Modeling Method for Leveraging Data Quality in Healthcare Big Data

Madhu H. K., D. Ramesh

Abstract: ***An accurate diagnosis of the healthcare-based Big Data will always demand a significant level of quality in its input data itself, which is a serious level of concern in the area of healthcare analytics. Review of existing approaches shows that there has been various learning-based approaches being used for disease diagnosis which often ignores various issues viz. data aggregation, presence of error prone data, accuracy etc. Therefore, this paper presents a novel framework which offers cost effective modeling of the aggregation process of healthcare-big data followed by facilitating solution towards identifying and rectifying all the positions within a database system where there are presence of an error. The proposed system offer a mechanism where the error-prone data has been identified and substituted with data of better quality in order to offer better analytical outcomes. The study offers a strong baseline in order to leverage the data quality in healthcare big data.***

Keywords: ***Analytics, Big Data, Data Aggregation, Error, Healthcare, Medical Data.***

I. INTRODUCTION

The digitization of processes involved into healthcare industry leads to generation of data from various sub-systems of the healthcare. Taking a scenario of modern hospital, when a patient approach to the hospital either first time or repeat consultation, their details including personal and health related are updated into the digital patient record system. When the registered patient assigned to the respect medical professional, they update the patient dieses or illness history in the Electronic patient history module. Further, the patient is directed for procedure or radiologist, there also the results of different test are maintained digitally. The digital data stored in the various stages of patient care keep growing over period of time as for many reasons for future analysis the data need to kept in reposit for long period of time that are used for artificial intelligent based predictions or recommendations. This leads to grow the size of the medical data in various forms of text; images etc. and comply the one of the characteristics of the BigData called volume. The voluminous data require mechanism of distributed file system on the cluster of computers [1-4]. The popular adoptability is setting up cluster on a cloud now days. Another important aspect for the knowledge discovery paradigm on digital medical records is handling its heterogeneity as these records poses the formats of structured, semi-structured and unstructured that are generated in the forms of sensory data, text record data, and various radiological data. These heterogeneities of medical data are correlated with another characteristic of big data called variety.

At present, there are various challenges associated with the medical big data analysis. The first challenge associated with it is the need of a process that can ensure an effective cleaning of the data in order to retain informative contents within it. This represents that cleaning is an essential operation association with data quality which can also improve the accuracy directly. The next challenging factor is related to the storage problems. Usually, medical big data are higher in size as well as in dimension and it significantly occupies more space. Raw storage of data will lead to significant dependency of resources in order to carry out query processing, which will be always time consuming and may not be accurate. Hence, it is preferred to store the data in analyzed form so that faster query processing can be supported along with lesser time consumption. This will give rise to another challenge which is related to query processing. A better form of query processing demands a good indexing policy as well as meta-data management. Various conceptual practices e.g. semantics and ontology can also be utilized in order to improve the method of inference [5]. The next challenging part is to offer a significant level of accuracy in the prediction process. This is quite a challenging task

The distribution and collaborations of various units of medical establishments due the global compulsion of medical industry exhibits the flow of medical-BD (Medical Bigdata) that induces another important characteristic of the Big Data namely velocity. Once these characteristics are added together then obviously the data introduces uncertainty into it that finally leads to very special characteristics of BigData into medical-BD as veracity. Therefore, the mechanism of knowledge discovery and analytics for medical-BD requires a suitable and efficient framework. This paper introduces such distributed framework architecture that is capable of analyzing medical big data using cost effective solution. The core idea of this paper is to offer significant level of data quality [6-9]. The organization of this paper is as follow: The Section II describes the review of literature for the related work while highlight of research problem is presented in Section III.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

Madhu H. K.*, Assistant Professor, Department of Master of Computer Applications, Bangalore Institute of Technology, Bengaluru, India. E-mail: madhuhkbit@gmail.com

Dr. D. Ramesh, Professor and HOD, Sri Siddaratha Academy of Higher Eduction, Tumkur, India.

Section IV discusses about research methodology while algorithm implementation is discussed in Section V. Result Analysis is carried out in Section VI while conclusive remarks are written in Section VII.

II. RELATED WORK

This section is a continuation of the review work discussion in our prior work [10]. The most recent work carried out by Alkouz et al. [11] have developed a model that is capable of performing forecasting of the disease on the basis of social network platforms. The study makes use of the linear regression model over the data for improving the prediction performance. The work carried out by Kumar and Singh [12] has discussed about the different aspects of healthcare based big data with respect to the database. Study toward mining-based implementation has been carried out by Sun et al. [13] which deals with the knowledge discovery process associated with eth chronic disease over heterogeneous network. Zayood et al. [14] have presented a probabilistic learning approach towards medical big data where the technique deals with a unique learning process. The technique makes use of the log files in order to offer inference towards the mined data right from the workflow.

The work carried out by Yu et al. [15] have discussed about the significance of the learning mechanism applied over medical big data. According to the study, it says that irrespective of challenges associated with the deep learning mechanism, it is still the best solution. Nearly, similar category of the review work has also being carried out by Ye et al. [16] which also concludes that adoption of artificial intelligence is one of the best option for analysis of disease condition from medical data. A unique implementation of the medical data has been also reported to be used in preserving privacy factor associated with it. The author Yang et al. [17] has used homomorphic encryption approach in order to perform extraction of the factors that are demanded for training purpose. The study also makes use of bloom filter for performing the prediction operation. Study towards formulating an analytical approach was presented by Jindal et al. [18] where a clustering process has been carried out towards medical data while fuzzy logic is applied for improving the classification performance. Significance of analytical approach was also presented by Wu et al. [19] where the study has been carried out by incorporating omic-data to perform analysis of medical data. The study has presented different modeling approach along with their essential functionalities. Adoption of hierarchical mining of patterns of data is seen in the work of Shah et al. [20] where the emphasis is offered towards contextual data. The authors have used association rule with cross level as well as heavy hilters of hierarchical form for pattern mining. The idea of this implementation is to explore the similarity factors occurring between the items.

Another unique study towards pattern identification in healthcare analytics has been carried out by Yassine et al. [21]. The authors have used frequent pattern mining approach along with applying clustering operation where the prominent idea is to analyze the energy patterns associated with the appliance level. Shakhgeldyan et al. [22] have discussed about the importance of statistical data and its individual processing level in healthcare sector. The works of Hossain and Muhammad [23] have presented an integrated use of features associated with voice based big data for assessing the presence of pathology. The study also implements learning approach of Gaussian mixture model, extreme learning, and support vector machine for performing an effective classification. Huda et al. [24] have presented a mechanism that can perform selection of an effective feature in order to perform diagnosis of critical brain disease. The prime aim of the work is mainly to perform faster analysis in presence of unstable clinical data where the wrapper filter is utilized in order to carry out feature selection. The work of Yang et al. [25] have presented a mechanism of self-learning that is capable of updating the knowledge model as per the data stream in order to carry of diagnosis. The study has also implemented semantics in order to facilitate better inference of the disease identified. Ho et al. [26] have developed a scheme where the data associated with air pollution has been used for offering event-based alert. Another work carried out by Forkan et al. [27] has presented contextual based tracking system for the patient over the cloud environment.

The authors, Kuo et al [28] has designed a framework for the analysis of the real-time health care related Big Data by using a high performance multi-core cluster of the computers to handle hospital records of large dimension through a custom warehouse system. The critical customization emphasis on the changes in the indexing aspect and typical configuration of traditional big Data management tool of HBASE The framework basically handles the storage aspects and its fault-tolerance by maintaining a provision of replications and limited query options which is optimized through additional layer configuration with an objective of the patient data migration to the traditional HBASE with patient detail denotification. This approach of the framework does not really handle the distribution health care data integration and handling the unstructured data handling capacity which limits its applicability in the analytics capacity. In health care industry the data size is continuously growing and the existing data management tool integration architecture is not adequate for the real-time processing that limits its use for the critical cases. In the work of Benhlima et al. [29] steam as well as the batch processing-based architecture is proposed that ensures reliability for the alter generation in critical cases but the work is an optimization in the existing tools like mango-DB and Spark to customize fast search process which does make the work qualify a desired fame work.

The adoption of machine learning is an obvious option for the analytics development but the challenges for developing machine learning models for the health care related big data are distinguished. In this direction, B. Xu et al. [30] proposes an framework to the purpose of the analysis of the data from the distributed sources using machine learning but the aspects of the data purity as veracity factor is not considered.

Therefore, it can be seen that there are various approaches in existing times that emphasizes over performing analysis of the healthcare big data. There are various approaches addressed in existing work that target different set of problems which are mainly associated with either classification or prediction.

Learning-based approaches are found to be dominant in existing practices towards analytical operation. The next section discusses the research problem identified after this review.

III. RESEARCH PROBLEM

There are various issues associated with the big data problems especially when it related to the healthcare sector. This section discusses about the research problems that has been identified after reviewing the existing approaches and is considered to be addressed in current study. Following are the identified research problems:

- *Low emphasis on data aggregation*: With the adoption of distributed enterprises all over for speeding up the process in application, it is essential to manage such higher scale of different variants of data. The present work directly takes the database and start analyzing it without even mapping the distributed method of data generation from various terminals of source of data origination point. If the distributed data are not aggregated in one place, it is nearly impossible to execute cost effective analytical operation on the top of it. It will save lot of time and conserve various resources in order to perform processing and analysis of data. Moreover, the data generated in distributed manner are high unstructured and cannot be directly subjected for analysis.
- *Less focus on source of error*: There are various reasons of incorporation of error over the data. The first possibility could be human-related error while the second possibility could be use of flawed system which captures the streamed data. Even if these errors are rectified, the third possibility will be network-related error especially the application is hosted over wireless network. Until and unless these possibilities are not considered in modeling a better form of data quality cannot be ensured which is highly demanded in the area of healthcare sector.
- *Lack of in-depth review on data*: Every data which is sourced from the healthcare unit has various information fed within it. However, not all the data are essential for catering up a specific query system. It all depends upon what kind of knowledge is required to be mined. Even if all data are not essential, it is really mandatory that data should be 100% complete in its own form. It will mean that all the cells must be filled with logical value associated with the electronic health record. Therefore, before carrying out data analysis, this charecteristics of the significance of data is required to be considered, which is not found in any existing approaches.
- *Reduced emphasis on accuracy*: Essentially, the information obtained from healthcare big data are used for predictive analysis of any specific disease. Therefore, accuracy is highly essential while performing prediction. However, this accuracy at the last stage of prediction is deeply affected if the data that are considered as an input is not free from errors. Basically, errors will mean difference between original forms of value added data with the error-prone data. These errors are required to be found from the massive input data before even subjected to analytical operation. Existing studies has no consideration about the existence of such form of errors which significant affect the accuracy of analysis.

All the above mentioned research problems are identified and addressed in proposed solution discussed in next section.

IV. RESEARCH METHODOLOGY

The proposed research work is carried out considering an analytical model that is being framed up exclusively for performing knowledge discovery related to healthcare sector. For this purpose, a computational framework is designed for storage and programming model virtually. The proposed work contributes to construct a healthcare-based knowledge repository system and development of search engine that also enhances the capability of search optimization in diagnostic data on the top of BigData. The initial part of the implementation focuses on developing a virtual platform that enables domain applications to access live streams of healthcare data. A large number of streams of healthcare data arriving continuously upon a data-center from hundreds and thousands of clusters located in distributed manner can swamp even the most robust clusters in cloud. In order to support the historical data feeds, the framework is also required to store data from feeds in a format easily compatible in cloud environment. In order to store a continuous feed along with queries from large numbers of clusters requires highly scalable and efficient storage and retrieval mechanisms. Therefore, the next part of implementation focuses on developing a technique for streaming data to cloud and store as universal format for fulfilling the goal of the study. This also solves the problem associated with processing healthcare big data. The conceptual framework of proposed system introduces a buckets (B) corresponding to the various different locations of the medical-BD streaming $B_n = \{U_1, U_2, \ldots U_i\}$, where $i=n$ and U_i is the ith healthcare establishments from where data is generated and supposed to be streamed to the Analytical Unit (AU) which is a centralized system as shown in the Figure 1.

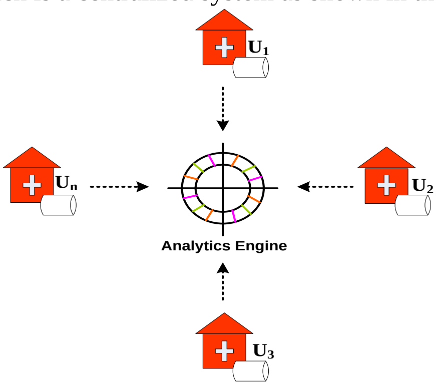

Fig.1 Bucket distribution for healthcare units

The Data Analytics Unit AU is connected to the Aggregation Center (AC) where all the data stream from the various departments are aggregated i.e. $D=\{D_1, D_2, D_3, D_4\}$, where D_1=Pharmacy, D_2=Laboratory, D_3= Radiology, and D_4=Narrative are fetched into the respective data storage name node-bucket as shown in the process architecture of the framework in the Figure 2.

It is to be noted that proposed system model considers one direction of the traffic stream which is from the healthcare sector to the cloud considering different forms of the data.

As the data will arrive from the network, there are fair possibilities of various inconsistencies owing to network-related errors. The process of data quality enhancement is carried out over cloud.

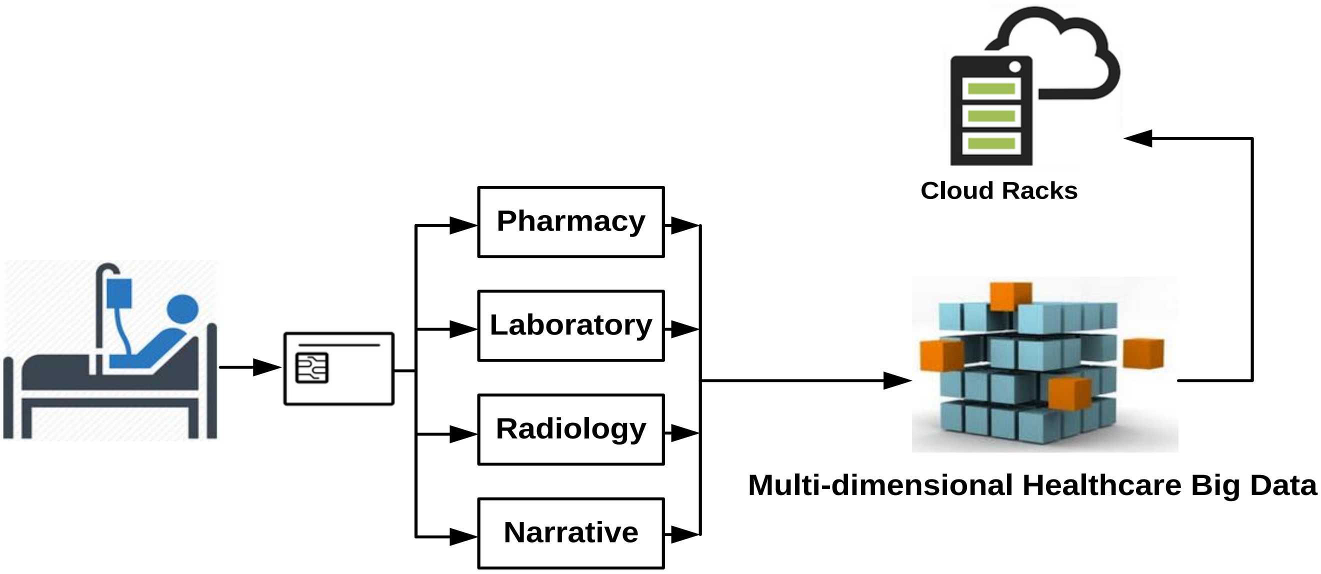

Fig.2 Data Stream and aggregation from different department process architecture of the framework

V. MODEL VALIDATION

The initial model of the framework is design to setup a distributed cluster of the nodes for the various department data fetching, in order to minimize the replication process, the existing data-buckets are removed in case of initialization of the new datastore creation after a data archival, by a custom function *f*(D) with all the contents, then depending upon the number of functional department, the data bucket generates. The framework generates a synthetic data for the healthcare sector as HBD, with the set-up record set as shown below.

Algorithm for Synthetic Big Data Generation

```
Input: Ds :Size of the Data in GB
Output: HBD(healthcare big data)
Process
Start
    1. HBD.ID$\leftarrow$f($T_f$, P)
    2. for : each value of the One GD Data as N
              T1 : Token-1 : Token ID → Value-1
              T2 : Token-2 : Token ID → Value-2
              .
              Tk : Token-k : Token ID → Value-k
              TN : Token-N : Token ID → Value-N

       end
    3. HBD$\leftarrow$write($T_1$,$T_2$,..$T_k$,..$T_N$)
```

The process module to generate the synthetic data HBD, utilizes a custom function and takes target file (T_f) and the permission (P) to generate an ID of the HBD→T_f(Line-1), In case of error in the HDB process held to chock condition. The HBD is updated with a syntax of the data as : TN : Token-N : Token ID → Value-N, (Line2-3)where N is the size in GB. The generated HBD, maintains a temporary replication in the reposition of datastore input. This is essential in order to support the virtualized environment over the cloud interface as well as to offer better data availability for the user. The next part of the proposed system is to offer data quality in order to address the problem where data could be possibly corrupted during the transmission over the networks. For this purpose, the proposed system formulates another algorithm which further carries out processing as well as improving the quality of the data. The steps of this algorithm are as follows:

Algorithm for Data Processing & Data Quality

```
Input: HBD (healthcare big data)
Output: $M_{final}$ (matrix with quality data)
Start
    1. For D:1:$D_{max}$
    2. Extract info→$c_1$, $c_2$, $c_3$
    3. entries→card(info)
    4. M$\leftarrow$(entries)$^{pos}$
    5. $M_{test}$$\leftarrow$$g$(M)
    6. identify $d_{imp}$→pos($M_{test}$)
    7. cell($d_{imp}$)→corr(sem(M))
    8. $M_{final}$$\leftarrow$$arg_{max}$((cell(dimp))
    9. End
End
```

The algorithm takes the input of HBD (healthcare big data) which after processing leads to an outcome of M_{final} (matrix with quality data). The algorithm considers all the respective data within the HBD (Line-1) which is followed by all the information row-wise in the input HBD. The study considers that a single row considers three types of contents i.e. c_1, c_2, and c_3 (Line-2), where c_3 is considered as main information that is associated with respective categories c_1 while c_2 acts as a connecting bridge between c_1 and c_3.

The algorithm than computes the cardinality *card* of this information *info* and store them in another matrix called as *entries* (Line-3). A new matrix is constructed which stores all these entries with respect to their individual position (Line-4). The next part of the algorithm executes a function *g*(x) which randomly introduces errors within the matrix and now the matrix is called as test matrix M_{test} (Line-5). By error, it will mean that certain information associated with c3 will be missing or will be intentionally rendered redundant. The algorithm finds out the position *pos* of such entries from M_{test} matrix (Line-6). This information is now stored in another matrix called as impure data matrix d_{imp} (Line-6). The algorithm then runs a correlation operation in order to find the error prone value on its respect cell position with other entries in super matrix M (Line-7). While performing this operation, the algorithm executes a semantic-based operation where the semantics of the words present in c_3 is obtained. During the search for the best matched result, the algorithm stops its search once it gets the best match. The obtained value is then substituted to the final matrix M_{final} (Line-8). Therefore, it can be seen that the proposed system can successfully find the error factors that degrades the quality of data and then can substitute them with quality data. The next section discusses about the outcome obtained.

VI. RESULT ANALYSIS

The complete scripting of the proposed system is carried out using MATLAB over normal windows machine. The proposed implementation is carried out by reviewing standard big data that is publically available [31]. The evaluation of the outcome obtained is carried out on the basis of three performance parameters i.e. time taken for processing, accuracy obtained, and data quality obtained. The analysis is carried out for 1000 simulation trials considering 10% of HBD, 50% of HBD, and 100% of HBD. The total size of test HBD is 150 GB. The database consists of categorical information which is fixed headers for all the entries and its respective values. The increase of simulation round will allow uneven and random traffic stream towards the analytical engine to find the impact of dimension of data over the performance parameter in order to map with real world application.

A. Analysis of Processing Time

Processing time is calculated as overall time consumed for successfully running the algorithm. Table 1 highlights the numerical outcome of it to show that processing time increases with increase with the size of data (HBD). However, a closer look on this numerical trend will show that differences of increment of processing time with all consecutive iteration in seconds are quite low and insignificant. Although, lower size of HBD shows faster processing; however, they are not practical as big data will always have massiveness in the size of the data. Therefore, the outcome shows considerably less processing time for proposed system.

Table.1 Analysis of Processing Time

Simulation Trial	Time Consumed (s)		
	10% HBD	50% HBD	100% HBD
100	0.001	0.006	0.008
200	0.004	0.008	0.010
300	0.005	0.009	0.011
400	0.006	0.010	0.012
500	0.007	0.011	0.016
600	0.010	0.015	0.018
700	0.012	0.018	0.021
800	0.013	0.019	0.021
900	0.015	0.021	0.024
1000	0.018	0.022	0.026

B. Analysis of Accuracy

The proposed system basically computes the error which is numerical difference between the clean data and unclean data. Accuracy is derived from it for better representation of the identification of the cell position which has error-prone information. The logic of this analysis is that if the detection of cell location can be carried out effectively than the system will have more leverage towards fault identification followed by replacement of error free data. The numerical outcome shown in Table 2 highlights that accuracy is slightly affected for complete size of test HBD whereas it is better when the size reduces down. It is quite natural as with the increase of data size, the extent of errors will be quite higher. However, the effectiveness in accuracy of the proposed for full size of data is 82.1% while that of half the size of data is 91.4%. The outcome for smaller size of HBD is witnessed with 100% of accuracy; however, this outcome is not considered as this size is quite smaller in contrast to the higher proportion of the data.

Table.2 Analysis of Accuracy

Simulation Trial	Accuracy (%)		
	10% HBD	50% HBD	100% HBD
100	100	100	100
200	100	100	100
300	100	98	95
400	100	96	91
500	99	94	90
600	98	92	89
700	97	89	75
800	95	86	62
900	92	81	60
1000	91	78	59

C. Analysis of Data Quality

For any analytical algorithm to be effective, it is necessary that data should be of higher quality. Basically, quality data in this case will mean a complete data where the algorithm contributes to extract and substitute the error prone data with new data. The matrix filled with new data can be referred to as quality data. Table 3 highlights the numerical analysis of data quality over various size of HBD.

Table.3 Analysis of Data Quality

Simulation Trial	Data Quality (%)		
	10% HBD	50% HBD	100% HBD
100	100	100	100

200	100	100	100
300	100	100	100
400	100	100	100
500	100	100	100
600	100	100	99
700	100	100	99
800	100	99	98
900	100	99	98
1000	100	99	98

The above outcome shows that proposed system offers 99.2% of data quality for complete data while it can offer 99.7% of data quality for 50% of HBD. It eventually means that it can always maintain a better data quality.

VII. CONCLUSION

This paper has presented a very unique and cost effective solution towards addressing the problems associated with data quality in healthcare sector. The contribution of the proposed study are as follows: i) it supports data aggregation process from different healthcare units which is necessary in order to apply faster query processing, ii) the study applies data analysis over aggregated data collected from distributed sources and not over individual data, this save significant time to carry out data analysis with higher probability of accuracy. Hence, they are cost effective, iii) the model can identify the position of error-prone data and can automatically substituted the error data with computed data. Therefore, the proposed operation carried out is meant for involuntary identification of artifacts and rectifying them unlike any existing system. Our next work will be focus towards further optimizing the process with an inclusion of new challenges in analyzing healthcare big data.

REFERENCES

1. Jiang, Ping, Jonathan Winkley, Can Zhao, Robert Munnoch, Geyong Min, and Laurence T. Yang. "An intelligent information forwarder for healthcare big data systems with distributed wearable sensors." IEEE systems journal 10, no. 3 (2014): 1147-1159.
2. Nepal, Surya, Rajiv Ranjan, and Kim-Kwang Raymond Choo. "Trustworthy processing of healthcare big data in hybrid clouds." IEEE Cloud Computing 2, no. 2 (2015): 78-84.
3. Zhang, Yin, Meikang Qiu, Chun-Wei Tsai, Mohammad Mehedi Hassan, and Atif Alamri. "Health-CPS: Healthcare cyber-physical system assisted by cloud and big data." IEEE Systems Journal 11, no. 1 (2015): 88-95.
4. Ekblaw, Ariel, Asaph Azaria, John D. Halamka, and Andrew Lippman. "A Case Study for Blockchain in Healthcare:“MedRec” prototype for electronic health records and medical research data." In Proceedings of IEEE open & big data conference, vol. 13, p. 13. 2016.
5. Mezghani, Emna, Ernesto Exposito, Khalil Drira, Marcos Da Silveira, and Cédric Pruski. "A semantic big data platform for integrating heterogeneous wearable data in healthcare." Journal of medical systems 39, no. 12 (2015): 185.
6. Lee, Choong Ho, and Hyung-Jin Yoon. "Medical big data: promise and challenges." Kidney research and clinical practice 36, no. 1 (2017): 3.
7. Anuradha, J. "A brief introduction on Big Data 5Vs characteristics and Hadoop technology." Procedia computer science 48 (2015): 319-324.
8. Yeh, Hseng-Long, Chin-Sen Lin, Chao-Ton Su, and Pa-Chun Wang. "Applying lean six sigma to improve healthcare: An empirical study." African Journal of Business Management 5, no. 31 (2011): 12356.
9. Lee, Choong Ho, and Hyung-Jin Yoon. "Medical big data: promise and challenges." Kidney research and clinical practice 36, no. 1 (2017): 3.
10. Madhu H K, Prakash B R, “A Survey: Big Data Ethics and Challenges in Healthcare Division”, International Journal of Computer Science and Engineering, Vol. 7, Issue.3, pp. 16-24, 2019
11. Alkouz, Balsam, Zaher Al Aghbari, and Jemal Hussien Abawajy. "Tweetluenza: Predicting flu trends from twitter data." Big Data Mining and Analytics 2, no. 4 (2019): 248-273.
12. Kumar, Sunil, and Maninder Singh. "Big data analytics for healthcare industry: impact, applications, and tools." Big Data Mining and Analytics 2, no. 1 (2018): 48-57.
13. Sun, Chenfei, Qingzhong Li, Lizhen Cui, Hui Li, and Yuliang Shi. "Heterogeneous network-based chronic disease progression mining." Big Data Mining and Analytics 2, no. 1 (2018): 25-34.
14. Zayoud, Maha, Yehia Kotb, and Sorin Ionescu. "$\beta $ Algorithm: A New Probabilistic Process Learning Approach for Big Data in Healthcare." IEEE Access 7 (2019): 78842-78869.
15. Yu, Ying, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. "Clinical big data and deep learning: Applications, challenges, and future outlooks." Big Data Mining and Analytics 2, no. 4 (2019): 288-305.
16. Ye, Mao, Hangzhou Zhang, and Li Li. "Research on Data Mining Application of Orthopedic Rehabilitation Information for Smart Medical." IEEE Access 7 (2019): 177137-177147.
17. Yang, Xue, Rongxing Lu, Jun Shao, Xiaohu Tang, and Haomiao Yang. "An Efficient and Privacy-Preserving Disease Risk Prediction Scheme for E-Healthcare." IEEE Internet of Things Journal 6, no. 2 (2018): 3284-3297.
18. Jindal, Anish, Amit Dua, Neeraj Kumar, Ashok Kumar Das, Athanasios V. Vasilakos, and Joel JPC Rodrigues. "Providing healthcare-as-a-service using fuzzy rule based big data analytics in cloud computing." IEEE journal of biomedical and health informatics 22, no. 5 (2018): 1605-1618.
19. Wu, Po-Yen, Chih-Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Hoffman, and May D. Wang. "–Omic and electronic health record big data analytics for precision medicine." IEEE Transactions on Biomedical Engineering 64, no. 2 (2016): 263-273.
20. Shah, Zubair, Abdun Naser Mahmood, Michael Barlow, Zahir Tari, Xun Yi, and Albert Y. Zomaya. "Computing hierarchical summary from two-dimensional big data streams." IEEE Transactions on Parallel and Distributed Systems 29, no. 4 (2017): 803-818.
21. Yassine, Abdulsalam, Shailendra Singh, and Atif Alamri. "Mining human activity patterns from smart home big data for health care applications." IEEE Access 5 (2017): 13131-13141.
22. Shakhgeldyan, Karina J., Nikolay A. Stepanov, and Boris I. Geltser. "The use of big data for extraction and processing of statistical data of the private healthcare sector." In 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC), pp. 162-165. IEEE, 2017.
23. Hossain, M. Shamim, and Ghulam Muhammad. "Healthcare big data voice pathology assessment framework." iEEE Access 4 (2016): 7806-7815.
24. Huda, Shamsul, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, and Michael Buckland. "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis." IEEE access 4 (2016): 9145-9154.
25. Yang, Shuo, Ran Wei, Jingzhi Guo, and Lida Xu. "Semantic inference on clinical documents: combining machine learning algorithms with an inference engine for effective clinical diagnosis and treatment." IEEE Access 5 (2017): 3529-3546.
26. Ho, Kin-Fai, Hoyee W. Hirai, Yong-Hong Kuo, Helen M. Meng, and Kelvin KF Tsoi. "Indoor air monitoring platform and personal health reporting system: big data analytics for public health research." In 2015 IEEE International Congress on Big Data, pp. 309-312. IEEE, 2015.
27. Forkan, Abdur Rahim Mohammad, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari. "BDCaM: Big data for context-aware monitoring—A personalized knowledge discovery framework for assisted healthcare." IEEE transactions on cloud computing 5, no. 4 (2015): 628-641.
28. M. Kuo, D. Chrimes, B. Moa and W. Hu, "Design and Construction of a Big Data Analytics Framework for Health Applications," 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, 2015, pp. 631-636
29. Benhlima L. Big data management for healthcare systems: architecture, requirements, and implementation. Advances in bioinformatics. 2018;2018.
30. B. Xu and S. A. Kumar, "Big Data Analytics Framework for System Health Monitoring," 2015 IEEE International Congress on Big Data, New York, NY, 2015, pp. 401-408
31. How To Get Experience Working with Large Datasets, http://www.bigfastblog.com/how-to-get-experience-working-with-large-datasets, retrieved on 17-02-2020

AUTHORS PROFILE

Mr. Madhu H. K., is a research scholar at Sri Siddhartha Institute of Technology, Tumkur(SSAHE), having nineteen years of Teaching expereance at Department of M C A, in Bangalore Institute of Technology, Bengaluru. His research interest includes Data Mining and Big Data Analytics.

Dr. D. Ramesh, Professor and HoD from Sri Siddaratha Academy of Higher Eduction,Tumkur, India. His vision is to emerge as a centre of excellence for imparting technical knowledge in the field of computer applications, nurturing technical competency and social responsibility among budding software professionals. He has around 28 years of techning experience.