International Journal of Engineering & Technology, 7 (4.36) (2018) 1189-1193

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Classification of Big Data: Machine Learning Problems and Challenges in Network Intrusion Prediction

Yasser Mohammad Al-Sharo[1], Ghazi Shakah[2] ,Mutasem Sh.Alkhaswneh[3] , Bajes Zeyad Aljunaeidi[4], Malik Bader Alazzam[5]

[1,2,3,4,5] *Faculty of Information Technology, Ajloun National University, Ajloun, 26810, Jordan*

Abstract

Centre of attraction of paper is on the main complication on classification of Big Data on network encroachment on traffic. It also explains the disputes this system faces that is bestowed by the Big Data difficulties that are correlate with the network interruption forecast. Forecasting of an attainable interruption in a network entails a prolonged accumulation of traffic information or data and being able to get the concept on their features on motion. The constant accumulation in the network of traffic data thereafter ends with Big Data difficulties that as a result of the large amount, change and possessions of Big Data. In order to learn the features of a network, one needs to have the skills in the machine techniques that are always able to capture world skills and knowledge of the traffic to be in order. The properties of Big Data will always end to an important system disputes to be able to apply machine learning foundation. The paper also discusses the disputes and problems in the way of taking care of Big Data categorization representing geometric techniques of learning along with the existing technologies of Big networking. The study particularly explains challenges that have a relationship with the combined directed by the techniques one learns, machine long learning techniques, and representation-learning techniques and technologies that are related to Big Data for example Hive, Hadoop and Cloud that are basics that enhances problem-solving that gives relevant solutions to classification problems in traffic networking.

Keywords *Big Data, machine learning, Hadoop distributed file systems, encroachment discovery.*

1. Introduction

Currently, Big Data is being described using its three basic features that are variety, velocity, and its volume.[1] The description means that some of the points when the velocity, capacity, and difference of the data and information has an increment,[2] the technologies, and its current technologies may sometimes not be able to grip storing of data and its transformations processes of information at this stage thus explained as large data.

The study that has been done on Data in terms of Big Data is usually explained or described as a method of evaluating and having the concept on the properties of big size datasets by drawing out important geometric and statistical arrangements[3]. Perfectly the three mentioned characteristics of a dataset also create an increment in the complexity of the data and therefore making sure the contemporary methods or techniques and sciences stop working as it is expected within a given processing period. For instance, net work trafficking, risk analysis, geospatial classification, and business forecasting are some of the applications that experience Big Data difficulties. Time-sensitive applications are applications that require a highly persuasive Big Data technologies and techniques in order for them to equip with the difficulties on the fly[4]. Some examples of this applications include Network intrusion detection and prediction systems.

Network intrusion detection and prediction are examples of such sensitive applications. Cloud technology, Hadoop Distributed File Systems[5], and Hive database techniques can be grouped together in order to the difficulties like the Classification of Big data. Nevertheless, the software which need a constant development in the rule of Big Data that includes a system of intrusion forecast and geospatial have high chances of suffering from Big Data difficulties automatically[6].

The difficulties and disputes in this paper that are correlated with embedding of up-to-date technologies of networking and techniques of machines learning that are used in giving solutions to Big Data categorization difficulties for network intrusion prediction are also considered[7].

Firstly, disputing issues rests on the new description of Big Data; that shows the data traffic network content the attributes of Big data for the categorization of Big Data[8]. Ideally, it is the earliest in order significant task to deal with it so that in order to make the analytics of Big data efficient and effective with the price. The first discovery of the features or attributes of Big Data can help in supporting price-effectiveness strategies - to many businesses to prevent the unwanted distribution of technologies of Big Data[3]. The data in the science of logical analysis of some data and information may not need Big Data techniques and technologies. The new or strong entrenched technologies may, therefore, be enough to take care of activities such as storage of data and processing of data. Therefore an earlier evaluation plus comprehension is required in data characteristics classification[9][10].

1.1. The Proof of Big Data

The new examination or study in the Big Data discipline has disregarded the first discovery of the Big Data attributes. An example for in study is the existing description of the Big Data described space of on 3D, V3 that is made by the three variables that are volume capacity, velocity, and also variety that cannot support an appropriate podium for the first discovery of attributes of Big Data for its stratification[11].

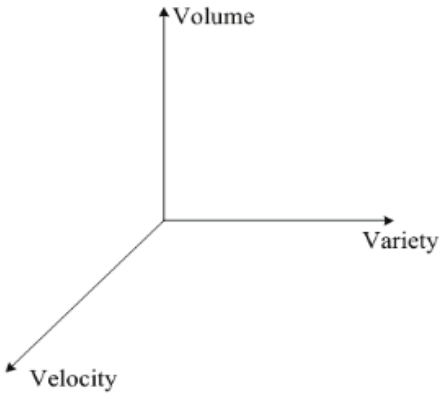

Fig.1: Definition of V3 Characteristics of Big Data

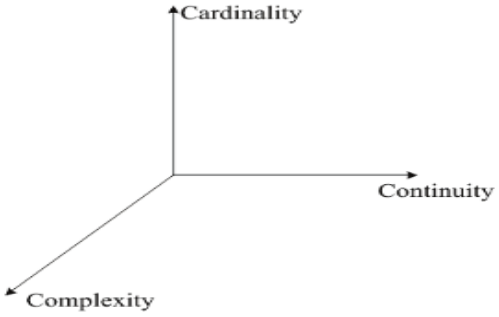

Fig.2: Definition C3 of characteristics of Big Data

1.2. The Definition of Big Data Features

Volume in Big Data, it describes or development of data volume, the arbor of velocity stands for addition in acceleration such that data should be prepared, arbour of the various aspects shows development in the extra data types. If we assume that the dataset has some numbers of zero, a portion of m of this numbers, n numbers of twos and so on such that there is an infinity growth that does not end. Therefore, the V3 space will require it as Big Data for instance by the fact that it is small data[12].

Fig. 1: Big data example

Examining in contrast to describing a measure of the metric in the attributes of metric to measure of attributes of Big data in the V3 space, this is recorded to make it easier to establish C3 metric space by employing mathematical tools and some important tools in statistics[13].

The unending describes to attributes that are named below: prolonged enhancement in the size of data with respect to the needs plus representing data by use of progressive functions. Also, the complicatedness describes three attributes and include: big differences in data type , high spatial dataset; along with the acceleration of data which is being processed should be high.[14]

1.3. Big Data Management.

The cardinality limit usually counts on the need for an efficient appropriated file system that will enable the capturing of data, storing and evaluating the network traffic for intrusion prediction. Therefore complicatedness and unending parameters will always add more problems to the task of being in charge or taking control over the Big Data. Hence the network geography must be constructed in a manner that the Bid Data Analytics difficulties can be taken care of effectively with the goal of cost-effectiveness[15].

1.4. Network Topology

The up-to-date modern computer technologies, such as public cloud and HDFS, can play a role in relieving the cardinality difficulties in Big Data Analytics[16]. They can also **be** joined in order to build a big and adaptive network topology that contains a store that is be able to make adjustments depending on processing needs of big data. In some situations, this joined model can initiate several disputes that should be taking care of effectively. This type of a joined model is made up of four units that are: NTRS, User Interaction and Learning System (UILS) AND Cloud Computing Storage System (CCSS). The Network Traffic Recording System (NTRS) unit in the joined model enables capturing of network traffic and flows it over to the traffic data to the CCSS unit in order for the unit to give additional data and information[17]. Most of the Cloud Computing Storage System is capable of using the Hive database in storing its data. The User Interaction and Learning System (UILS) is the central unit that takes control of everything including storage of data and data requirements[18].

Fig.2: Embracing Big Data

2. Communication Challenges

Research in networking of computers and its uses, communication cost is set as the mains responsibility examining in contrast with the processing activity cost of the same data in the mentioned topology. The main concern dispute in this is how to minimize the rate of communication cost while at the same time be able to meet the needs of the more data requirements from the public cloud for Big Data processing[19]. There are two major network facial characteristics that will essentially has effect on communication with the clients and the cloud server. They include bandwidth or latency[8]. Communication interference

h right, as in (1). Your equation should be typed using the Times New Roman font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

between clients and cloud server is a large problem that has associated itself with disputes to search results that unfavourably interferes with needs of the Big Data processing at the Cloud Computing Storage System unit and also at the UILS. In order for the

machine learning techniques that are using the technologies to be improved, the disputes and difficulties should be kept in mind[20].

2.1. Security Challenges

In cloud technology, the security mechanisms are generally weak. Its weakness in the security system usually tampers the data in the public cloud that has a link to the clients, therefore, being invertible and also having a big concern. In order to find a strong security measure for the reason of using the public cloud, for example, Cloud Computing Storage System is challenging difficulty. Cloud technology having its poor security systems, hackers can easily tamper with the data that currently at the situation it's being transferred and exchanged between the Cloud Computing Storage System and the HDFS performance evaluation review and the Network Traffic Recording System (NTRS) units. The hacker also called attackers can trick the reply that between this unit and therefore being able to shut the cloud server; Cloud Computing Storage System using an attack of DOS.These difficulties transgress to disputes in initializing the tool of Big Data Analytics along wuith the network topology that is proposed[18].

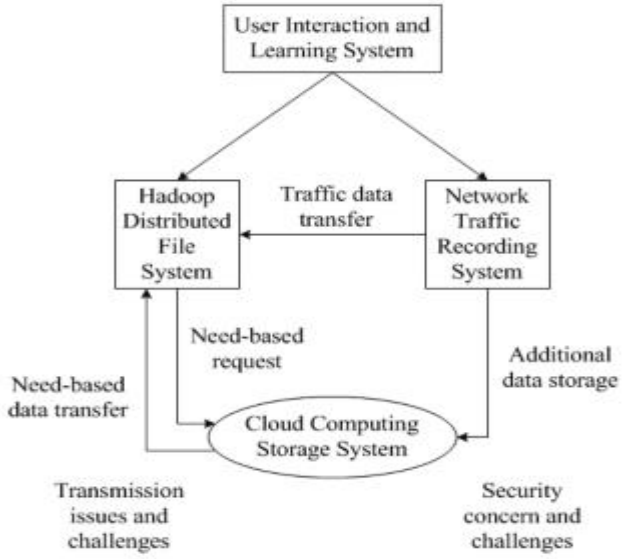

Fig.3: Proposed network topology for Big Data analytics

2.2. Knowledge of Big Data.

These hurdle parameters entail other numerous authoritative attributes of information or data that are of a higher spatial dataset, a bigger figure on types of data, required acceleration in which the data should take a transformation process and also data that is unstructured. The complicatedness limits along with its outcome difficulties should be forwarded by use of techniques of machine learning.This dispute resets in order to keep a continuous improvement in the on-going learning techniques to be able to handle the big information categorization difficulties.[21]

Before, there existed a machine that was called the traditional learning Machine method that has been improved plus utilised for getting information that is of help from the information by practising and also validating by using the datasets that are named.

Learning Machine being a suitable technique, there are some problems that still find ways of having it done and therefore not able to solve or give solutions to Big Data classification problems. These problems include (i) An Learning machine that has successfully qualified on a specific dataset, or the rule of data maybe unable to be fit to suit for a different set of data or information which the catagerization cannot be able to strengthen over the various sets of data. (ii) A Learning Machine technique that is generally prepaid using certain types of classes, having a huge variety of types of class that are located in the dataset that is growing ,therefore, creating inaccurateness in results classification. (iii) An ML technique is through the growth that bases on single simple learning tasks and activities. For this reason, they are not suitable for learning since it cannot be used in learning skills along with the requirements of knowledge of analysing Big Data. Its only supervised algorithms that help in the classification of the prediction of network traffic data intrusion. Numerous algorithms have been created; they have have their importance in intrusion traffic among greater machines that SVM that usually have attention[22]. The computerized cost of Support Vector Machine is, therefore, higher than other classification methods. To leave this difficulty, Support Machines has afterward developing in machine learning research. For this reason, the support vector machine is not capable and suitable in Big Data Analytics. Support Vector Machine is sometimes accurate and therefore they are preferred to be excellent. Since it has the ability to adopt Support Vector Machine then they are highly preferred[4]. The challenge here comes when finding the solution to make an improvement to the Support Vector technique.

3. Health and Medical Information

An individual can be capable of dealing with larger volumes of data that is structured and unstructured which comes from various sources. Big data analytical tools have the capacity if keeping the promise of studying the results on the population of a large-scale which is based on studies that are longitudinal and also capturing the tendencies and models that are proposed that are predictive that get the data which is obtained from electronic health records of various patients. There is a rare chance which lies at the centre of two medical informatics; the traditional and mobile health which assists in offering information on acute diseases and chronic illnesses in a manner that has not been observed before[23].

An electronic health record (EHR) assists in elaborating the treatment of patients and also in offering the final results that are well informed. Majority of the traditional health data centres normally capture and store data in large amounts of the data in a structured manner and this is with respect to a lot of data such as laboratory tests for all patients, clinical data and the medication of the patients. Doctors use the health records of the patients[24], natural language processing plays a critical role in the organized examination and suggestion of grammatical content that are underlying the in the data. Mining[16] EHRs is an essential tool in the creation of clinical data and keeping up with clinical study such as identifying information of phenotypes[19]. Mining of local information which is part of HER data has already assured of efficient management of a wide range of healthcare issues such as support of diseases and management, creation of modes to anticipate the evaluation of health risks, enhancing education on the rates of survival and healing recommendations.[25] It has also been used in finding our comorbidies and also developing support structures for enrolling patients for new trials and new clinical tests. A lot of work in health facilities is always cantered on examining large and complex data from the patients that was collected in the past[26]. Nonetheless, majority of the clinical databases offer low temporary information since there is struggle in gathering quality long-term data. In order to solve this issue, current clinical databases should be enhanced by linking mobile health platforms, community centres and other health facilities so that other data can be shared through the system in facilitating decision making and solve clinical queries that have not been solved[27]. The intriguing part takes the direction that will develop models that are patient-specific by use of the available records in books and databases. Subsequently, the model can be updated with data which can be collected outside the health facilities from patients that cannot get to the hospital. Some illnesses that are incessant show acute happenings which cannot be predicted easily fully in hospitals[26]. People that have connective tissue disorders are mostly predis-

posed to aortic aneurysms or tear. Screening many people for this illness will be helpful in identifying those that have the likelihood of developing aortic dissection.

Even if an original model created from explanation can offer better insight into this issue, this does explain the continuous hemodynamic differential over time along with its impact of daily activities of people. Through integrating ambulatory BP models, the possibility of creating results of simulation lead to a longitudinal model that spans over a duration of time to understand the progress of the illness well[25]. One of the primary roles of telemedicine is health as it aims to connect patients with doctors past the clinic to enable the doctors check the whereabouts of the patients. The processes of communication have been enhanced due to social networks, to new levels of interaction in the society plus making peace[6]. The newly developed feature created new abilities and also opened new abilities of communication among patients. A quarter of the patients suffering from chronic illnesses like cancer, diabetes and heart illness are now making use of social media to share their experiences with other patients that are going through the same issues, this therefore offers another source of important Big Data[28]. Moreover, relocation and social applications offer an extra feature to comprehend the tendencies plus the patient's social demographics, while dodging studies that use a lot of resources and costly studies in large sampling statistically. This benefit has already been used by various studies of epidemiological studies fields such as outbreaks of influenza, smoking dynamics and in the misuse of antibiotics. The posts and texts that are shared on the online platforms such as the social media are very essential since they are important sources of information. In comparison to traditional techniques like fluctuations, emotion regulation, surveys[10], behaviours and thoughts are analysed through the social media platforms provide new chances for analysis that is real time or the mood that is expressed as well as its context. Internet searches along with social media can also be joined with data from the environment like information on the air quality so as to foretell the immediate rise of asthma linked emergency visits[29]

4. Representation Learning.

Learning algorithms help in representation techniques enable one attain an elevated level of classification that is accurate in computational effectiveness. Therefore, data transformation as the original features are maintained is possible[7]. Another rule used, is in bringing out the accuracy of these algorithms and minimize complexity of computation which also enhances acceleration of the process activity. Big Data classification sometimes requires multi-rule and other methods and of its large domain.

5. Big Data with User Interaction

Big Data has many challenges in its classification[26][30]. Another challenge is the utilization of the time of data and processing procedures. Also being able to detect and interaction that is amongst the parameters of Big Data[31] , that is complexity, continuity is observed to be a challenge that requires user interactions[22]. When the usage of the machine is long term learning, then there be should a user interaction which will help the Big Data classification importantly.

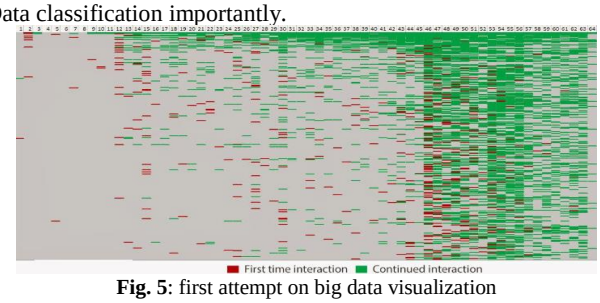

Fig. 5: first attempt on big data visualization

6. Data Visualization

Data visualization is being kept as a challenging task by the attributes of the Big Data. Dimension reduction and data projection are among the recent techniques that just offer an intangible picture of the data. In many cases, the intangible picture of the data normally doesn't usually produce the spatial illustrations of the information or the data. Consequently, this is an issue to this data visualization. In some model, algorithms provide a unit circle that represents intrinsic traffic and also regular traffic that can thus cut down the issue of visualization of Big Data, hence, plotting the big numbers of points of data to unit circle in order to find the solution[32]. If this problem is solved, it can be beneficial to the User Interaction Learning System (UILS). This will enable proper storage of data[33]

7. Data Ambiguity

Exchange of information has a problem that is brought about when the activities are transferred amongst UILS, CCSS and also NTRS units in the networking having greater delay in the data or maybe it can bring the loss of data[2][23]. The missing-data problem will cause data uncertainty in the system that will occur in the UILS unit. It therefore grows and continuous bring up complexity to the Big Data definition. Different techniques should be presented earlier so that it can be considered in the growth and creation of precise information from the data that is incomplete. For this reason, the units of UILS should make a step in the way it will handle this situation with a user and also give it significant disputes to user[34].

8. Conclusion

The paper desires that an integration of the modern technologies that is the Cloud Technologies, Hadoop Distributed File Systems that consist of representations, learning methods and also help in supporting vector machines to be able to see the future network interruption in Big strategy of Data. It is also suggestive that adopting machine lifelong learning framework for getting solutions to issues that the classification of Big Data for a network system. Intrusions that in reality, they do not have any real experience in Big analytics of data. It further suggests different description of the V3 of Big Data ranging up to C3 to enable the professionals explain and giving the understanding of the both techniques[15]. The studies on the techniques of Big Data plus other technologies that come out[35], handles the disputes that emerge, therefore there is still hope that there is a development of better techniques toward the ways of getting into solutions towards the Big Data classification problems.

9. Funding

This research is funded by the Deanship of Research and Graduate Studies in Ajloun National University Ajloun, 26810, Jordan

References

[1] R. Kitchin, "The real-time city? Big data and smart urbanism," *GeoJournal*, vol. 79, no. 1, pp. 1–14, 2014

[2] A. B. Wei Fan, "Mining big data: current status, and forecast to the future," *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, pp. 1–5, 2012.

[3] H. V Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and Its Technical Challenges," *Assoc. Comput. Mach. Commun. ACM*, vol. 57, no. 7, p. 86, 2014.

[4] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[5] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[7] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[8] H. Chen and V. C. Storey, "B Usiness I Ntelligence and a Nalytics : F Rom B Ig D Ata To B Ig I Mpact," *Mis Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[9] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2561–2573, 2014.

[10] C. Snijders, U. Matzat, and U. Reips, "'Big Data': Big Gaps of Knowledge in the Field of Internet Science," *Int. J. Internet Sci.*, vol. 7, no. 1, pp. 1–5, 2012.

[11] H. R. Varian, "Big Data: New Tricks for Econometrics," *J. Econ. Perspect.*, vol. 28, no. 2, pp. 3–28, 2014.

[12] K. U. Jaseena and J. M. David, "B Ig D Ata M Ining," no. August, pp. 131–140, 2014.

[13] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Commun. ACM*, vol. 57, no. 3, pp. 78–85, 2014.

[14] E. Junqué de Fortuny, D. Martens, and F. Provost, "Predictive Modeling With Big Data: *Is Bigger Really Better* ?," *Big Data*, vol. 1, no. 4, pp. 215–226, 2013.

[15] O. Tene and J. Polonetsky, *Big data for all: Privacy and user control in the age of analytics*, vol. 11, no. 5. 2013.

[16] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[17] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii Int. Conf. Syst. Sci.*, pp. 995–1004, 2013.

[18] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!," *... 14th Int. Work. Data ...*, pp. 101–104, 2011.

[19] J. Lee, H. A. Kao, and S. Yang, "Service innovation and smart analytics for Industry 4.0 and big data environment," *Procedia CIRP*, vol. 16, pp. 3–8, 2014.

[20] M. A. Just, L. Pan, V. L. Cherkassky, D. Mcmakin, C. Cha, M. K. Nock, and D. Brent, "Emotion Concepts Identifies Suicidal Youth," pp. 911–919, 2018.

[21] D. Kononenko, Y. Ganin, D. Sungatullina, and V. Lempitsky, "Photorealistic Monocular Gaze Redirection Using Machine Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–15, 2017.

[22] Y. M. Al-sharo, "Comparative Study of Neural Network Based Speech Recognition : Wavelet Transformation vs . Principal Component Analysis," vol. 5, no. 1, pp. 1–5, 2015.

[23] M. Doheir, B. Hussin, A. Samad, H. Basari, and M. B. Alazzam, "Structural Design of Secure Transmission Module for Protecting Patient Data in Cloud-Based Healthcare Environment," *Middle-East J. Sci. Res.*, vol. 23, no. 12, pp. 2961–2967, 2015.

[24] M. B. Alazzam, A. Samad, H. Basari, and A. S. Sibghatullah, "Trust in stored data in EHRs acceptance of medical staff : using UTAUT2," vol. 11, no. 4, pp. 2737–2748, 2016.

[25] A. Mamra, A. S. Sibghatullah, G. P. Ananta, M. Bader, Y. H. Ahmed, M. Doheir, A. Mamra, A. S. Sibghatullah, G. P. Ananta, B. Alazzamm, Y. H. Ahmed, and M. Doheir, "Theories and factors applied in investigating the user acceptance towards personal health records : Review study Theories and factors applied in investigating the user acceptance towards personal health records : Review study," *Int. J. Healthc. Manag.*, vol. 0, no. 0, pp. 1–8, 2017.

[26] J. Forrester-sellers, "Classifying Ancient West Mexican Ceramic Figures Using Three-Dimensional Modelling and Machine Learning," pp. 19–24, 2017.

[27] S. M.Alazzam, BASARI, "EHRs Acceptance in Jordan Hospitals By UTAUT2 Model: Preliminary Result," *J. Theor. Appl. Inf. Technol.*, vol. 3178, no. 3, pp. 473–482, 2015.

[28] M. Rasmi, M. B. Alazzam, M. K. Alsmadi, A. Ibrahim, R. A. Alkhasawneh, and S. Alsmadi, "Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study) Healthcare professionals ' acceptance Electronic Health Records system : Critical literature review (Jordan case study)," *Int. J. Healthc. Manag.*, vol. 0, no. 0, pp. 1–13, 2018.

[29] M. R. Ramli, Z. A. Abas, M. I. Desa, Z. Z. Abidin, and M. B. Alazzam, "Enhanced convergence of Bat Algorithm based on dimensional and inertia weight factor," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018.

[30] S. Nikou and H. Bouwman, "The Diffusion of Mobile Social Network Service in China: The Role of Habit and Social Influence," *2013 46th Hawaii Int. Conf. Syst. Sci.*, pp. 1073–1081, Jan. 2013.

[31] M. B. Alazzam, Y. M. Al-sharo, and M. K. Al-, "DEVELOPING (UTAUT 2) MODEL OF ADOPTION MOBILE HEALTH APPLICATION IN JORDAN E- GOVERNMENT," vol. 96, no. 12, 2018.

[32] C. Hair, Joseph F, Anderson, Rolph E, Tatham, Ronald L, & William, *Multivariate data analysis*. 1998.

[33] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," *Proc. - IEEE Symp. Secur. Priv.*, pp. 3–18, 2017.

[34] A. Mamra and A. Mamra, "A Proposed Framework to Investigate the User Acceptance of Personal Health Records in A Proposed Framework to Investigate the User Acceptance of Personal Health Records in Malaysia using UTAUT2 and PMT," *Int. J. Adv. Comput. Sci. Appl.*, no. March, 2017.

[35] D. B. Fridsma, "Moving beyond the physician's EHR," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 6, pp. 1277–1277, 2015.