Big Data in Healthcare: Are we getting useful insights from this avalanche of data?

Kayode I. Adenuga
School of ICT, ICT University
Messasi, Yaoundé
Cameroon
Kayodeadenuga@yahoo.ca

Idris O. Muniru
Department of Biomedical Engineering
University of Ilorin, Kwara State, Nigeria
idrismunir@gmail.com

Fatai I. Sadiq
Computer Science Dept. Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
sfatai2011@gmail.com

Rahmat O. Adenuga
University Hospital Southampton, Hampshire, United Kingdom
rahmatnuga@gmail.com

Muhammad J. Solihudeen
Information System Dept.
Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
Jamiumuhammed@yahoo.com

ABSTRACT
The benefits of deriving useful insights from avalanche of data available everywhere cannot be overemphasized. Big Data analytics can revolutionize the healthcare industry. It can also ensure functional productivity, help forecast and suggest feedbacks to disease outbreaks, enhance clinical practice, and optimize healthcare expenditure which cuts across all stakeholders in healthcare sectors. Notwithstanding these immense capabilities available in the general application of big data; studies on derivation of useful insights from healthcare data that can enhance medical practice have received little academic attention. Therefore, this study highlighted the possibility of making very insightful healthcare outcomes with big data through a simple classification problem which classifies the tendency of individuals towards specific drugs based on personality measures. Our model though trained with less than 2000 samples and with a simple neural network architecture achieved mean accuracies of 76.87% (sd=0.0097) and 75.86% (sd=0.0123) for the 0.15 and 0.05 validation sets respectively. The relatively acceptable performance recorded by our model despite the small dataset could largely be attributed to number of attributes in our dataset. It is essential to uncover some of the many complexities in our societies in relations to healthcare; and through many machine learning architectures like the neural networks these complex relationships can be discovered

CCS Concepts
• Information systems→Big Data Analytics • Computing methodologies→Soft Computing.

Keywords
Big Data, Analytics, Benefits, Challenges.

ICSIE 2019, April 9-12, 2019, Cairo, Egypt

ACM ISBN 978-1-4503-6105-7/19/04…$15.00
https://doi.org/10.1145/3328833.3328841

1. INTRODUCTION
All around the world, improving health care delivery is hinged on research and evidence based practices, and these have evolved on the premise of years of research on how patients care were rendered resulting in positive outcome. It is a known fact that accurate, timely, accessible data and records is a major tool used in improving these outcome. The benefits of deriving useful insights from avalanche of data available everywhere cannot be overemphasized. Big Data analytics can revolutionize the healthcare industry. It can also ensure functional productivity, help forecast and suggest feedbacks to disease outbreaks, enhance clinical practice (as shown in Figure 1.) and optimize healthcare expenditure which cuts across all stakeholders in healthcare sectors [1]. Big Data has the tendency to minimize expenses and offer great advancement in the time required to perform a spreadsheet exercise [2, 3].In the recent times, advanced health information system are applied to move from the traditional record keeping method to automated electronic means for more efficient storage and maintenance of patients' records. Examples of these include electronic medical record (EMR) and personal healthcare record systems (PHR), [4]. Despite immense capabilities available in the general application of big data, studies on derivation of useful insights that can enhance medical practice has received little academic attention. Therefore, this study explores how useful data can be derived from a data set, analyze the outcome and proffer suggestion that can guide future research directions.

Figure 1. Big Data Analytics in healthcare (Source: www.123rf.com)

2. PROBLEM BACKGROUND

One of the challenges of healthcare management in developing countries is lack of data to obtain new insights that can enhance its practice. By having access to this data; researchers for instance can mine the data to identify what treatment procedures are most suitable for particular conditions, determine sequences associated with drug reactions and identify drugs that can provide cure for certain diseases [5]. This in turn helps to obtain other useful information that can promote healthcare delivery and optimize healthcare costs in general [4]. Presently, companies and humans produce enormous quantities of data at an increasing rate nearing the dimension of an exabyte. This development has brought about the need to use some tools to mine useful information from these data in order to derive new insights and establish a pathway for business growth and development. In other words; this will help corporate organizations to gain deeper knowledge of their operations and improve on their day to day activities. Nevertheless, it is a general belief that related information acquired using big data technologies will enhance business productivity and bring about a competitive edge [6]. In developed countries, Big Data application has recorded a huge success. It has been applied in astronomy, online sales, search engines and politics [7]. Furthermore, healthcare decision makers and stakeholders can now have access to this favourable drifts in knowledge. Big data; though comes as an information not only for its magnitude but due to its complexity, heterogeneity and readiness. Healthcare providers and pharmaceutical industries can now apply outcome of analysis obtained from big data to provide recommendations that can help their respective organizations'. Although these efforts are still in their nascent stages, they could collectively help the industry solve problems related to differences in healthcare quality and escalating healthcare expenditure. Learning from the aforementioned, healthcare sectors of developing countries can also harness these benefits to improve the health sectors most especially towards predicting incidence of a particular disease and offer good planning that can predict and forestall any eventual disease outbreak. Notwithstanding these immense capabilities available in the general application of big data, studies on derivation of useful insights from healthcare data that can enhance medical practice have received little academic attention. Therefore, this study demonstrates how useful data can be derived from a large volume of data, analyze the outcome and proffer suggestion that can guide future research directions in developing countries.

3. LITERATURE REVIEW

3.1 What is Big Data?

Big data are sizeable composite datasets that extend beyond the capacities of conventional data management systems which is applied to organize, manipulate and store data in an expedient and cost-effective manner. They are classified as structured, semi-structured or unstructured and often in petabytes. It is an innovative phenomenon and still at a nascent stage of adoption for many sectors but it is obvious that exploiting its full potentials can provide a number of overwhelming advantages [1].

3.2 Characteristics of Big Data

Big data is referred to as the development of data that are hard to store, organize, and examine using a normal database system. It is commonly classified as the Four Vs and these Vs. are presented below as explained by Feldman, Martin [8] and Patil, Raul [4] and as shown in Figure 2.

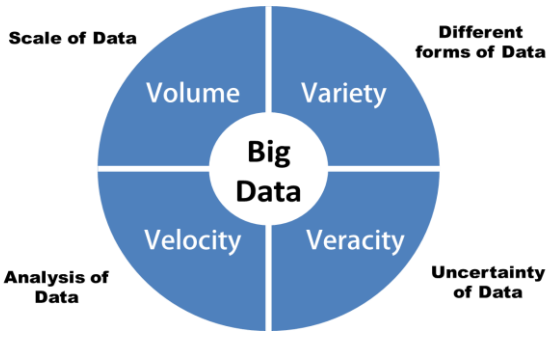

Figure 2. Big Data Four V's (Source: www.nextview.nl)

Volume: The magnitude of world's data is rising quickly, *in* 2005 data was estimated to be 130 exabytes, this figure has ri*sen to 7,910 exabytes in 2015* and by 2020, it is expected to have increased to 35 zettabytes (10^{21} *bytes)*. From these numbers, only 20% of this data is structured *i.e.* data suitable for computer processing while unstructured is growing at about 15 times more th*an the rate of the structured data*. This development emerges from digitizing existing healthcare data such as electronic medical records, radiology images and current huge byte of data emanating from 3D medical imaging and biometric sensing data also contributes to this size. Therefore it poses a big task to store such a complex data. Manipulating data of this nature requires scalable storage and also support for these distributed queries across the data sources. It was recommended however that health care systems should be able to examine and recognize this data in this enormous data arrangement.

Variety: Data stored in healthcare systems can be structured, semi structured and unstructured. This huge variety of data makes healthcare data both compelling and demanding. Literally, the point of care in medical settings generated mostly unstructured data which includes paper prescriptions, office medical records, handwritten notes by clinicians (Doctors and nurses), x-rays, hospital admission and discharge records, MRI scan and other images. It is unlike structured data which can be easily stored, retrieve, examine and manipulated by computer. Studies have shown that relatively little of this data can currently be assessed, preserved and organized in a form that can be manipulated by computers and examined for useful information. It was therefore recommended that healthcare applications require more effective techniques to integrate and translate varieties of data, including instinctive conversion from structured to unstructured data. Although, recent technological advancement has brought about various technologies to handle this highly variable data.

Velocity: The steady flow of new data gathering at unparalleled dimensions present new concerns. The recent changes in the volume and variety of data collected and stored, has brought about a corresponding velocity at which the data is generated and the speed needed to retrieve, examine and make decisions using the outcome. Conventionally, most healthcare data come in form of paper files, X-ray files and so on. But in some medical situations, real-time data such as operating room monitors for anesthesia, bedside heart monitors become inevitable. The application of this real time data coming from ICU can help detect related diseases early enough and apply appropriate treatments. This in turn reduces patient morbidity and mortality in the public and private hospitals at large. It is important that data are explored in such a manner that the main cause of a particular ailment is also examined

using predictive analysis technique. It is therefore mentioned that ability to carry out real-time analytics against high-volume data in motion could bring about rapid change in healthcare industry. Nevertheless, data stored in healthcare systems is updated on daily, weekly and even monthly basis so it is very necessary that data stored is accurate and free from any error. Consequently, it is also important to leverage data to manage and lower the healthcare expenses.

Veracity: Veracity implies data of varying quality, relevance and meaning. Data quality issues pose a great challenge in healthcare for two obvious reasons: (1.) in critical situations, healthcare resolutions are determined by availability of accurate information. (2.) Challenge of unstructured nature of healthcare data and which are mostly unreliable. Veracity issues applicable to healthcare can be in form of: "*are diagnoses, treatments, prescriptions, procedures, outcomes correctly captured"?* Having access to high quality data can improve health care management, improve diagnosis accuracy and reduce healthcare costs. However, high Variety and Velocity prevents the need to treat data before investigating and making some inferences, bringing up issues of data "trust." There are unrevealed statistical issues afflicting large data sets, not just "garbage in, garbage out". Therefore, in order to provide effective and correct solution, the data to be analyzed must be accurate in all aspects. Consequently if the data has good quality, the outcomes will be useful towards making some productive conclusions that can aid healthcare management and consequently reduce healthcare cost.

3.3 Related Work on Big Data

Mathew and Pillai [9] study focuses on big data issues, approaches and solutions in relation to health care industry. Various big data solutions and framework for managing divergent clinical data were discussed. It was mentioned that proper implementation could enhance healthcare decision making and bring about productive outcomes. In addition, their work highlighted review of applications of big data in healthcare, factors affecting its adoption, possibilities of big data and architecture to manage healthcare data. The findings revealed that proper selection of tools to do analytics on health care data can provide promising results. The varying adoption bottlenecks mentioned in their study are some of the issues considered as an area of research for future studies. Jiang, Winkley [10] presented a wearable sensors system with capability for continuous monitoring of the elderly and advancing the data to the big data systems. Their arguments were based on the fact that since big data system controls the volume, variety and velocity of information from various roots. In their study, Hidden Markov model for human behavior recognition was applied to a wearable sensor system with intelligent data forwarder. This ensures that meaningful data from sensor and conditions of user is fed to the system and consequently improves the quality of care. Salih and Abraham [11] examined and applied different algorithms to produce novel intelligent ensemble health care and decision system to control and enhance healthcare using wearable sensors. This Novel Intelligent Ensemble method was designed based on Meta classifier voting in combination with three base classifiers J48, Random Forest and Random Tree algorithms. Similarly, Vemuganti [12] mentioned the importance of Meta data management to support big data analytics and in carrying out their study, Meta data management framework and its importance were examined while Muni Kumar and Manjula [13] explained how Big Data Analytics are needed to revolutionize the rural healthcare by deriving invaluable insights from their clinical data and to efficiently make the right decisions. Therefore in view of their recommendations on the use of proper big data analytical tools; the study applied Keras deep learning neural network framework. Due to its flexibility of minimizing cognitive load. It has a consistent deep learning framework supported with simple application program interface (API). It is also known to reduce the frequency of user actions for regular use cases and also provides coherent and practicable response upon user error. This model can be trained, installed across varying platforms and has capability to turn models into products.

4. RESEARCH METHOD

In this work, we extended the work of [14] by bringing out more insights from the database which was collected by Elaine Fehrman using anonymous online survey methodology (Fehrman E, Egan V. Drug consumption, collected online from March 2011 to March 2012). This is expected to show the possibility of making very insightful healthcare outcomes with big data through a simple classification problem which classifies the tendency of individuals towards specific drugs based on personality measures and biodata. We used the Drug consumption (quantified) Data Set [14]. This dataset presents "The Five Factor Model of personality" as related to drug consumption risk. It contains 1885 samples each with twelve (31) fields. The fields comprise personality measurements (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness-NEO-FFI-R, impulsivity- BIS-11, and sensation seeking – ImpSS), level of education, age, gender, country of residence and ethnicity. The other fields in the dataset are participants' responses concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, volatile substance abuse and Semeron. According to the author of the dataset, *Semeron*, which is a fictitious drug, was introduced to identify over-claimers. We built a simple Volatile Substance Abuse classifier with the keras deep learning framework using the above dataset. Our model architecture contains 30 input nodes, one hidden layer with eight (8) nodes and 7 output nodes. We trained the model for a hundred (100) epochs, 5k cross validation and with five percent (5%) and fifteen percent (15%) validation set respectively.

5. ANALYSIS AND DISCUSSION

The performance parameters of our model after 100 epochs with 0.05 and 0.15 validation sets are shown in Table 1 and Table 2 respectively. Our model though trained with less than 2000 samples and with a simple neural network architecture achieved mean accuracies of 76.87% (sd=0.0097) and 75.86% (sd=0.0123). The small values of the standard deviation show that the model accuracies for each of the cross-validation episode are close. The validation accuracies indicate that the model generalizes better on the validation set as compared to the training sets (77.53% against 63.16%). However, this performance has the tendency to improve with more data samples.

A deeper look at the dataset shows that it consists three (3) different parts; biodata, personality measure, and drug use history. Although, these parts could have been gathered from different sources, the deep learning algorithm we applied on the dataset identifies the complex relationships amongst the thirty (30) entities to create the classifier model (Volatile Substance Abuse Predictor). The output of the classifier, if linked to disease risks will empower healthcare practitioners to better manage incidents and improve healthcare policies formulation, decision making and facilitate proper planning.

6. CONCLUSION

The outcome of the study revealed that using the right tools, many insights can be derived from large volume of data. Although, the relatively acceptable performance recorded by our model despite the small dataset could largely be attributed to number of attributes in our dataset. Multivariate data give possibilities for multiple inferences and it is essential to uncover some of the many complexities in our societies in relations to healthcare. And through many machine learning architectures like the neural networks these complex relationships can be discovered.

7. REFERENCES

[1] Nambiar, R., et al. *A look at challenges and opportunities of big data analytics in healthcare*. in *Big Data, 2013 IEEE International Conference on*. 2013. IEEE.

[2] Davenport, T., *Big data at work: dispelling the myths, uncovering the opportunities*. 2014: Harvard Business Review Press.

[3] Birjali, M., A. Beni-Hssane, and M. Erritali, *Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume.* Procedia computer science, 2017. **113**: p. 280-285.

[4] Patil, P., et al., *Big Data in Healthcare.* International Journal of Research in Information Technology, 2014. **2**(2): p. 202-208.

[5] Groves, P., et al., *The 'big data'revolution in healthcare.* McKinsey Quarterly, 2013. **2**(3).

[6] Akoka, J., I. Comyn-Wattiau, and N. Laoufi, *Research on Big Data – A systematic mapping study.* Computer Standards & Interfaces, 2017. **54**: p. 105-115.

[7] Murdoch, T.B. and A.S. Detsky, *The inevitable application of big data to health care.* Jama, 2013. **309**(13): p. 1351-1352.

[8] Feldman, B., E.M. Martin, and T. Skotnes, *Big data in healthcare hype and hope.* Dr. Bonnie, 2012. **360**: p. 122-125.

[9] Mathew, P.S. and A.S. Pillai. *Big Data solutions in Healthcare: Problems and perspectives*. in *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*. 2015. IEEE.

[10] Jiang, P., et al., *An intelligent information forwarder for healthcare big data systems with distributed wearable sensors.* IEEE systems journal, 2016. **10**(3): p. 1147-1159.

[11] Salih, A.S.M. and A. Abraham, *Novel Ensemble Decision Support and Health Care Monitoring System.* Journal of Network and Innovative Computing, 2014. **2**(2014): p. 041-051.

[12] Vemuganti, G., *Metadata Management in Big Data.* Big Data: Countering Tomorrow's Challenges, 2013. **3**.

[13] Muni Kumar, N. and R. Manjula, *Role of Big data analytics in rural health care-A step towards svasth bharath.* International Journal of Computer Science and Information Technologies, 2014. **5**(6): p. 7172-7178.

[14] Fehrman, E., et al., *The Five Factor Model of personality and evaluation of drug consumption risk*, in *Data Science*. 2017, Springer. p. 231-242.

Table 1. Performance Parameters of the Model

	Training Performance		Validation Performance	
Validation Set	Loss	Accuracy	Loss	Accuracy
0.05	0.6097	0.7884	0.9030	0.6316
0.15	0.6034	0.7861	0.6195	0.7753

Table 2. Validation Set of the Model

	Model Accuracy	
Validation Set	Mean	Standard Deviation
0.15	76.87%	0.97%
0.05	75.86%	1.23%