International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Design and Implementation of Machine Learning and Big Data Analytics models for Cloud Computing platforms

[1]Osama Yassin Mohammed, [2]Hather Ibraheem Abed, [3]Nawar A. Sultan

Submitted: 16/02/2023 **Revised:** 20/04/2023 **Accepted:** 09/05/2023

Abstract: There are concerns regarding the protection of significant digital assets because it is noticeable that cyber attackers are outstripping defenses. AI models need specialized cybersecurity and defense keys in order to lower dangers, enhance information privacy, and enable a secure federated knowledge situation. The growth of "artificial intelligence" has controlled the rise of several new fields, including "machine learning" (ML), "natural language processing" (NLP), CPU vision, and several more. Massive volumes of information are produced in the "Internet of Things" (IoT) age and are acquired from a variety of heterogeneous sources, including "mobile devices", "sensors", and "social media". Big Data faces significant challenges in terms of processing, storage, and analytical capabilities. Firewall security has proven to be insufficient as a result of significant restrictions against external attackers. Given that computer worms and viruses, which are intelligent semi-autonomous agents, are responsible for the majority of network-centric cyberattacks, it has become necessary to combat them with intellectual semi-autonomous mediators that can identify, evaluate, and respond to cyberattacks. Specified that the majority of network-centric cyberattacks are produced by CPU worms and illnesses, which are intelligent semi-autonomous agents, it has become important to combat them with intelligent semi-autonomous agents that can recognise, assess, and respond to cyberattacks. The study aims to determine the Big Data Analytics and Machine Learning paradigms for usage in cybersecurity and to make use of big data analytics and machine learning.

This study used a case study research methodology. This is because each statistics analytics model for cybersecurity is observed as a unique case that needs to be examined in its own context. Case studies have been used a lot in prior research on cybersecurity. The investigator knows two data analytics models or frameworks through a review of the literature and a study of the 8-person sample. There were eight people interviewed in total. Even if there might not be much data, it is sufficient for the current stages of this investigation. Future studies may look at other publications to discover other data and analytics models that are pertinent to cybersecurity.

Table 4 illustrates the overall show of our CART procedure in predicting scores on the banks on the training dataset. On the testing data, the precision of the approach was 83.1%. On the test dataset, the approach got a kappa of 88.76% and a reliability of 92.7%. On the training dataset, the SVM figure's overall accuracy for predicting bank health remained at 79.1%. The Kappa statistic and Kappa SD were 67.6% and 0.14, combined. On the testing data, the approach had an efficiency of 92.7% and a kappa of 88.59%. The efficiency of the randomized forest in the training phase was 85.57 %. Large amounts of information are is for connections and trends, leading to the creation of algorithms for these kinds of relationships and patterns. Passive sources of information include laptop data such as IP address, information security certifications, keypad usage, clickstream trends, and WAP data.

Keywords*: large statistics, engine learning, natural linguistic processing, procedure, cybersecurity.*

1. Introduction

Data protection is a combination of policies, individual behavior, and technological advances designed to protect electronic information assets. Clearly, intrusion detection systems are exceeding barriers, which poses concerns about the safety of vital digital products [1]. Deep education tools could be used to make multifaceted models for malware classification, cyber threat sensing, and intrusion discovery in a mobile net. To reduce risks, improve information privacy, and enable a secure federated learning environment, AI models need specialized cybersecurity and protection solutions. ML, NLP, computer vision, and many more fields have emerged as a result of the development of "artificial intelligence" [2]

Big Data analytics is an emerging field that attempts to use machine learning and large datasets to solve complex problems. Central to these analyses is the development of predictive models, which can be used for risk assessment or fraud detection. However, many organizations still have limited resources in data management and/or computational power for big data analytics purposes [3]. This introduces the question of cloud computing platforms as a solution. Cloud computing is an approach to computing whereby resources are provided as a service over the Internet, rather than on-premise. It involves

[1]*The specialty of Computer Science, Northern Technical University, Mosul, Iraq*
Email: osama.yassin@ntu.edu.iq
[2]*The specialty of Computer Science, Northern Technical University, Mosul, Iraq*
Email: hather@ntu.edu.iq
[3]*The specialty of Computer Science, Northern Technical University, Mosul, Iraq*
Email: nawarabd@ntu.edu.iq

"enabling organizations to pay only for the resources they need, while they use their own applications and data to deliver value", which can be achieved through various types of cloud services. The objective of big data analytics is to use available resources efficiently and predictively. The cloud provides a dynamic environment in which all these components are present, thus creating a readymade, economically efficient solution for such problems [4,5].

The “Internet of Things” (IoT) era produces enormous capacities of figures that are gathered from numerous diversified bases, like sensors, social media, and mobiles. The storage, processing, and analytical capacities of this Big Data face enormous obstacles. In order to support the storage of Big Data and the operation of applications of data, cloud computing offers a practical and affordable alternative. IoT involves the utilization of “artificial intelligence” (AI) for data analytics and information mining, as well as a cloud computing environment to handle data interchange and processing. However, AI offers valuable improvements to traditional cybersecurity that is hampered by the networking of IoT devices and cloud weaknesses. Unhappily, hackers are also using AI to undermine cybersecurity. The intricacy of cybersecurity vulnerabilities and the amount of cloud-based data increases astronomically with the exponential growth of IoT device numbers. The problem is made worse by the fact that IoT devices lack proper cybersecurity protections. Vulnerabilities in IoT devices create a huge number of opportunities for cybersecurity threats and other types of crimes, particularly among the networked devices that are already commonplace in homes.

Cloud computing and cybersecurity platforms are the main focus of the study paper. Evaluation of the Big Data Analytics and Machine Learning paradigms for Cybersecurity is the goal of the project. This is relevant as we examine the effectiveness of effective and affordable cloud computing platforms and services due to the quick advancements in “deep learning” (DL) and “machine learning” (ML). This research paper's main focus is on evaluating defense systems and attacks using Big Data paradigms and ML.

Due to severe constraints against external attacks, firewall security has proven to be insufficient. Given that computer worms and viruses, which are intelligent semi-autonomous agents, are responsible for the majority of network-centric cyberattacks, it has become necessary to fight them with intellectual semi-autonomous managers that can recognize, evaluate, and respond to cyberattacks. The majority of organizations now must update their cyber defense strategy due to the fast growth of calculating and digital equipment. As a result, safety system managers must be more flexible, agile, and capable of offering strong cyber defense systems that can detect cyber threats in real-time. Assessing “big data analytics” (BDA) and “machine learning” (ML) examples for use in cybersecurity is the main issue.

Aim of the Study

1. The study aims to assess the “Big Data Analytics” and “Machine Learning” paradigms for usage in cybersecurity.
2. Create a cybersecurity system that makes use of big data analytics and engineering knowledge.

Questions of the Study

1. Which Big Data and Machine Learning paradigms are best for creating a cybersecurity system?
2. How are the big data analytics and machine learning paradigms employed in cybersecurity?

Significance of the Study

The study is conducted mainly to analyze the significance of machine learning and big data analytics while using cybersecurity.

2. Literature Review

All information-providing systems created for the use of humanity, including computers, phones, the internet, and others, are vulnerable to criminal activities. Computer invasions, infringement of logical property rights, online extortion, financial espionage, global money laundering, disaster in delivering products and services, etc. are all considered cybercrimes [6]. Systems for detecting and preventing intrusions (IDPS) encompass all preventative measures, the detection of potential incidents, and the analysis of incident log data. suggests that several security control mechanisms be used in an organization. The constant rise in a variety of cyber threats and viruses clearly shows how inadequate the current defenses are for protecting computer networks and resources [7]. It is desired to do a study in “artificial intelligence”, extra specifically device education, to address the issues with traditional cyber security methods.

In order to facilitate company continuity and genuine demands, high-performance computing has grown to be a significant sector for business and computer technology. To guarantee the system's availability at all times, several top-tier business and technology groups are still working to improve high performance and traffic resiliency. A significant development in computer technology is machine learning, which supports administration through forecast and categorization mechanisms built on past information. The idea of incorporating advanced computer technology with “artificial intelligence” methods for device learning on cloud platforms was proposed and integrated through research. To validate, forecast, and categorize traffic and performance patterns,

as well as to guarantee the performance of the system and uninterrupted flow of traffic for resiliency decisions, networking and computational presentation information are used. The study found that when compared to currently use non-machine learning-based model models, our engine learning combined project reproduction outcomes demonstrate its movement resilience operates effectively 38.17% quicker with regard to the failing point's restoration as well as 7.7% business cost savings [8].

Author Voros (2021) presented a concept 2021 that uses cloud-based dockers and microservices to distribute high-end hardware-specific resources [9]. Similarly to this, Sarangarajan et al. (2021) explore how machine learning models can be used to automate training and competency for businesses and customers in order to attain high computational performance [10]. Author Walker et al. (2021) make a distinction between the idea that allows for the selection of a high show and high flexibility in the same year. While maximum performance is an efficient wonder in industrial, research, and imitation technical domains, resilience is a more fundamental and accepted occurrence in the majority of trade, medicinal, and real-time areas [11].

A study was conducted on the two frontiers- cloud computing and big, and it reviews the consequences and advantages of intercepting big data analytics using cloud computing. The investigation reveals the following findings: (i) Big Data, spatiotemporal thoughts, and diverse application contexts motivate the progression of cloud technology and related technologies with new standards. (ii) Big Data and cloud computing enable scientific findings and developing applications. (iii) Inherent spatial-temporal fundamentals of Big Data and geospatial fields of science can provide a source for discovering technical and significant remedies for Big Data [12].

3. Methodology

Research design: In order to comprehensively characterize the truths and personalities of big statistics analytics simulations for cybersecurity, the researcher uses a descriptive study design. The study's main goal is to describe the models in great detail.

Research method: This study used a case study methodology. In this regard, the individual statistics analytics classic for cybersecurity is treated as a distinct instance that must be looked into within a context all of its own. In previous studies on cybersecurity, case studies have frequently been used. To allow for comparison, the researcher creates a control case that takes into consideration the optimum data analytics approach for cybersecurity.

a. *Sampling and population:* All statistics analytics prototypes for cybersecurity that have been advised and advanced in works of mechanism, papers, conference minutes, and employed identifications make up the study individuals for this knowledge. This is in line with an earlier study that involved a thorough literature review.

   Sampling: From a survey of the literature and an analysis of the 8-person example, the investigator exposed two statistics analytics models or frameworks. Interviews with a total of eight people were conducted. Even though there may not be much information, it will be enough used for the resolutions of this education at this time. Forthcoming research might study other papers to find more cybersecurity-relevant information and analytics models.

b. *Data types and sources:* In order to discover the procedure of numbers analytics models in cybersecurity, the academic habits are secondary statistics.

c. *Data analysis:* The researcher refers to the assets of a perfect information analytics model for cybersecurity when examining the various information analytics models for cybersecurity. The researcher integrates many literature sources when creating an ideal model. Big data, analytics, and insights make up the three main building blocks of the fundamental big data analytics paradigm for cybersecurity. However, prediction (or predictive analytics) may be recognized as a fourth element. The following Figure 1 shows this:

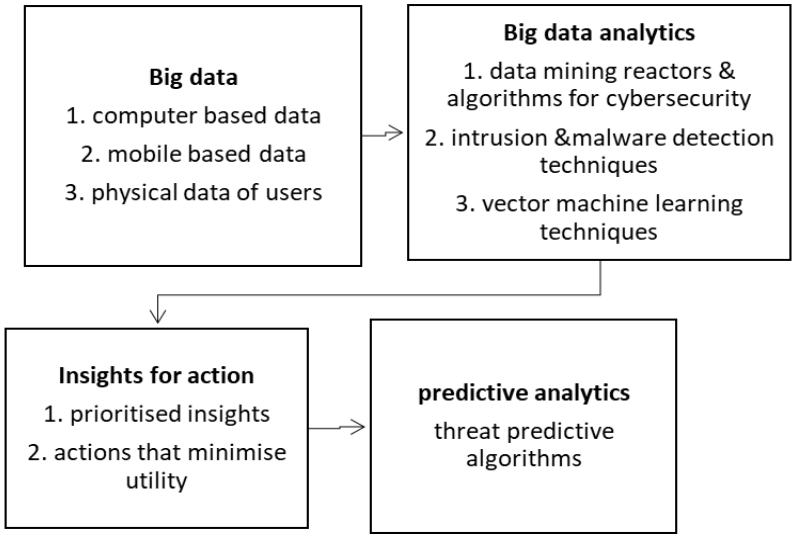

Fig 1: Model for Big Data Analytics in Cybersecurity huge data

Big data: The presence of big data about cyber is the basis of the framework for big data analytics in safety. Standard kinds of big data include system information and security assessments. Data from desktops, portable devices, individuals' actual movements, and data from human resource departments, data encryption, identifiers, and social networks are now also sources of big data for protection. Several authors have referred to corporate mail, access control systems "Customer relationship management" (CRM), "human resource management" structures, a range of pullers in relational data networks, internal networks, and "IIoT/IoT", collectors and processors in social media networks, and foreign news tapes as providers of safety big data.

"Big data analytics": To address the problems with big data security, data gathering and computational mathematics have been used to make more credible data insights for big data. Through big data analytics, humans are using data mining nuclear power stations and computational methods, ways to find computer viruses and intrusions, and methodologies for cybercrime vector deep learning. Still, it has observed that types of malware tend to change their behavior by trying to adapt to the reactor designs as well as machine learning techniques made to find them. However, systems for intrusion detection face problems like endless trends, data structure presents, unequal time lags, uniqueness, rising false alarm rates, and Russian interference attacks, which require a multilayer or multi-dimensional strategy for cybersecurity.

Predictive analysis: Applying a big data analytics model for cybersecurity to current cybersecurity data to determine the possibility of a cybersecurity event occurring in the future is known as predictive analytics.

d. *Reliability and validity:* In order to establish the internal validity of the study, the researcher asked for input from peers on the developing findings as well as comments to clarify the researcher's biases and assumptions. The researcher also provides consistency or reliability in research findings by thoroughly outlining the underlying hypotheses and theories.

e. *Possible outcomes:* Table 1 below, which displays the international benchmark, illustrates the projected accuracy rate for the research.

Table 1 is the taken parative Discovery correctness rate (%), the period taken to shape the model, and the wrong fright rate

Classifier	Detection accuracy (%)	Time taken to build the model in seconds	False alarm rate (%)
Decision trees (J48)	82.31%	**	**
Random forest	81.20%	**	**
AdaBoost	90.54%	**	3.45

SVM	70.12%	**	**
Naive Bayes	77.23%	**	**
Multinomal Naive Bayes updateable + N2B	39.12%	1.3	28.1
Mutlinomal Naive Bayes + N2B	39.02%	0.75	28.0
Discriminative Multinomal Bayes + RP	82.34%	2.45	13.10
Discriminative Multinomal Bayes + PCA	95.12%	119.12	4.6
Discriminative Multinomal Bayes + N2B	96.7	1.34	3.5

** not applicable

4. Results

The straightforward guidelines for the examination of attacks are detailed in Figure 2 below. In an anomaly-based intrusion detection system, an alarm is set off on the IDS when an unusual behavior on the network occurs.

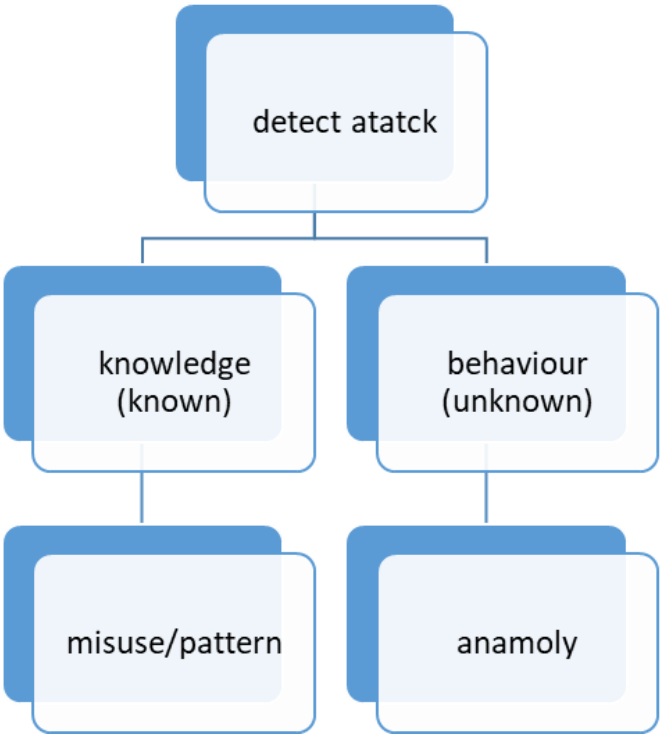

Fig: 2 analysis of the attack

The outcomes of the CART algorithm in predicting financial catastrophe on the exercise data set are presented in Table 2 under. The process's efficiency on the training examples was 83.1%. The ideal model's optimal variable for tweaking or difficulty was 0.069. The learner was successful, as evidenced by the Kappa statistic of 75.2% and the Kappa confidence interval of 0.08 for the identification of bank types. The approach had a kappa of

88.76% and a reliability of 92.7% on the testing data. The system misclassified only two cases as medium and one occasion as excellent.

Table 2: CART typical performances

Complexity parameter	Accuracy	Kappa	AccuracySD	KappaSD
0.06849325	0.8275089	0.7519497	0.04976462	0.07072574
0.15753436	0.7783155	0.6683232	0.07720891	0.14039945
0.42465786	0.5222352	0.1148594	0.08183356	0.1873426

As per Table 3, the SVM figure's precision for financial institution stability based on the training data was 79.1%. The Kappa statistic and Kappa SD were 67.6% and 0.14, respectively. On the testing data, the approach had an efficiency of 92.7% and a kappa of 88.59%. Especially in comparison to the CART algorithm, the method was misidentified as severe in only three occurrences.

Table 3: provision vector engine presentation in terms of truth and kappa

Sigma	c	Accuracy	Kappa	AccuracySD	KappaSD
0.050387	0.27	0.783225	0.678538	0.095596	0.140316
0.050387	0.52	0.776009	0.661357	0.087869	0.132556
0.050387	1.03	0.791394	0.678698	0.080343	0.126468

Table 4: lined discriminant procedure presentation

Accuracy	Kappa	AccuracySD	KappaSD
0.8042396	0.7038135	0.1016817	0.159308

As presented in table 5, the “LDA” achieved a truth equal to 80.2% on the exercise dataset. The “Kappa statistic” was 70.4%, and the “Kappa SD” was 0.15. The process's exactness equal on the test dataset was 90.1%, and its kappa value was 84.67%. Only 4 instances were misclassified as moderate by the system, and their performance was subpar compared to the CART method.

Table 5: “K-NN” procedure presentation

K	Accuracy	Kappa	AccuracySD	KappaSD
5	0.5988648	0.3698934	0.1280372	0.2158110
7	0.6268868	0.4072927	0.1564921	0.2703506
9	0.6621974	0.4715554	0.1747905	0.2881392

This diagram random forest's exactitude on the exercise established was 85.57%, as presented in table 6.

Table 6: casual forest presentation

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.8272529	0.7421422	0.10396456	0.15420082
14	0.8554216	0.7829893	0.06069715	0.09303132
16	0.8482786	0.7718937	0.06455252	0.09881994

5. Discussion

“Machine Learning”, and “Big Data Analytics” models for Cloud Computing platforms are a step ahead to evolving our lives. It brings the setting of Artificial Intelligence (AI) into the public eye. AI is not just a buzzword; it has been taking place for many decades for both people and machines. They help in processing huge amounts of data on Cloud platforms without causing any bottlenecking, which can be carried out with the use of machine learning and analytics. This leads to the agreement of cloud platforms on selling machine learning and big data analytics models as a service [13].

Cloud companies look at this service as a business model, where there is no pressure on sales teams. In this same light, they provide the programs to execute it in their data centers through “Machine Learning” and “Big Data Analytics” models. With the assistance of these two services, companies can make sure that there is no exposure to security issues or other issues and therefore are not prone to public backlash on privacy issues. They can be led to the conclusion that they have installed a machine that is a step ahead of our technology [14].

It brings the setting of “Artificial Intelligence” (AI) into the public eye. AI is not just a buzzword; it has been taking place for many decades for both people and machines. They help in processing huge amounts of data on Cloud platforms without causing any bottlenecking, which can be carried out with the use of machine learning and analytics [15]. With the use of cloud computing, most people are more than willing to purchase a machine with more storage than what can be stored at home. This leads to the agreement of cloud platforms on selling machine learning and big data analytics models as a service [16].

6. Conclusion

Machine learning techniques are a category of artificial intelligence. They can be controlled, unmonitored, semi-supervised, or encouragement supervised learning. Trends and connections in large amounts of data are instantly looked for, which gives rise to the formation of modeling techniques for those trends. In addition to the dimension of the data, big data analytics looks at how much information there is, how different it is, and also how fast it changes. Big data is defined by its quantity, speed, wide range, factuality, honesty, vocabulary, compliance to distinct framework, modeling techniques, and conceptual frameworks, and valuation, which explains the value and cost of big data. The book talks about just how big data is big as well as changes quickly. Wide range is a word that shows how different big data is. Because of big data, tools and methods for big data mining and large data analytics have been made. The word "big data analytics" makes reference to a mix of very well methods and techniques, such as machine learning and data mining, that can be used to use vital information that is usually covered in big data and create an interface in the form of straightforward and data visualisation.

Big Data Analytics analyses a range of unorganized and semi-structured data, including online posts, call logs, virtual server logs, and web click streaming data. Data gathering, computer vision, machine intelligence, analytics, and processing of natural language are all utilized in big data analytics.

Computer-based data, such as geographic IP location, computer security health certifications, keyboard typing, clickstream patterns, and WAP data are examples of passive data sources. The use of firewalls, encryption, secure protocols, antivirus software, and other measures can help protect data transmitted via networks. However, hackers can always come up with creative ways to access network infrastructure.

References

[1] Zeadally, S., Adi, E., Baig, Z., & Khan, I. (2020). Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/access.2020.2968045

[2] Kong, L. J. (2013). An improved information-security risk assessment algorithm for a hybrid

model. *International Journal Of Advancements In Computing Technology*, *5*(2).

[3] Liu, X., Lu, R., Ma, J., Chen, L., & Qin, B. (2016). Privacy-Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification. *IEEE Journal of Biomedical and Health Informatics*, *20*(2), 655–668. https://doi.org/10.1109/jbhi.2015.2407157

[4] Shah, S. K., Tariq, Z., Lee, J., & Lee, Y. (2021). Event-Driven Deep Learning for Edge Intelligence (EDL-EI). *Sensors*, *21*(18), 6023. https://doi.org/10.3390/s21186023

[5] Gupta, C., Johri, I., Srinivasan, K., Hu, Y.-C., Qaisar, S. M., & Huang, K.-Y. (2022). A Systematic Review on Machine Learning and Deep Learning Models for Electronic Information Security in Mobile Networks. *Sensors*, *22*(5), 2017. https://doi.org/10.3390/s22052017

[6] BERMAN, D.S., Buczak, A.L., Chavis, J.S., and Corbett, C.L. (2019). "Survey of Deep Learning Methods for Cyber Security", Information 2019, 10, 122; doi:10.3390/info10040122

[7] SARKER, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big Data. https://doi.org/10.1186/s40537-020-00318-5

[8] BRINGAS, P.B., and Santos, I., (2010). Bayesian Networks for Network Intrusion Detection, Bayesian Network, Ahmed Rebai (Ed.), ISBN: 978-953-307- 124-4, InTech, Available from: http://www. intechopen.com/books/bayesian-network/bayesiannetworks-for-network-intrusion-detection

[9] BLOICE, M. & Holzinger, A., 2018. A Tutorial on Machine Learning and Data Science Tools with Python. Graz, Austria: s.n

[10] 2. Wilson, B. M. R., Khazaei, B., & Hirsch, L. (2015, November). Enablers and barriers of cloud adoption among Small and Medium Enterprises in Tamil Nadu. In: 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) (pp. 140-145). IEEE.

[11] HASHEM, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. In Information Systems. https://doi. org/10.1016/j.is.2014.07.006.

[12] SITI Nurul Mahfuzah, M., Sazilah, S., & Norasiken, B. (2017). An Analysis of Gamification Elements in Online Learning to Enhance Learning Engagement. 6th International Conference on Computing & Informatics.

[13] MENZES, F.S.D., Liska, G.R., Cirillo, M.A. and Vivanco, M.J.F. (2016) Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression. Expert Systems with Applications, 69, 62-73. https://doi.org/10.1016/ j.eswa.2016.08.014

[14] HASSAN, H. (2017). Organisational factors affecting cloud computing adoption in small and medium enterprises (SMEs) in service sector. Procedia Computer Science. https://doi.org/10.1016/j.procs. 2017.11.126

[15] BOU-HARB, E., & Celeda, P. (2018). Survey of Attack Projection, Prediction, and Forecasting in Cyber Security. September. https://doi.org/10.1109/ COMST.2018.2871866