

A New Framework of Honeypots Network Security Using Linear Regression Decision Algorithm

Avijit Mondal*

Research Scholar MAKAUT and Assistant Professor, Department of Computer Science and Engineering, Techno International New Town, Kolkata

E-mail: avijitmondal88@yahoo.com

ORCID iD: <https://orcid.org/0000-0003-2433-4729>

*Corresponding Author

Radha Tamal Goswami

Director, Techno International New Town, Kolkata, West Bengal, India

E-mail: rtgoswami@tict.edu.in

ORCID iD: <https://orcid.org/0000-0001-5307-3666>

Soumita Sen

Assistant Professor, Department of Computer Science and Engineering, Techno International Batanagar, Kolkata

E-mail: sou.2a8.mita@gmail.com

Received: 17 June, 2023; Revised: 02 August, 2023; Accepted: 16 September, 2023; Published: 08 December, 2023

Abstract: The expansion of the Internet and shared networks aids to the growth of records generated by nodes connected to the Internet. With the development of network attack technology, all Internet hosts have become targets of attack. When dealing with new attacks (such as smart ongoing threats) in a complex network environment, existing security strategies are powerless. Compared to existing security detection techniques, honeypot systems (IoT research) can analyze network packets or log files being attacked, and automatically monitor potential attack. Researchers can use this data to accurately capture the tactics, strategies, and techniques of threat actors to create defense strategies. However, for general security researchers, the immediate topic is how to improve the honeypot mechanism that attackers do not recognize and quietly capture their actions. Honeypot technology can be used not only as a passive information system, but also to combat zero-day and future attacks. In response to the rapid development of honeypot recognition with machine-learning technology, this paper proposes a new model of machine learning based on a linear regression algorithm with application and network layer characteristics. As a result of the experiment, we found that the proposed model was 97% more accurate than other machine learning algorithms.

Index Terms: Honeypots, Linear Regression, application-layer feature, network-layer feature

1. Introduction

Based on to GDATA [1], the new threat and attacks is growing rapidly, while traditional tools are being used for a lot of information or new types of attacks. The main disadvantage of intelligent based learning is that it takes time, and it is difficult to classify attacks [2].

Understanding the motivations, goals, and techniques used by others to access the system without permission will not only prevent and protect the system from attacks, but also affect our own systems. This is the key to analyzing and predicting different attacks. Honeypots technology is a powerful information system configured to monitor, detect and analyze malicious activities [3], as ineffective intrusion-detection-systems and log files, as it does not detect much information, emissions and new cyber-attacks [4].

A honeypot is a source of safety used for investigations, attacks or destruction [5]. The cyber-attack automatically detected interactions as malicious action, the supervisor's network uses statements produced by spiteful resources to understand the individuality, motivation, and practices used by the attacker to access the system.

Honeypot used to improve the level of attack detection of the company, defined by the new fraud technology for network security defense [6]. In other words, the honeypot is used to guide the attacker in contact with it and collect information for analysis and research on the attacker's attack pattern [7 - 9].

Depending on the level of interaction of honey, it can be divided into three categories: low-interaction honeypot, medium-interaction honeypot, and high-interaction honeypot [10-12].

The honeypot detection technique should also be improved based on the widespread use of deceptive technology, but it makes the honeypot more realistic than researchers who are difficult to identify to deceive potential attackers. Therefore, many researchers began to pay attention to intelligent detection of honeypot and honeypot framework [13 - 17].

The main contribution of this paper as follows,

1. Intelligent honeypot detection technology proposed two groups of attributes that can accurately distinguish between general servers and honeypot servers. These attributes can be summarized at the application and the network layer features.
2. This paper suggests an automated detection model that uses machine learning functions and algorithms to identify cyber attack in honeypot. In order to improve finding the attackers, Linear Regression machine learning algorithm is proposed.
3. To emphasize the honeypot detection, this paper compares the Accuracy, Recall and F1 score of the 4 different machine learning algorithms and validated by 10-fold cross-validation.

The structure of the article is as follows: Section-2 is the concerned article presenting current developments in honeypot-detection research. Section-3 is the anticipated framework that describes the details of the proposed machine learning algorithm. Section 4 is the result of experiment. Section-5 is the conclusion and future-work.

2. Related Work

The anomalous characteristics aroused a lot of interest. Everyone tries to protect themselves from unauthorized use of their data and harmful access to computer systems. Over the past ten years, many security solutions have been proposed, but the results are still limited [18]. The latest work rarely uses data collected by information technology (such as honeypots), but instead relies on machine learning algorithms.

The authors of [19] have proposed a smart honeypot that can improve the security of IoT devices based on machine learning. Research IoT devices that can access the Internet to remember each device's response, provide IoT scanners to detect malicious Internet interactions, and use a model called optimization model of an IoT student to respond to attackers.

The author of [20] proposed an automated classification method based on the analysis of social spam machines (SVM), which uses social honeypots to collect information related to malicious profiles to classify community networks such as Facebook and Myspace.

The author of [21-22] proposed a honeypot-based link defense system to overcome the limitations of existing tools.

The attacker can easily determine if the server has implemented the honeypot service. Research to address these threats will make honeypot services realistic and internal structure and external connected interfaces of the honeypot framework [23-25].

The characteristics of network monitoring and perceive the virtual environment to identify honeypots. Many interactive honeypots are always distributed across firewalls and IDS [26].

The proposed method to focus on activities and services at honeypot implementation and designed an experiment to demonstrate the effectiveness of this method by adding another feature set, TCP / IP printing [27].

The honeypot feedback extracted data from Linux in user mode and VMWare and proposed a honeypot detection framework [28]. For UML, one can use the information from the fingerprint "dmesg" command or view the / etc / fstab file.

For VMWare recommended as getting the physical address and Media Access Control address. According to IEEE standard, , VMWare is assigned to the same MAC address as "00-0E-23-xx-xx-xx". Send-Safe Honeypot Hunter is the world's first anti-honeypot technology [29].

3. Approach

3.1 Honeypot

The purpose of information security policies is generally to establish mechanisms that assurance services in relations of integrity, confidentiality, authentication, and access control. The attack is constructed on cross-network scanning tools of susceptible systems [23]. Therefore, the novelty of the honeypot is that the approach presents itself as a weak resource and can automatically attract the attention of the attacker.

The usual persistence of honeypots is to convince them that they can control the actual running machine. This allows managers to see how the attackers have been compromised, protect themselves from new attacks and provide

more response time. Honey pots have very flexible and diverse forms. Most work is divided into honey pots into two methods. The first category classifies honey pots based on the interactions that the honey pot allows, and the second category classifies honey pots based on their practicality.

3.2 Framework

To identify honey pots more effectively, this document provides an automated identity framework shown in Figure 1. The framework contains three important components: feature acquisition, labeling methods and discovery models. Main acquisition function is to collect functions on the target server. The functions are divided into two categories: application layer, network layer. The values of the obtained attributes are calculated in the definition of strings or integers.

The succeeding important portion is the scoring method. As the name indicates, the aim of the Intrusion system is to collect tag information, including the IP addresses of common songs and hosts. Cyberspace search engines like Shodan, FOFA and other common platforms have been deployed to collect honey pot servers. After scanning the actual data on these platforms through the application interface, manual control methods will be used.

After retrieving function data and label data, they are sent through two methods of measuring educational data. At the study stage, each data has a data label. More specifically, the record of each training data is explained in feature data and label data. Use machine learning algorithm to train machine learning model data. The algorithm includes the next *Random Forest*, *SVM*, *kNN* and *naïve Bayes*, where the discovery phase training data is first loaded.

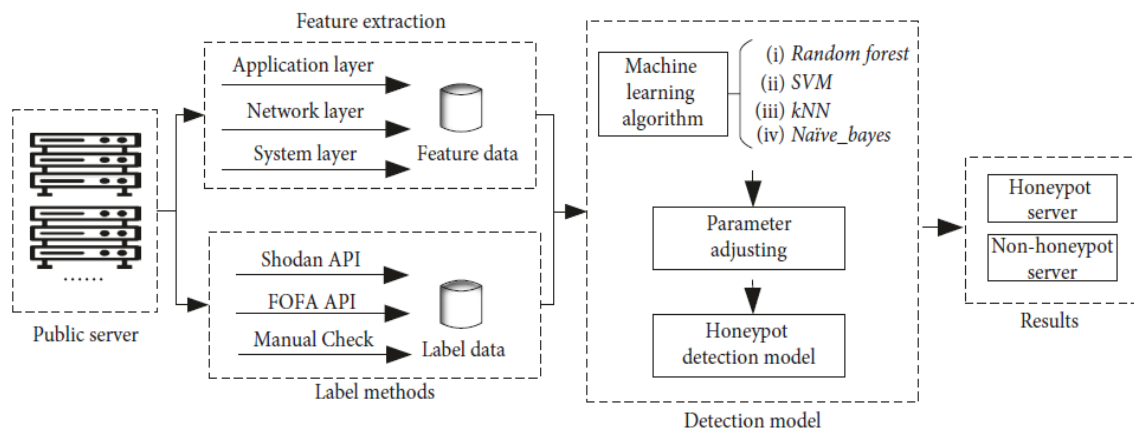


Fig. 1. The proposed work flow of Honey pot Detection

3.3 Feature Extraction

Good functionality is the foundation of a well-trained machine learning prototype. In this article, features are split into 2 parties: application-level features, network-layer features.

3.4 K-means clustering

Clustering technology is used to group data into similar classes to create user profiles. Profiles are classified as attack or non-attack profiles based on other classification algorithms. The K-means clustering algorithm, the data is grouped into identical classes to create a configuration file. Linear regression is used to model each class and provide a more meaningful and consistent representation of the data.

The qualitative data and the proposed volumetric model form a controversial contour. Crystal (crystalline phase) uses the measurement of the distance between the training model (training phase) and the suspicious patterns of the attacker by placing the suspicious path on the path.

The research stage (*Algorithm 1*) requires three constraints: the number of groups, the average start, and the constraints of the linear function. The accurate calculation of these parameters, the more accurately the model is fitted.

The algorithm returns information about *profile creation and classification* based on the attacker model (algorithm 2). In the Euclidean sense, $HCL_i[m]$ is the closest vector to a solution based on the release of a new hacker profile.

3.5 Learning Algorithm

3.5.1 Input

k// no. of clusters
 $V_{pn} = (U_1, U_2, \dots, U_n)$ // hack information

3.5.2 Algorithm

Start
 $Q_{ab} = Q1(V_{pn})$ // qualitative adaption

```

Qcd = Q2(Vpn) // quantitative adpation
m = space dimension(d) - 1;
for i = 1; i < d; i ++
    (Ci(k), Ri(k) = k means (Nid, k) // cluster creation
CLi(m) = LR (Ci(k), Ri(k), Nid) //LR
End

```

3.5.3 Output

```

Qab
CLd(m) // LR coefficients

```

3.6 Decision Algorithm

3.6.1 Input

```

k// no. of clusters
Vpn = (U1, U2,..., Un) // user information
HMab, HCLd[m]

```

3.6.2 Algorithm

```

Start
HMab = Q1(Vpn) // qualitative adpation
Ncd = Q2(Vpn) //quantitative adpation
m = space dimension(d) - 1;
for i = 1; i < d; i ++
    (Ci(k), Ri(k) = k means (Nid, k) // cluster creation
HCLi(m) = LR (Ci(k), Ri(k), Nid) //LR
End

```

3.6.3 Output

```

Distance (HCLd(m), UCLj(m)
Euqal (HMab, UMab)

```

4. Results and Discussions

4.1 Experiment

To manage the expected results more reliable plus accurate, the research will focus on real decoy servers plus conventional servers built on remote inspecting technology. In the experiment, the *scikit – learn library* was used to train ML models. The system configuration is i3 CPU, 32 GB of memory.

4.2 Dataset

Shodan and *Fofa* are famous search-engines cast-off to discover network-devices in cyberspace. The honeypot servers can be found in the following ways: First, display multiple IPs in the decoy and then randomly select multiple IPs on the internet. Second, manually check all *IP addresses* to determine if they are honeypot. Then scan all IP addresses with socket expertise and to finalize the data set which is used for experiment analysis. Finally, 2,407 IP records were collected in the experimental dataset. The number of *IP addresses* in the honeypot is 805. The actual number of systems *IP addresses* is 1602, shown in the Table 1.

Table 1. Honeypot Dataset

Data Set Tpe	Length
Honeypot Records	805
Real System Records	1602

4.3 Experiment Design

To improve model accuracy, this paper introduces three steps in detailed experimentation: size integration, parameter adjustment and cross-validation.

4.4 Size Integration

When training the model, certain key functions will affect the result. For example, Table 2 has three records. Feature2 size is larger than Feature1 size. Therefore, when modeling the model, large factors can dominate the expected

results. Therefore, the fourth record company can expect 0. But the truth is one. This is why the dimensions must be constant.

Table 2. Records for example

Feature 1	Feature 2	Feature 3	Label
0.02	3100	14	1
0.061	46000	60	0
0.001	13685	18	1
0.03	8200	24	0

4.5 Parameters Adjusting

By model training have chosen the best parameters of the model. The Linear Regression algorithm has two important parameters, which are used to find out the accuracy. The parameters are "n_estimators" and "max_depth". "N_estimators" is the number-of-decision trees and "max_depth" is the largest depth allowed for each tree. If $n_estimators$ is too large, the outcome may be redundant. However, if "n_estimators" are too small, the findings may not be sufficient. Therefore, an appropriate value for "n_estimators" is very important.

{110, 140, 210, 320, 600, 900, 1300} assumed for "n_estimators," and {6, 9, 16, 27, 32, 90} was prepared for "max_depth." These two sets were tested separately, and the best values were selected.

4.6 Cross Validation

In verifying this article, 10 cross-validation methods are used to overcharge and avoid problems. *Cross-validation* is a evaluation and verification method [26] of honeypot detection. The structure of *cross-validation* is shown in the Figure 2. The framework of cross-validation consists of three parts. First, D data is divided into 10 identical subsets. These subsets come together exclusively when data is broken down, so split data can be passed on to training and testing. In the *training* and *testing* stages, 9 subsets are used for training, and the preceding subset is used for verification. Therefore, after validated, there are 10 results found and each classifier provides the expected results. Calculate the result from 10 results.

4.7 Comparing Experiment

This comparison experiment emphasizes the advantages of this method on white paper. Several machine learning studies, SVM, kNN, Naive bayes, J48 with the proposed Random Forest algorithm. Then, it uses the five algorithms of this machine to train four models, each of which is derived from the same set of data.

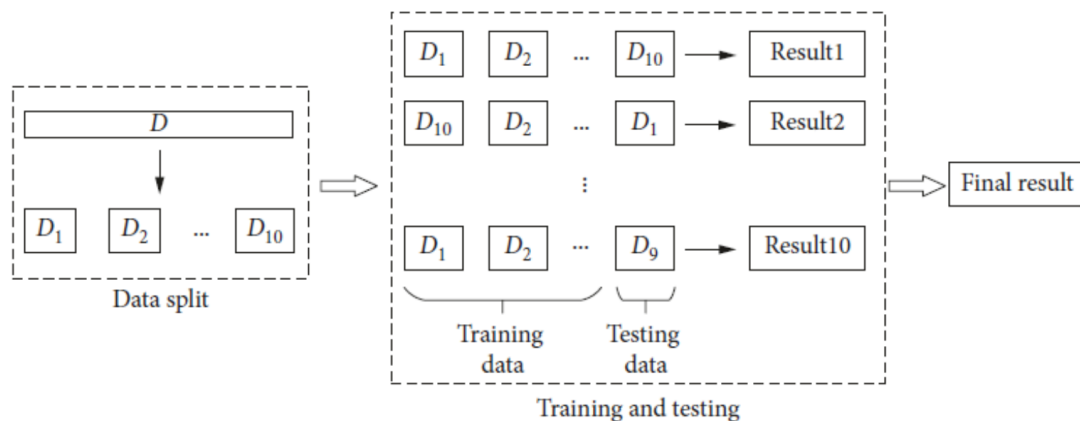


Fig. 2. Verification by 10-fold cross validation

4.8 Experiment Outcome

This portion shows the experimental results. First, tested each value of "n_estimators" and "max_depth". Figure-3 shows the effect of the other values of "n_estimators" on accuracy. The effect of the other "max_depth" values on accuracy is shown in Figure 3 and Figure 4. The testing data and training data are shown in different color. In the simulation result shown that when "n_estimators" are 210, the testing accuracy is very high as 0.91. When "max_depth" is 32, highest accuracy of 0.90.

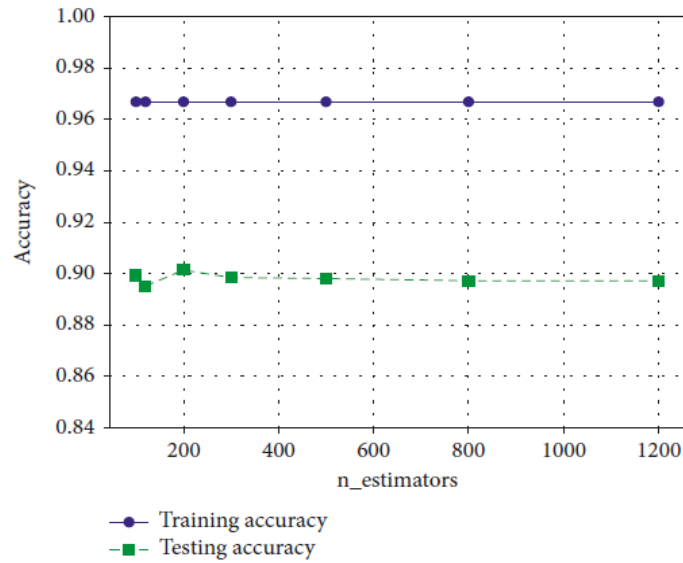


Fig. 3. Accuracy of “n_estimators”

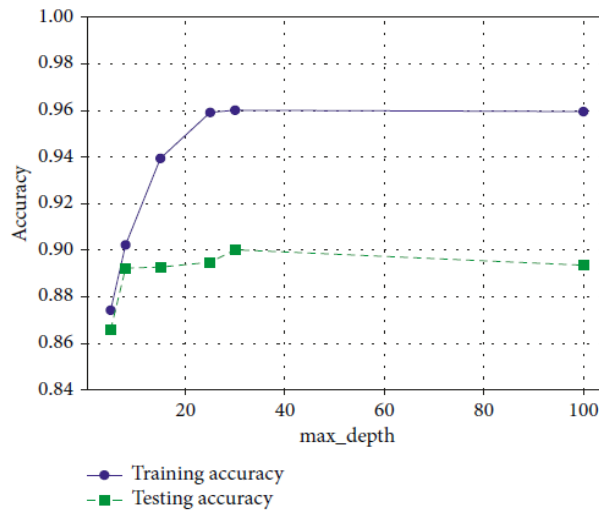


Fig. 4. Accuracy of “max_depth”

The accuracy of the linear regression algorithm proposed in Table 3 is 97%, the real positive rate is 84%, the F1 score is 79, and the recall rate is 82. Indicators show that the final model is high discovery rate and efficient technology in generalization.

Figure 5 shows the simulation performance metrics of various conventional algorithms such as J48, Naive Bayes, SVM, kNN, and Random Forest with the proposed Linear Regression (LR) decision algorithm. The proposed LR algorithm, based on linear shrinkage decision making can achieve up to 97% accuracy. The proposed algorithm is more effective by reducing the calculation cost and calculation time, and overcharging and avoiding this problem.

Table 3. Performance metrics of Existing and Proposed Machine learningmodel

Machine Learning Algorithm/ Performance metrics	Accuracy	TPR	Recall	F-Measure
SVM	86	84	82	79
Naive Bayes (NB)	90	87	83	82
kNN	91	90	88	86
J48	92	91	90	89
Random Forest (RF)	94	93	92	92
Linear Regression (LR) (Proposed)	97	96	96	95

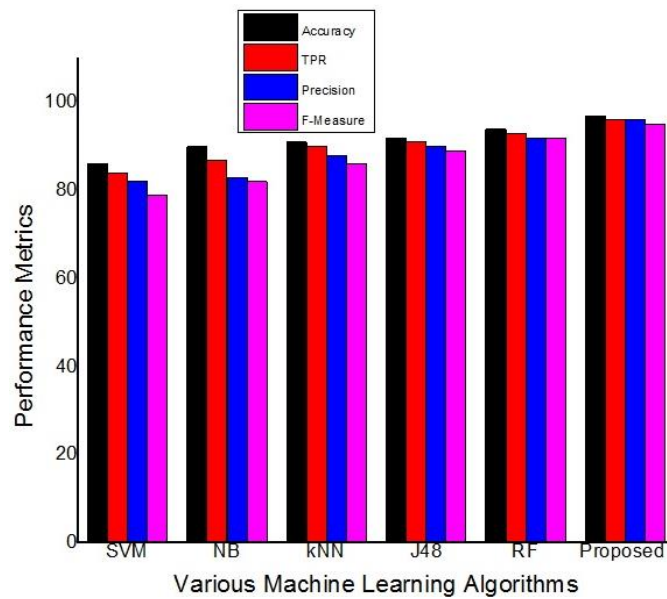


Fig. 5. Performance Metrics of ML Algorithms

5. Conclusions

In this paper, the proposed Linear Regression learning techniques provided the honeypot security network security solutions. The emergence of honeypot as an active strategy in defense of network security has further reduced the business space of attackers. Based on the research made on security attacker by honeypot detection, proposed a novel intelligent honeypot detection using different machine learning algorithm. This technology uses machine learning technology to retrieve and collect the various characteristics of honeypot to achieve accurate honeypot detection. The proposed algorithm can restore two important information that are used to create an configuration file, and the other is used to sort the configuration file. In fact, the combination of several honeypot-based solutions constitutes a robust modeling system and prediction for the identification and classification of suspicious profiles. Experimental results promoted the development of honeypot technology by providing reference materials for enhanced honeypot.

Declaration Section

Author's contribution:

A.M. contributed to technical and conceptual content, architectural design. R.T.G. contributed to guidance and counselling on the writing of the paper. S.S. contributed in discussing the result with other authors and revising the draft of manuscript.

Availability of data and materials

The data generated or analyzed during the current study is not publicly available due to restrictions in the ethical permit but may be available from the corresponding author on request.

Funding

No funding received by any government or private concern.

Conflict of interest:

The authors declare that they have no conflict of interest.

References

- [1] GData, *Malware Numbers*, <http://www.gdatasoftware.com>, 2017.
- [2] P. Owezarski, "Unsupervised classification and characterization of honeypot attacks," in *Proceedings of 10th International Conference on Network and Service Management (CNSM) and Workshop*, pp. 10–18, Rio de Janeiro, Brazil, November 2014.
- [3] S. Dowling, M. Schukat, and E. Barrett, "Improving adaptive honeypot functionality with efficient reinforcement learning parameters for automated malware," *Journal of Cyber Security Technology*, vol. 2, no. 2, pp. 75–91, 2018.
- [4] M. M. Matin and B. Rahardjo, "Malware detection using honeypot and machine learning," in *Proceedings of 2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Bandung Institute of Technology, Bandung, Indonesia, pp. 1–4, November 2019.
- [5] L. Spitzner, *Honeypots: Tracking Hackers*, Addison-Wesley, Clemson, SC, USA, 2003.

- [6] Mokube and M. Adams, "Honey pots: concepts, approaches, and challenges," in *Proceedings of the 45th Annual Southeast Regional Conference*, pp. 321–326, ACM, Winston-Salem, NC, USA, March 2007.
- [7] L. Spitzner, "The honeynet project: trapping the hackers," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 15–23, 2003.
- [8] O. Thonnard and M. Dacier, "A framework for attack patterns' discovery in honeynet data," *Digital Investigation*, vol. 5, pp. 128–139, 2008.
- [9] W. Fan, Z. Du, M. Smith-Creasey, and D. Fernandez, "HoneyDOC: an efficient honeypot architecture enabling all round design," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 683–697, 2019.
- [10] K. Sadasivam, B. Samudrala, and T. A. Yang, "Design of network security projects using honeypots," *Journal of Computing Sciences in Colleges*, vol. 20, pp. 282–293, 2005.
- [11] M. Mansoori, O. Zakaria, and A. Gani, "Improving exposure of intrusion deception system through implementation of hybrid honeypot," *The International Arab Journal of Information Technology*, vol. 9, no. 5, pp. 436–444, 2012.
- [12] G. Portokalidis and H. Bos, "SweetBait: zero-hour worm detection and containment using low- and high-interaction honeypots," *Computer Networks*, vol. 51, no. 5, pp. 1256–1274, 2007.
- [13] W. Fan, Z. Du, D. Fernandez, and V. A. Villagra, "Enabling an anatomic view to investigate honeypot systems: a survey," *IEEE Systems Journal*, vol. 12, no. 4, pp. 3906–3919, 2017.
- [14] M. L. Bringer, C. A. Chelmecki, and H. Fujinoki, "A survey: recent advances and future trends in honeypot research," *International Journal of Computer Network and Information Security*, vol. 4, no. 10, pp. 63–75, 2012.
- [15] P. Wang, L. Wu, R. Cunningham, and C. C. Zou, "Honeypot detection in advanced botnet attacks," *International Journal of Information and Computer Security*, vol. 4, no. 1, pp. 30–51, 2010.
- [16] K. Papazis and N. Chilamkurti, "Detecting indicators of deception in emulated monitoring systems," *Service Oriented Computing and Applications*, vol. 13, no. 1, pp. 17–29, 2019.
- [17] W. Fan and D. Fernandez, "A novel SDN based stealthy TCP connection handover mechanism for hybrid honeypot systems," in *Proceedings of the 2017 IEEE Conference on Network Softwarization (NetSoft)*, pp. 1–9, IEEE, Bologna, Italy, July 2017.
- [18] T. Luo, Z. Xu, X. Jin, Y. Jia, and X. Ouyang, "Iotcandyjar: towards an intelligent-interaction honeypot for iot devices," in *Proceedings of the Black Hat*, Las Vegas, NV, USA, 2017.
- [19] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots+ machine learning," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR'10*, pp. 435–442, the ACM Digital Library, New York; NY, USA, July 2010.
- [20] G. Feng, C. Zhang, and Q. Zhang, *A Design of Linkage Security Defense System Based on Honeypot: Trustworthy Computing and Services*, Springer, Berlin, Heidelberg, Germany, 2014.
- [21] L.-j. Li and H. Peng, "A defense model study based on IDS and firewall linkage," in *Proceedings of 2010 International Conference of Information Science and Management Engineering*, pp. 91–94, IEEE, Xi'an, China, August 2010.
- [22] J. Papalitsas, S. Rauti, and V. Leppanen, "A comparison of record and play honeypot designs," in *Proceedings of the 18th International Conference on Computer Systems and Technologies*, pp. 133–140, ACM, Ruse, Bulgaria, June 2017.
- [23] R. M. Campbell, K. Padayachee, and T. Masombuka, "A survey of honeypot research: trends and opportunities," in *Proceedings of the 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pp. 208–212, IEEE, London, UK, December 2015.
- [24] F. Y.-S. Lin, Y.-S. Wang, and M.-Y. Huang, "Effective proactive and reactive defense strategies against malicious attacks in a virtualized honeynet," *Journal of Applied Mathematics*, vol. 2013, Article ID 518213, 11 pages, 2013.
- [25] O. Surnin, F. Hussain, R. Hussain et al., "Probabilistic estimation of honeypot detection in Internet of things environment," in *Proceedings of the 2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 191–196, IEEE, Honolulu, HI, USA, February 2019.
- [26] Z. Wang, X. Feng, Y. Niu, C. Zhang, and J. Su, "TSMWD: a high-speed malicious web page detection system based on two-step classifiers," in *Proceedings of the 2017 International Conference on Networking and Network Applications (NaNA)*, pp. 170–175, IEEE, Kathmandu City, Nepal, October 2017.
- [27] D. Wenda and D. Ning, "A honeypot detection method based on characteristic analysis and environment detection," in *2011 International Conference in Electrics, Communication and Automatic Control Proceedings*, pp. 201–206, Springer, Berlin, Germany, 2012.
- [28] N. Provos, "A virtual honeypot framework," in *Proceedings of the USENIX Security Symposium*, vol. 173, pp. 1–14, San Diego, CA, USA, August 2004.
- [29] N. Krawetz, "Anti-honeypot technology," *IEEE Security & Privacy Magazine*, vol. 2, no. 1, pp. 76–79, 200

Authors' Profiles



Mr. Avijit Mondal is currently working as an Assistant Professor in the Computer Science and Engineering department of Techno International New Town, Kolkata. He has Completed his B.Tech in Information Technology from CIEM Tollygunge (under WBUT) and Completed his M.Tech in Computer Science from BIT Mesra . He has submitted his PhD Thesis in Computer Science from Maulana Abul Kalam Azad University of Technology (MAKAUT) (Registration No: PhD/Tech/CSEIT056/2018). He has Total 14 years of experience. His area of interest is Network Security, Cloud Security.



Dr. Radha Tamal Goswami Director Techno India College of Technology (Techno International New Town), Professor in the Department of Computer Science and Engineering, Newtown, Kolkata India. Dr. Radha Tamal Goswami has received his Ph.D. in Technology from Birla Institute of Technology Mesra Ranchi India. He is having 23 years of experience in the field of academics and research. He was the professor in Computer Science and Engineering and also the Director of BIT Mesra Kolkata Campus since 1995. He joined Techno India College of Technology Kolkata as a Director in September 1, 2016 on 2 years' lien from BIT Mesra. His-research interest in the field of Network Security and BigData. He has conducted almost 30 MDP and FDP program. He has guided more than 100 students in UG and PG projects. He is the visiting faculty of ten Institutions and member of ACM, IEEE, CSI and NIPM. Published almost 30 research papers. He chaired many National and International conferences.



Ms. Soumita Sen is currently working as an Assistant Professor in department of Computer Science and Engineering of Techno International Batanagar, Kolkata. She has Completed her B.Tech in Information Technology from MCKV Institute of Engineering (under WBUT) and completed her M.Tech in Computer Technology from Jadavpur University, Kolkata . She has total 8 years of experience. Her area of interest is Network Security.

How to cite this paper: Avijit Mondal, Radha Tamal Goswami, Soumita Sen, "A New Framework of Honeypots Network Security Using Linear Regression Decision Algorithm", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.13, No.6, pp. 23-31, 2023. DOI:10.5815/ijwmt.2023.06.03