gives high precision and a low rate of false positives.

4.4. Malware Automatic analysis

This paper [11] uses Sandbox and machine learning method to automate malicious codes identification.

Malicious programs are able to steal personal and business information, Denial of service attack. To provide the security of a computing environment malware analysis performed the first step in this process is malicious code classification. All the malicious code analyzed in the OS before execution of the artifact. Runtime actions monitoring of malware is also performed in the research. Sandbox features includes monitoring of created or modified files, Access or system registry key modifications, Dynamic loaded libraries, Virtual memory accessed areas, created process, instanced network connections, Data transmitted over the network.

To implement the Sandbox with machine learning attributes stored in files were engineered, Development of dictionary terms created, Vector models created. Once all the steps done ML techniques applied in the attribute vectors.

He evaluation criteria used in the paper includes Accuracy, False positives, false negatives. After implementation of machine learning methods [12] results suggest that Random Forest & J48 gives best output.

4.5. Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder

This research paper [13] discusses use of Machine Learning techniques to detect automatically cyberbullying messages in social media. Cyberbullying [14] can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication. It has negative, insidious and sweeping impacts children and outcome may even be tragic as the occurrence of self-injurious behavior or suicides.

Three kinds of information i.e. text, user demography & social network features can be used in cyberbullying detection within social media.

Text is most reliable, so used in this paper. Numerical representation for the text is done using Semantic marginalized stacked denoising auto-encoder. One of the deep learning methods used to learn robust representations, bullying feature set used; Result shows more correlation between reconstructed words.

4.6. A survey of Data Mining and Machine Learning methods for Cyber Security Intrusion Detection

Machine Learning and Data Mining for cyber security applications [15] to detect the intrusion in the system works in a dynamic way. Set of recommendations on the best methods to use depending on the characteristics of the intrusion. Comparatively analysis of the usage and best fitted machine learning methods is required as each attack deals with different parameters and domain.

To focus on clustering, decision tree and rule mining methods, Cover wired and wireless network.

Cyber Security Data Sets used in the paper are:

- Packet level data
- Netflow Data

Most all of the ML/DM algorithms covered for the cyber security intrusion detection [16] and [17]. Comparison criteria used in this research paper includes Accuracy, time for training a model, time for classifying an unknown instance. Based on the empirical comparison Bagged trees, Random Forests and ANNs give the best results.

One of the research work [21] developed security analytics for cyber-attack detection using deep learning. The security analytics architecture shown in fig. 3 performs on gigabytes of data and accurately models the highly complex data.

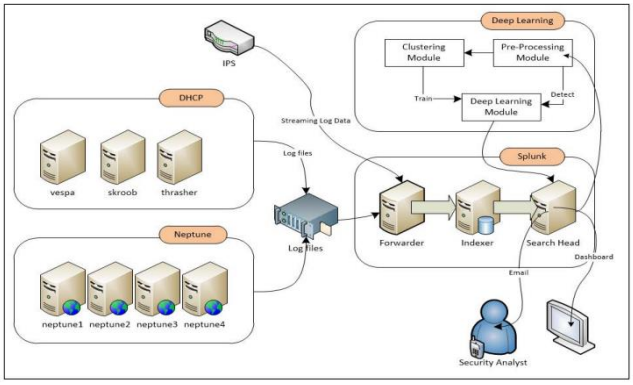

Fig.3 Deep learning-based architecture to cyber-attack detection

The results from fig 3 implementation states that, Intelligent and forecasting requirements of high-dimensional learning is achievable through deep leaners to create better models**.**

V. FINDINGS

In this systematic review the focus was to cover different areas in which network forensics can be used with intrusion detection systems. Anomaly detectors in the systems design covers, sequence of the event set for all exchanged frames using session key and frame type which will be used to generate the representation map. All the malicious events can be

extracted using session generator to classify the extracted events. But the data used in the research papers [18] are mostly offline data and training data set details are not shared fully. Most of the systems wired as well as wireless cover the prevention mechanism for attacks and to modify them based on the new findings. So, there is a need to design and develop an intrusion detection system, which should work on a real time scenario and adapt itself based on the identified anomalous behavior by the users with multi-layer and intelligent analysis.

An intelligent system with web robots or botnets proposed as a solution to cater network attacks. This will implement multilayer protection and detect anomaly by attracting the suspicious systems/users.

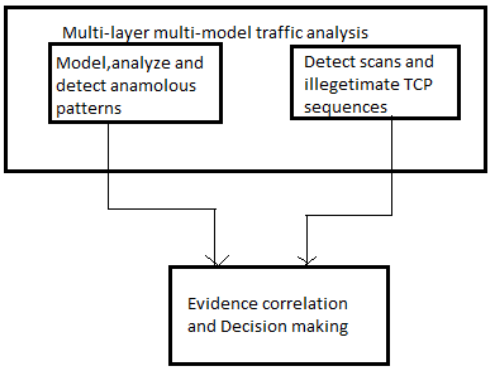

Fig. 4. Proposed framework with multi-layer intelligent system

The data collected from the traffic using artificial intelligence compatible will be very helpful for the network forensic applications. Hacking or stealing of data by different mechanisms for example ransom-ware attack, social networking attacks needs prevention of the system more than after effect mitigations.

Network threat detection at the early stage with latest tools and technology will change the scenario in India as it is considered as a green field for cyber-attacks. So, the study of different mechanisms to detect intrusion gives a vast knowledge to work and explore the network forensics, where footprints of each network log data serve as the part of digital evidence.

VI. Conclusion

Study and review of research work facilitates the need of deep learning in intrusion detection to get better results in terms of security enhancement of the existing systems as well as to produce the digital evidence from the cybercrime. Visualization of the intrusion traces will also explicitly add the value of findings as analysis majorly influences the connection between the nodes and log data.

Deep learning-based approaches in cyber-attacks and forensics reduces the limitations of information processing of huge data and also provides results with advanced analysis capabilities.

So, implementation of the proposed work given in fig [4] will give a major goal to achieve the performance based multilayered network intrusion detection system.

References

[1] Chuanlong Yin,Yuefei Zhu,Jinlong Fei, and Xinzheng He, A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, IEEE Access,Oct 12 -2017.

[2] Hamid Alipour, Youssif B. Al-Nashif, Pratik, and Salim Hariri,Wireless Anomaly Detection Based on IEEE 802.11 Behavior Analysis, IEEE Transactions on Information Forensics and Security,Vol10.2158-2170,October 2015.

[3] EI Bouzekri EI Idrissin Younes,EI Mendili Fatna, Maqrane Nisrine, A security approach for social networks based on Honeypots. IEEE, 2016.

[4] Cesar Augusto Borges de Andrade,Claudio Gomes de Mello, Julio Cesar,Malware Automatic analysis, BRICS Congress on Computational Intelligence & 11th Brazilian Congress on Computational Intelligence,2013.

[5] Rui Zhao, Kezhi Mao,Cyberbullying Detection Base on Semantic-Enhanced Marginalized Denoising Auto-Encoder, IEEE Transactions on Affective Computing,Vol 8. NO. 3,July-September 2017.

[6] Anna L. Buczak and Erhan Guven,A survey of Data Mining and Machine Learning methods for Cyber Security Intrusion Detection, IEEE Communications surveys & tutorials,vol 18,2016.

[7] H.V. Zhao et al., "Behavior Modeling and Forensics for Multimedia Social Networks: A Case Study in Multimedia Fingerprinting," IEEE Signal Processing Magazine, Jan. 2009, pp. 118-139 SOCIAL MEDIA REF.

[8] V.H. Bhat,"A Novel Data Generation Approach for Digital Forensic Application in Data Mining," *Proc. 2nd Int'l Conf.on Machine Learning and Computing* (ICMLC 10), IEEE, 2010, pp. 86-90.

[9] J. Yang, Y. Chen, W. Trappe, and J. Cheng, "Detection and localization of multiple spoofing attackers in wireless networks," *IEEE Trans.Parallel Distrib. Syst.*, vol. 24, no. 1, pp. 44–58, Jan. 2013.

[10] Aminnezhad A, Dehghantanha A, Abdullah MT. A Survey on Privacy Issues in Digital Forensics. International Journal of Cyber-Security and Digital Forensics (IJCSDF). 2012; 1:311-23.

[11] Thapliyal M, Bijalwan A, Garg N, Pilli ES. A Generic Process Model for Botnet Forensic Analysis. 2013.

[12] M. Tavallaee, E. Bagheri, W. Lu, and A. A. A. Ghorbani, ''A detailed analysis of the KDD CUP 99 data set,'' in Proc. IEEE Symp. Comput. Intell Secur. Defense Appl., Jul. 2009, pp. 1–6.

[13] B. Ingre and A. Yadav, ''Performance analysis of NSL-KDD dataset using ANN,'' in Proc. Int. Conf. Signal Process. Commun. Eng. Syst., Jan. 2015, pp. 92–96.

[14] J. Zhang, M. Zulkernine, and A. Haque, ''Random-forests-based network intrusion detection systems,'' IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 38, no. 5, pp. 649–659, Sep. 2008.

[15] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, ''A deep learning approach for network intrusion detection system,'' presented at the 9th EAI Int. Conf. Bio-inspired Inf. Commun. Technol. (BIONETICS), New York, NY, USA, May 2016, pp. 21–26.

[16] R. R. Reddy, Y. Ramadevi, and K. V. N. Sunitha, ''Effective discriminant function for intrusion detection using SVM,'' in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Sep. 2016, pp. 1148–1153.

[17] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, ''A new intrusion detection system based on KNN classification algorithm in wireless sensor network,'' J. Elect. Comput. Eng., vol. 2014, Jun. 2014, Art. no. 240217.

[18] N. Farnaaz and M. A. Jabbar, ''Random forest modeling for network intrusion detection system,'' Procedia Comput. Sci., vol. 89, pp. 213–217, Jan. 2016.

[19] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, "Learning from bullying traces in social media", *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, pp. 656-666, 2012.

[20] J. Sui, "Understanding and fighting bullying with machine learning", 2015.

[21] Security Analytics: Using Deep Learning to Detect Cyber Attacks, UNF Graduate Theses and Dissertations, Student Scholarship at UNF Digital Commons.