CYBERHACKINGIDENTIFICATIONUSINGMACHINELEARNING

P.Vemulamma,AssistantProfessorCSE,VaagdeviCollegeofEngineering(Autonomous), India
T.Vivek,UGStudent,CSE, VaagdeviCollegeof Engineering(Autonomous),India
K.Nandini,UGStudent,CSE, VaagdeviCollege ofEngineering (Autonomous), India
B.Soniya,UGStudent, CSE,VaagdeviCollege of Engineering(Autonomous),India
K.Nagaraju,UGStudent, CSE, VaagdeviCollege ofEngineering(Autonomous), India

Abstract:
Cyber hacking breaches prediction is one of the emerging technologies and it has been a quite challenging task to recognize breaches detection and prediction using computer algorithms. Making malware detection more responsive, scalable, and efficient than traditional systems that call for human involvement is the main goal of applying machine learning for breaches detection and prediction. Various types of cyber hacking attacks any of them will harm a person's information and financial reputation. Data from governmental and non-profit organizations, such as user and company information, may be compromised, posing a risk to their finances and reputation. The information can be collected from websites that can trigger cyberattack. Organizations like the healthcare industry are able to contain sensitive data that needs to be kept discreet and safe. Identity theft, fraud, and other losses may be caused by data breaches. The findings indicate that 70% of breaches affect numerous organizations, including the healthcare industry. The analysis displays the likelihood of a data breach. Due to increased usage of computer applications, the security for host and network is leading to the risk of data breaches. Machine learning methods can be used to find these assaults. By research, machine learning models are utilized to protect the website from security flaws. The dataset can be obtained from the Privacy Rights Clearinghouse. Data breaches can be decreased by educating staff on the use of modern security measures. This can aid in understanding the attacks knowledge and data security. The machine learning models like Random Forest, Decision Tree, k-means and Multi- layer Perceptron are used to predict the data breaches.

INTRODUCTION

Databreachesareone ofthemostdevastatingcyberincidents.ThePrivacyRightsClearinghouse [1] reports 7,730 data breaches between 2005 and 2017, accounting for 9,919,228,821 breached records. The Identity Theft Resource Center and Cyber Scout reports 1,093 data breach incidents in 2016, which is 40% higher than the 780 data breach incidents in 2015. The United States Office of Personnel Management (OPM) [2] reports that the personnel information of 4.2 million current and former Federal government employees and the background investigation records of current,former,andprospectivefederalemployeesandcontractors(including21.5millionSocial Security Numbers) were stolen in 2015. The monetary price incurred by data breaches is also substantial. IBM [3] reports that in year 2016, the global average cost for each lost or stolen record containing sensitive or confidential information was $158. NetDiligence reports that in year2016,themediannumberofbreachedrecordswas1,339,themedianper-recordcostwas $39.82, the average breach cost was $665,000, and the median breach cost was $60,000 [4]. While technological solutions can harden cyber systems against attacks, data breaches continueto beabig problem. This motivates us to characterizetheevolution ofdata breach incidents. This not only willdeep ourunderstanding ofdatabreaches, but also shed lighton otherapproaches for mitigating the damage, such as insurance [5]. Many believe that insurance will be useful, but the development of accurate cyber risk metrics to guide the assignment of insurance rates is beyond the reach of the current understanding of data breaches (e.g., the lack of modeling approaches). Recently, researchers started modeling data breach incidents. The statistical properties of the personal identity losses in the United States between year 2000 and 2008. They found that the number of breach incidents dramatically increases from 2000 to July 2006 but remains stable thereafter.

LITERATURESURVEY

1. Our paper predicts cyber hacking breaches. In addition to posing a threat to personal and financial security, data breaches can be costly for organizations that keep large amounts of personal data. Researchers and practitioners alike have argued for robust and innovative cyber- insurance pricing models to manage residual IT security risks. However, the accuracy of premiumsremainsanopenquestion.In2011,the paperdevelopedacyber-insurancemodelusing the emerging copula methodology, filling an important scholarly gap. In 2015, we identified two distinct spatiotemporal patterns based on macroscopic analysis of attack traffic flows: deterministic and stochastic patterns. In this approach, a gray box model is recommended to accommodate statistical properties/phenomena exhibited by the data. The methodologies we use in our prediction are often equally applicable to the analysis of any cyber attack data, eventhough the predictions are based on specific cyber attack data. There has been an increase in data breach incidents in 2015, leading to severe financial and legal repercussions for the affected organizations. Extreme values, extreme value theory, prediction, gray-box models, time series. Index Terms In 2015, many thousands of people have lost their private information as a result of data breaches as a result of the opportunity theory of crime, institutional anomie theory, and institutional theory. According to some reports, there have been alarming increases in the sizeand frequency of knowledge breaches. This has forced institutions worldwide to respond to what appears to be a worsening situation. The economy, human privacy, and even national security have been threatened by cyber attacks, which have become a drag. It is crucial that we have a solid understanding of cyber attacks from a variety of perspectives in 2017 before we can adequately deal with the issue. This issue can be difficult to model. A study of multivariate cybersecurity risks is presentedin this paper. Inourfirst statistical approach, weusevine copulas to simulate the multivariate dependence observed by real-world cyberattack data in 2018, using the Copula-GARCH model. Our current method of predicting breach size and interarrival time is a stochastic process model.

2. Cyber hacking identification is vital, however difficult, problems. during this paper, we tend to initiate the study of modeling and predicting cyber hacking breaches. In this study we tend to plan a theoretical account model to predict each hacking breach incident lay to rest arrival times and breach sizes, here we'll use each qualitative and quantitative analysis on the information set.

3. Analyzing cyber incident data sets is an important method for deepening our understanding of the evolution of the threat situation. This is a relatively new research topic, and many studies remain to be done. In this paper, we report a statistical analysis of a breach incident data set corresponding to 12 years (2005-2017) of cyber hacking activities that include malware attacks. We show that, in contrast to the findings reported in the literature, both hacking breach incident inter-arrival times and breach sizes should be modeled by stochastic processes, rather than by distributions because they exhibit autocorrelations. Then, we propose particular stochasticprocess models to, respectively, fit theinter-arrival timesand the breach sizes. Wealso showthat these models can predict the inter-arrival times and the breach sizes. In order to get deeper insights into the evolution of hacking breach incidents, we conduct both qualitative and quantitativetrend analyses on thedataset. We drawaset ofcybersecurity insights, including that the threat of cyber hacks is indeed getting worse in terms of their frequency, but not in terms of the magnitude of their damage.

4. Using cyber occurrence informational indexes as a means of expanding our understandingof the current threat situation is essential. More research is needed on this topic before it can be considered a finished research project .In this, we employ artificial neural networks (ANN) to forecast malware assaults and data breaches. Since autocorrelations are shown by both hacker breakoccurrenceandpenetrationsize,weshowthatstochasticprocesses ratherthanconveyances should be used to depict these data. Stochastic method models are proposed to fit the break sizes and between appearance times. The informational index is being subjected to both subjective and quantitative

pattern inspections in order to gather more information about the progression of hacker penetration occurrences. Despite the fact that cyber-attacks are becoming more frequent, they aren't getting much worse in terms of the damage they can cause.

PROBLEMSTATEMENT

The present study is motivated by several questions that have not been investigated until now, such as: Are data breaches caused by cyber-attacks increasing, decreasing, or stabilizing? A principled answer to this question will give us a clear insight into the overall situation of cyber threats.Thisquestionwasnotansweredbypreviousstudies.Specifically,thedatasetanalyzedin [7] only covered the time span from 2000 to 2008 and does not necessarily contain the breach incidents that are caused by cyber-attacks; the dataset analyzed in [9]is morerecent, but contains two kinds of incidents: negligent breaches (i.e., incidents caused by lost, discarded, stolendevices and other reasons) and malicious breaching. Since negligent breaches represent more human errors than cyber-attacks, we do not consider them in the present study. Because the malicious breaches studied in [9] contain four sub-categories: hacking (including malware), insider, payment card fraud, and unknown, this study will focus on the hacking sub-category (called hacking breach dataset thereafter), while noting that the other three sub-categories are interesting on their own and should be analyzed separately.Recently, researchers startedmodeling data breach incidents. Maillart and Sornette studied the statistical properties of the personal identity losses in the United States between year 2000 and 2008. They found that the number of breach incidents dramatically increases from 2000 to July 2006 but remains stable thereafter. Edwards et al. analyzed a dataset containing 2,253 breach incidents that span over a decade (2005 to 2015). They found that neither the size nor the frequency of data breaches has increased over the years. Wheatley et al.,analyzed a dataset that is combined from corresponds to organizational breach incidents between year2000 and 2015. They found that thefrequency of large breach incidents (i.e., the ones that breach more than 50,000 records) occurring to US firms is independent of time, but the frequency of large breach incidents occurring to non-US firms exhibits an increasing trend.

LIMITATIONOF SYSTEM

- Mentioningthebreachsize.
- Wedon’tknowhowitwas hacked.

PROPOSEDSYSTEM

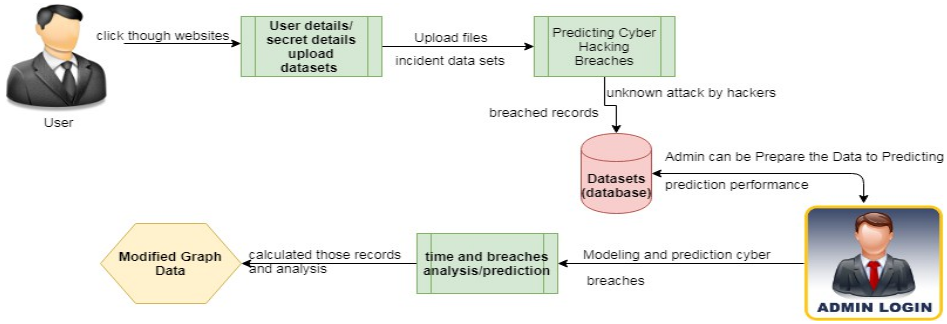
In this paper, we make the following three contributions. First, we show that both the hacking breach incident interarrival times (reflecting incident frequency) and breach sizes should be modeled by stochastic processes, rather than by distributions. We find that a particular point process can adequately describe the evolution of the hacking breach incidents inter-arrival times andthat aparticularARMA-GARCHmodel can adequately describetheevolutionofthehacking breach sizes, where ARMA is acronym for “AutoRegressive and Moving Average” and GARCH is acronym for “Generalized AutoRegressive Conditional Heteroskedasticity.”We show thatthese stochastic process models can predict the inter-arrival times and the breach sizes. To the best of our knowledge, this is the first paper showing that stochastic processes, rather than distributions, should be used to model these cyber threat factors. Second, we discover a positive dependence between the incidents inter-arrival times and the breach sizes, and show that this dependence can be adequately described by a particular copula. We also show that when predicting inter-arrival times and breach sizes, it is necessary to consider the dependence; otherwise, the prediction results are not accurate. To the best of our knowledge, this is the first work showing the existence of this dependence and the

consequence of ignoring it. Third, we conduct both qualitative and quantitative trend analyses of the cyber hacking breach incidents. We find that the situation is indeed getting worse in terms of the incidents inter-arrival time because hacking breach incidents become more and more frequent, but the situation is stabilizing in terms of the incident breach size, indicating that the damage of individual hacking breach incidents will not get much worse. We hope the present study will inspire more investigations, which can offer deep insights into alternate risk mitigation approaches. Such insights are usefulto insurance companies, government agencies, and regulators because they need to deeply understand the nature of data breach risks.

ADVANTAGES

- Knowingtheinter-arrivaltimesandthebreachsizes.
- Bothqualitativeandquantitativetrend analysesofthecyberhackingbreach

SYSTEMARCHITECTURE



IMPLEMENTATION

UPLOADDATA

Thedataresourcetodatabasecanbeuploadedby bothadministratorand authorizeduser.The data can be uploaded with key in order to maintain the secrecy of the data that is not released without knowledge of user. The users are authorized based on their details that are shared to admin and admin can authorize each user. Only Authorized users are allowed to access the system and upload or request for files.

ACCESSDETAILS

The access of data from the database can be given by administrators. Uploaded data are managedby admin and administheonlypersontoprovidetherightsto processtheaccessing details and approve or unapproved users based on their details.

USER PERMISSIONS

The data from any resources are allowed to access the data with only permission from administrator. Prior to access data, users are allowed by admin to share their data and verify the details which are provided by user. If user is access the data with wrong attempts then, usersareblocked accordingly. Ifuseris requested to unblockthem, based on therequests and previous activities admin is unblock users.

DATAANALYSIS

Data analyses are done with the help of graph. The collected data are applied to graph inorder to get the best analysis and prediction of dataset and given data policies. The datasetcan be analyzed through this pictorial representation in order to better understand of the data details.

ALGORITHMUSED

SUPPORTVECTOR MACHINE

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin),since in general thelargerthe margin thelowerthegeneralization errorofthe classifier.Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher- dimensional space, presumably making the separation easier in that space.

EXPECTEDRESULTS

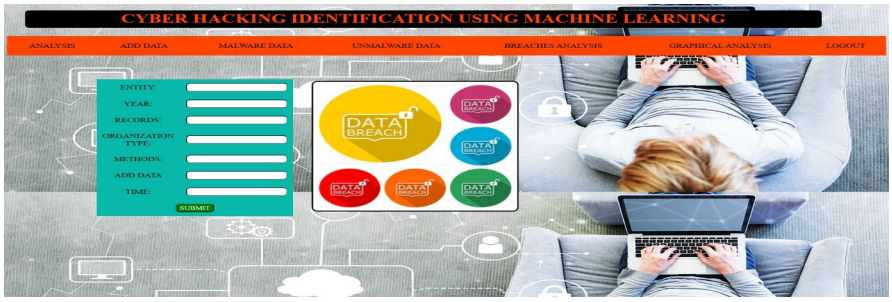


Fig8.1Home page

Fig8.2UserLoginpage

Fig8.3UserAccountcreation

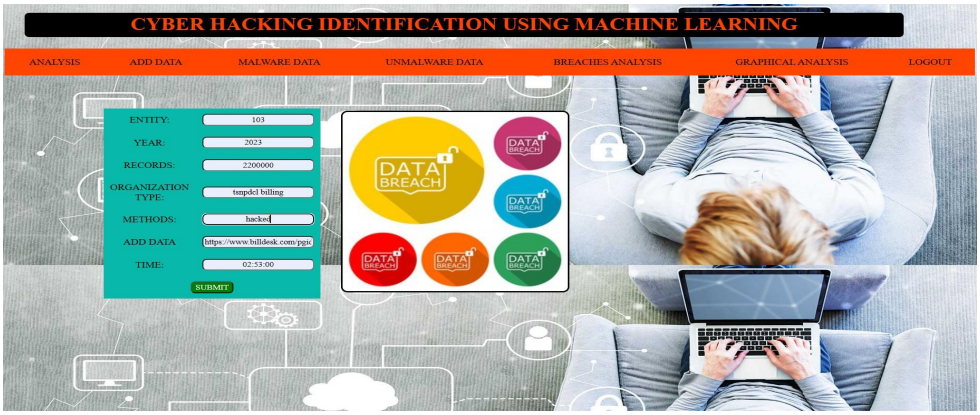


Fig8.4AddDatapage

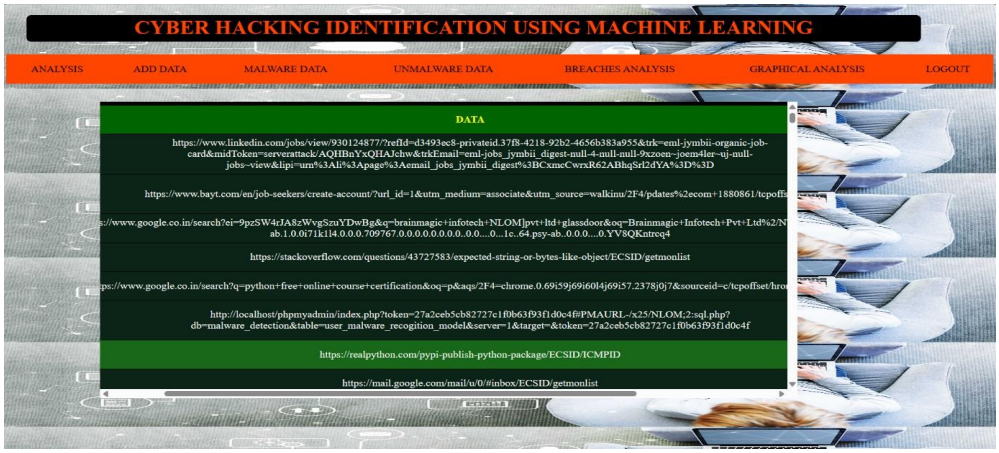


Fig8.5AnalysisPage

Fig8.6Unmalware Analysis

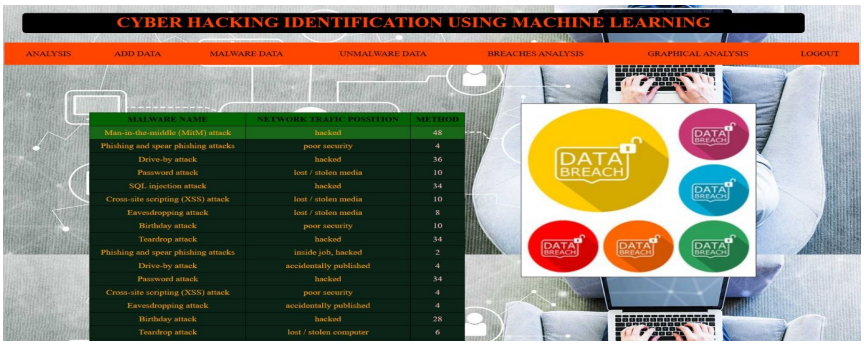


Fig8.7BreachAnalysisPage

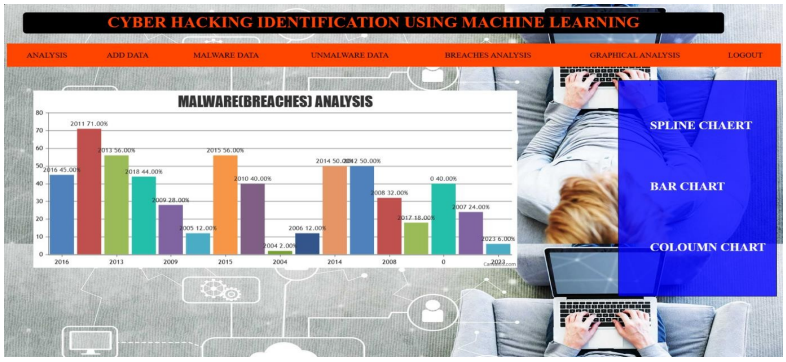


Fig8.8Barchart -Breach analysis

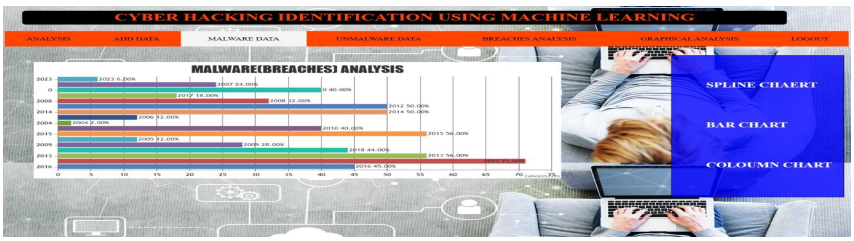


Fig8.9Columnchart- Breachanalysis

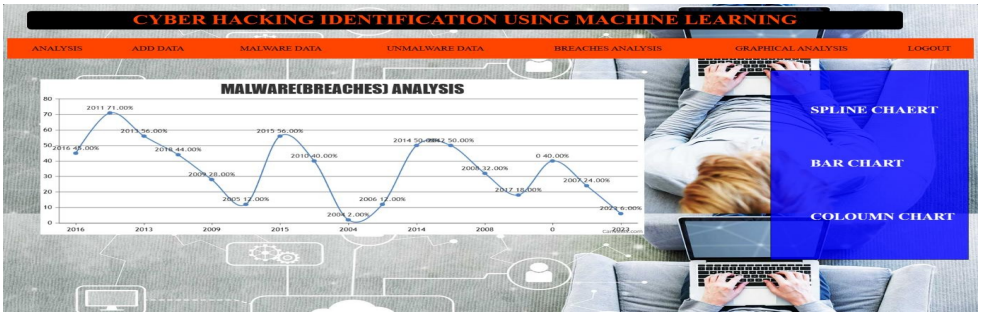


Fig8.10Splinechart-Breachanalysis

Fig8.11Admin Login

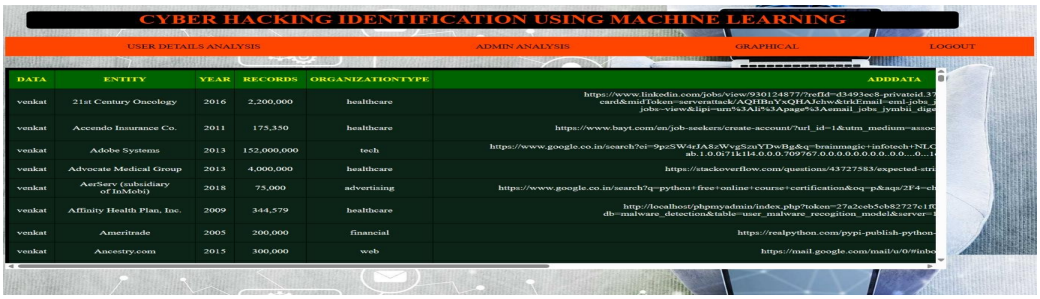


Fig8.12Userdetailsanalysis

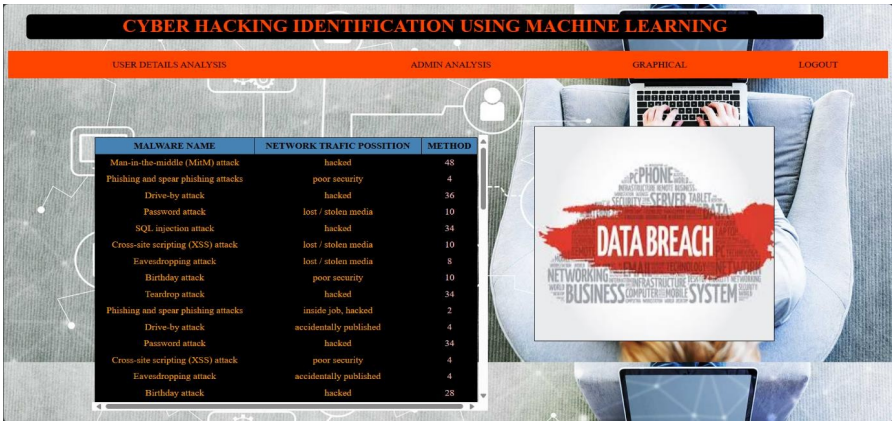


Fig8.13Admin analysis

CONCLUSION

We analyzed a hacking breach dataset from the points of view of the incidents inter- arrival time and the breach size, and showed that they both should be modeled by stochastic processes rather than distributions. The statistical models developed in this paper show satisfactory fitting and prediction accuracies. In particular, we propose using a copula-based approach to predict the joint probability that an incident with a certain magnitude of breach size will occur during a future period of time. Statistical tests show that the methodologies proposedin this paper are better than those which are presented in the literature, because the latter ignored both the temporal correlations and the dependence between the incidents inter-arrival times and the breach sizes. We conducted qualitative and quantitative analyses to draw further insights. We drew a set of cybersecurity insights, including that the threat of cyber hacking breach incidents is indeed getting worse in terms of their frequency, but not the magnitude of their damage. The methodology presented in this paper can be adopted or adapted to analyze datasets of a similar nature.

FUTURESCOPE

There are many open problems that are left for future research. For example, it is both interesting and challenging to investigate how to predict the extremely large values and how to deal with missing data (i.e., breach incidents that are not reported). It is also worthwhile to estimate the exact occurring times of breach incidents. Finally, more research needs to be conducted towards understanding the predictability of breach incidents (i.e., the upper bound of prediction accuracy

REFERENCES

[1] F.Y.Leu,J.C.Lin,M.C.Li,C.TYang,P.CShih,"IntegratingGridwithIntrusionDetection," Proc. 19thInternational Conference on Advanced Information Networking and Applications, pp. 304-309, 2005.

[2] Whitepaper,"IntrusionDetection:ASurvey," ch.2,DAAD19-01,NSF, 2002.

[3] K. Scarfone, P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," NIST Special Publication800-94, Feb. 2007.

[4] IBM Security.Accessed: Nov. 2017 [Online]. Available: https://www.ibm.com/security/databreach/index.html

[5] NetDiligence.The2016Cyber ClaimsStudy.Accessed:Nov.201710/P02_NetDiligence- 2016-Cyber-Claims-Study-ONLINE.pdf

[6] M. Eling and W. Schnell, "What do we know about cyber risk and cyber risk insurance?" J.Risk Finance, vol. 17, no. 5, pp. 474– 491, 2016.

[7] https://ieeexplore.ieee.org/document/8360172#:~:text=Modeling%20and%20Predicting%20Cyber%20Hacking%20Breaches%20Abstract%3A%20Analyzing,topic2C%20and%20many%20studies%20remain%20to%20be%20done.

[8] Okamgba, J., 2017. Online Fraud Drains Nigeria over N500 Billion in 7 Years. Retrieved from. https://cfatech.ng/online-fraud-drainsnigeria-over-n500- billion-in-7-years/.

[9] Okoh, J., Chukwueke, E.D., 2016. The Nigerian Cybercrime Act 2015 and its Implication for Financial Institutions and Service Providers. Financier Worldwide. Retrieved from. https://www.financierworldwide.com/thenigeriancybercrime-act-2015-and- itsimplications-for-financialinstitutions-and-serviceproviders#.

[10] Olasanmi, O.O., 2010. Computer crimes and counter measures in the Nigerian banking sector. J. Internet Bank. Commer. 15 (1), 1–10.

[11] Olawoyin,O.,2017.NorthKoreanHackersAttackBanksinNigeria,17OtherCountries– Kaspersky.PremiumTimesRetrievedfrom.https://www.premiumtimesng.com/news/topnews/ 22816 6-north-korean-hackers-attack-banks-innigeria-17-othercountries-kspersky.html.

[12] Olayemi, O.J., 2014. A socio-technological analysis of cybercrime and cyber security in Nigeria. Int. J. Sociol. Anthropol. 6 (3), 116–125.
[13] Omodunbi, B.A., Odiase, P.O., Olaniyan, O.M., Esan, A.O., 2016. Cybercrimes in Nigeria: analysis, detection and prevention. FUOYE J. Eng. Technol. 1 (1), 37–42.
[14] Omotubora, A.O., 2016. Comparative perspectives on cybercrime legislation in Nigeria and the UK-a case for revisiting the" hacking" offences under the Nigerian Cybercrime Act 2015.Eur. J. Law Technol. 7 (3), 1–15.
[15] Oni, A.A., Ayo, C.K., 2010. An empirical investigation of the level of users' acceptance of ebanking in Nigeria. J. Internet Bank. Commer. 15, 1–13.
[16] T Manikandan, B Balamurugan, C Senthilkumar, RRA Harinarayan, RR Subramanian, “Cyber War is Coming”, Cyber Security in Parallel and Distributed Computing: Concepts, Techniques, Applications and Case Studies, John Wiley & Sons, Inc, pp. 79-89, Mar. 2019 .

