A Novel Integrated Framework to Ensure Better Data Quality in Big Data Analytics over Cloud Environment

C. S. Sindhu[1], Nagaratna P. Hegde[2]
[1]JNIAS, Hyderabad, Global Academy of Technology, Bangalore
[2]Dept of CSE, Vasavi College of Engineering, Ibrahimbagh, Hyderabad, India

Article Info

Article history:

Received Mar 23, 2017
Revised Jun 18, 2017
Accepted Sept 11, 2017

Keywords:

Big data analytics
Data speed
Data uncertainty
Data variety

ABSTRACT

With advent of Big Data Analytics, the healthcare system is increasingly adopting the analytical services that is ultimately found to generate massive load of highly unstructured data. We reviewed the existing system to find that there are lesser number of solutions towards addressing the problems of data variety, data uncertainty, and data speed. It is important that an error-free data should arrive in analytics. Existing system offers single-hand solution towards single platform. Therefore, we introduced an integrated framework that has the capability to address all these three problems in one execution time. Considering the synthetic big data of healthcare, we carried out the investigation to find that our proposed system using deep learning architecture offers better optimization of computational resources. The study outcome is found to offer comparatively better response time and higher accuracy rate as compared to existing optimization technqiues that is found and practiced widely in literature.

Corresponding Author:

Name of Corresponding Author,
Department of Electrical and Computer Engineering,
National Chung Cheng University,
168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan, ROC.
Email: lsntl@ccu.edu.tw

1. INTRODUCTION

Cloud computing is not a new name in the industry from the past half a decade. There are various revolutionary enterprise applications that are highly optimized over cloud environment [1]. This optimization is mainly in terms of high performance which is normally scaled in the form of Service Level Agreement (SLA) and Quality-of-Experience (QoE). The biggest boon of cloud computing is its offering of the pervasive environment that offers accessibility of user data at any point of time [2]. However, this greatest contribution of pervasiveness of data by cloud computing has given rise to a biggest impediment in the area of data mining. It is because of such pervasiveness, multiple types of data bearing different forms of information, modalities, and formats are now increasingly populated over cloud server almost in every second [3]. Certain data are physically pushed to cloud by user but majority of the data that exists in cloud are generated by autonomous system e.g. sensors [4]. Eventhough there is abundant usage of sensors in automation and security systems, there is one more area where sensors play a vital role i.e. Healthcare sector [5]. The modern system exercised in majority of bigger chains of hospitals have perfectly automated all tedious and manual task using sensors and have significantly tuned the respective information over the cloud servers in explicit manner. Apart from the sensor, the hospital also adopts the usage of Electronic Health Record (EHR) and Electronic Medical Record (EMR), which captures all the related information about the patient e.g. i) prior clinical history record, ii) personal contact record, iii) medication list, iv) attending and referral doctor information, v) insurance details, vi) surgery-related information, etc. The amount of such data just for one patient is so massive that it is really difficult to perform analysis. The problem of data analysis worsens when complete hospital data is considered. Basically, the prime reason of this problem is obviously the size of the data which could be in terms of petabytes, however, storage is still not a bigger challenge in the era of cloud computing. The bigger challenges related to processing such data are i) data variety, ii) data uncertainty, and iii) data velocity. Data variety is related to heterogeneity in the forms of the data which makes the data highly non-applicable to be subjected to any forms of data analysis. Data uncertainty speaks about the possibility of missing data as well as presence of ambiguous data due to which the complete volume of data becomes unreliable. This problem is quite difficult to handle and is mainly caused due to defective data storage and retrieval processing. The final problem i.e. data velocity is something which is quite difficult to be solved as the data arrival rate is unknown. At present, there is no such model or parameter that has been reportedly used to gauge the rate of data arrival from a particular source in a highly distributed network system. Although, there are existing software frameworks like Hadoop, MapReduce, Cassandra, neo4j and many more, all of them are mainly open source, which is never said to be secured as the adversaries mainly uses open source to initiate attacks over cloud. Apart from this, there are many cases wherein existing frameworks have reported problems. Therefore, we review some of the existing techniques of data analytical application which points towards medical data processing. From this we learn that this is still in its nascent stage and hence there is scope for evolving up with a solution. A significant research gap of integrated framework towards addressing majority of problems in data analytics in cloud in found. We, therefore, present a novel framework that has the capability to address multiple problems in one platform in highly cost effective manner using big data approach. Section 1.1 discusses about the existing literatures where different techniques are discussed for energy harvesting followed by discussion of problem identification in Section 1.2. Section 1.3 briefs about the proposed contribution to address research problems. Section 2 elaborates about the algorithm implementation followed by result discussion in Section 3. Finally summary of the paper is discussed in Section 4.

1.1. Background

This section discusses about the existing technqiues towards big data analytics. Our prior work [6-9] have discussed about the existing technqiues pertaining to significance of classification approach, tools of big data analytics, and applicability of big data analytics over healthcare sector. Chen et al. [10] have presented architecture for supportability of agile methodologies over big data. Dabek and Caban [11] have introduced a technique where a visualization mechanism is designed for modeling the user interaction. Fiadino et al. [12] have presented a discussion towards using cellular network in the viewpoint of big data analytics. Ordonez et al. [13] have introduced a technique that makes use of matrix multiplication in order to perform summarization of big data using statistical modeling. The technique uses array-based operators exclusively for sparse as well as for dense dataset to show memory consumption and scalability accomplishment. Paul et al. [14] have used the big data analytics in order to solve the problem associated with human behaviour. The author discusses about introducing a system that could act as a communication bridge between the internet-of-things application and data mining technqiues over larger data scale.Sheng et al. [15] have emphasized on the applicability of big data analytics with respect to information theory and cyber-physical system. This paper presents a superior form of mathematical modeling connected with information theory followed by significant channel models that could be potentially helpful for extracting knowledge from bigger scale of data.Tawalbeh et al. [16] have discussed the generation of massive loads of data from healthcare sector and how such forms of data can be analyzed using big data analytics. The problems pertaining to data uncertainty is being recently discussed by Wang and He [17]. The authors point out 6 potential problems to be overcome for future application of big data analytics e.g. i) highly complex representation of data, ii) pervasive uncertainty, iii) extremely weaker relationship among data, iv) computation-scalability problems, v) extraordinary massive size of complex data, and vi) larger number of classes involved in mining process. Similar direction of study considering Electronic Health Records-based data and its applicability over mining approach is discussed by Wu et al. [18]. The discussion highlights that dimensionality reduction is one of the prominent problems along with processing capability. Castellano et al. [19] have presented a classification-based approach for discriminating critical condition of arrhythmia. The authors have presented a unique technique for involuntary clustering of electrograms in presence of cloud environment. The technique is also claimed to offer lower computational load. The study outcome witness around 90% accuracy with 2.5% of error in classification performance. Cavallaro et al. [20] have used learning algorithm for performing an effective classification of images. Lu et al. [21] have presented a modeling of a concept that allows performing big data analytics over the cloud environment. The next section discusses about the problem being identified from the existing literature.

1.2. Problem Identification

The existing technqiues towards big data analytics has associated advantages witth respect to various individual applications; however, it is also associated with limitations too. This section briefs about the open research issues in the line of various potential problems of data processing particularly relating to the big data analytics. The identified problems in the existing system are as follows:

- Majority of existing literature have not considered many cases from healthcare sector. The data arriving from heathcare sector is highly complex in comparison to other forms of bigger data. Such problems are less addressed in existing literature.
- Existing system introduces big data analytics with more focus on applying mining operation and very less focus on performing processing operation on the top of it. Majority of the processing is left of existing software frameworks, which already has reported pitfalls.
- There are no research attempts where the data before storing over cloud undergoes processing in order to eliminate problems. Moreover, there is integrated system which offers mitigation procedure for data variety, data uncertainty, and data speed in existing literature.

1.3. Proposed Solution

The proposed system is a continuation of our prior study [22-23], where we have presented an individual algorithm for solving the problems related to big data analytics e.g. data variety, data uncertainty, and data speed. Figure 1 highlights the architecture of proposed system.

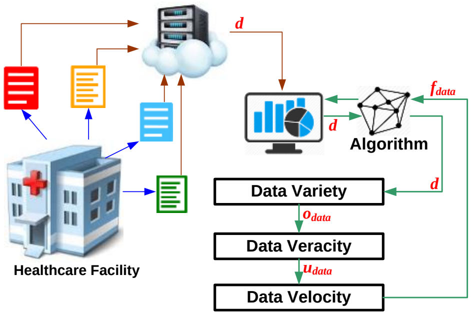

Figure 1 Proposed architecture

It is already known that there are different levels of complexities associated in addressing the problem related to medical big data analytics e.g. data variety, data uncertainty, and data speed. The proposed system offers an integrated framework where it is feasible to address all the three significant problems without using any sophisticated tool or expensive process. We assume that a healthcare facility already uses cloud server in order to stack out the generated information from the facility. However, we consider that our algorithm runs over such cloud server where before storing the data, the system implements a chain of algorithm in order to eliminate the problems associated in processing big data. We also assume that the data generated by the healthcare facility is highly unstructured and it should not be stored in this state that may further invite challenges during analytical operation. The unstructured data after arriving to cloud server it has to undergo processing from its first algorithm that removes the unstructuredness charecteristic of the data and make it more organized data. This data than further undergo processing with next chain of the algorithm to explore better substitution of the missing values or eliminiate the ambiguities. The obtained data from this part of the algorithm is then subjected to last part of the algorithm that makes the data modeling structure in such a way that the storage system will be able to store the arrived data irrespective of its time of arrival i.e. velocity. The next section further elaborated about such algorithms descriptively.

2. ALGORITHM IMPLEMENTATION

The framework discussed in the prior section mainly incorporates three types of algorithm targeting to countermeasures the problems associated with the medical data analysis of larger size. The algorithm takes the input of medical data which has the corresponding information about the patient e.g. patient id, hostpital name, Referral doctor, Date, Patient Name, Gender, Age, Attending Doctor, and Clinical History and perform a series of operation in order to overcome the problems associated with big data analysis.

2.1. Algorithm for Addressing the Variety Problem

The rationale of this algorithm is based on the fact that data with high score of variety are highly unorganized and hence it is required to provide a robust structure to this data so that further operation can be carried out. This algorithm is responsible for minimizing the problem associated with variety of medical data. The algorithm takes the input of unstructured data *d* which has different forms of variables of patient information residing under category *C* (Line-1). The algorithm read the text file and extracts textual contents on the basis of cardinality (*card*) of different entries as well as positions and finally stores it in a matrix (Line-2). The algorithm first finds the list of categories in the form of index and then takes one by one entry for creating new categories (Line-3). Finally, it is capable of finding the feasible number of array. The possibility of empty value in the data is also addressed in this system where it checks the size of all the matrix elements mat_{elem}. In case, a matrix element is found null than it is replaced by an integer value corresponding with the single matrix (Line-5-6). Finally, all the total entries are accumulated by exploring the size of the matrix. This process of data transformation finally generates organized data o_{data} free for data variety problem.

Algorithm for Addressing the Variety Problem

```
Input: $p_{id}, h_n, R_d, D, p_{name}, G, A, A_{doc}, C_{his}, C, d, E_{pos}$
Output: $o_{data}$
Start
1. init d, read C [$C=\{p_{id}, h_n, R_d, D, p_{name}, G, A, A_{doc}, C_{his}, C\}$]
2. Read & Store$\rightarrow$card (entries) & $E_{pos}$
3. Get $C_{index}$ & content
4. If $mat_{elem}=0$
5.         Substitute$\rightarrow mat_{elem}(Single_{Mat})$
6. $o_{data}\rightarrow mat_{elem}=[mat_{elem}\ ii]$
End
```

2.2. Algorithm for Addressing Uncertainty Problem

The design principle of this algorithm is based on the fact that uncertainty score in the medical big data may occur due to incompleteness of data or ambiguities in data sourcing model. Therefore, uncertainty problem in medical big data can be countermeasured if incomplete or data ambiguity is addressed in the micro-scale. The input to the study is the output from prior algorithm i.e. organized data, which is initially read (Line-1). In this case, all the field of the data sources are read and checked if all of them have undergone the process of removal of data variety problems. This process is then further followed by an iterative operation for checking the matched identification of the data corresponding to all the case studies of the hospital database. The main objective of this operation is to chek for scores of matched elements along with identity and respective categories. Hence, if any one of the field information is missing, the other key attributes e.g. atched elements along with identity and respective categories will assist to find the missing elements or even ambiguous elements in the matrix. Finally, based on the selected identity of the newly explored data, it is then stored. The mechanism also obtains the initial word from the columnar element and converts it to character followed by extraction of mean value. A new matrix is then formulated using unique element in order to generate a random number within a range of size of the matrix. This phenomenon will significantly avoid any kind of ambiguity in the data and only the matched data will be retrieved. Hence, if the category element do exists (Line-2) than it will use the same category identity (Line-3) or else it will generate a new category identity that corresponds with the size of category matrix (Line-5). The matrix with newly positioned data is then updated followed by calculation of data purity as the ranking mechanism of addressing uncertainty problems in medical big data (Line-7). We also compute error that could possibly occur during the substitution process to fill up missing or ambiguous data (Line-8); however, it is just a performance paramerer.

Algorithm for Addressing Uncertainty Problem
Input: o_{data} (Organized Data), C_{elem} (Category Element)
Output: u_{data} (updated data), d_{pur} (data purity), E (Error)

```
Start
1. Read $o_{data}$
2. If ($C_{elem} \neq 0$)
3.   use $C_{id}$
4. Else
5.   use $newC_{id} \rightarrow [C_{id}, size(C)]$
6. Update $mat_{elem}$
7. $d_{pur} \rightarrow size(unique(C), mat_{elem}) / card(C_{elem})$
8  E=E+size(newEntry)
9. Update $mat_{elem} \rightarrow u_{data}$
End
```

2.3. Algorithm for Addressing Speed Problem

The prime rationale of this algorithm is that data speed cannot be controlled by any means but we consider that if the speed of data is known to some extent than a proper data management could be done. Hence, we offer a very simple model where irrespective of any speed of arrival of the incoming data, the data could be efficiently stored in out database framework. The algorithm takes the output of prior algorithm where uncertainty problem is addressed. The algorithm initially read the column value of the incoming data (Line-1) and finds its numerical value. A classification of different matrix are generated depending on the last column of the incoming data (Line-2) followed by extraction of different labels too (Line-3). Basically, Labels corresponds to each individual cell in the matrix. We formulate a temporary matrix match that is responsible for identifying similar elements in incoming data and storage area. If the incoming data is found to have match than the data is directly discarded and only the column index is updated (Line-4). By this process, the algorithm ensures cost effective usage of data storage model by only considering unique incoming data to be stored. However, if the incoming data is found to be unique i.e. it has no matched elements between itself and data storage structure (Line-4), than it instantly update its label with respect to size of the label (Line-5). It will mean that a single label is sufficient enough to store the different frequencies of the incoming data, so that other part of the cells offers enough buffers for new arrival of massive data. Finally, all the explicit columnar information col is accomplished (Line-6) and cells are updated with respect to labels and output cells (Line-7). Hence, irrespective of any flow of the incoming data, the proposed system can offer significant room for data storage system in order to store the incoming data arriving from certain healthcare facilities.

Algorithm for Addressing Speed Problem
Input: u_{data}, col_{val}
Output: f_{data}

```
Start
1. find $\rightarrow col_{val}$
2. classify based on col
3. get Label
4. If match=0
5.    update Label $\rightarrow$ [1:size(label)]
6. Get col
7. $f_{data} \rightarrow$ update cells (Label, Output Cells)
End
```

Table 1 Notation used in Algorithm Design

Notation	Meaning
p_{id}	patient id
h_n	hostpital name
R_d	Referral Doctor
D	Date
p_{name}	Patient Name
G	Gender
A	Age
A_{doc}	Attending Doctor
C_{his}	Clinical History
C	Category
d	unstructured data
E_{pos}	Entry Position
o_{data}	Organized Data
C_{elem}	Category Element
u_{data}	updated data
d_{pur}	data purity
E	Error
col_{val}	numeric value of column
f_{data}	final data

3. RESULT ANALYSIS

This part of the study discusses about the results being accomplished from the proposed study. The implementation of the proposed system is carried out in Matlab. The constructed framework offer a highly comprehensive single environment for testifying the three significant problems associated with medical big data i.e. data variety, data uncertainty, and data speed. The reason for selecting Matlab for design and development is its easiness, convenient, and scalable. The implementation of the study was carried out using sythentic medical data of larger size. As the proposed system is nearly similar to optimize the performance of a framework for big data analytics, hence, it is wise enough to be compared with similar frequently used techniquesof optimization. We find that adoption of Support Vector Machine has been carried out by Cavallaro et al. [20] as well as by Singh et al. [24]. One of the optimization feature of SVM is its excellent classification approach that has capability to be applied over high-dimensioanl data and it is independent of any form of conventional feature selection procedures in over to resists the problem associated with larger dimensionality of big data. Different forms of regularization elements as well as extracted features are concatenated during the process of learning. In order to make the SVM works efficiently for big data, it is required for the class boundary to be very close to the outcome of the anticipated samples of training. Apart from SVM, we also find the adoption of neural network in optimizing big data analytics. The most recent work carried out by Chung et al. [25] has discussed the usage of neural network for high performance computation of big data. One of the significant advantages of applying neural network over big data analysis is its capability of approximation of any forms of function. This operation can be significantly helpful while exploring for the sub-space clustering in high dimensional data. A significant sigmoid function can be further fine tuned in order to arrive into ellite outcome using neural network in big data analysis. For simplicity, we implement support vector machine as the training algorithm to address the problem discussed in our paper. We take the feed-forward algorithm as the learning approach for neural network too. Hence, our existing system is a combined result arrived from implementing support vector machine and neural network. The comparative assessment of the study was carried out considering accuracy parameter and computational response time.

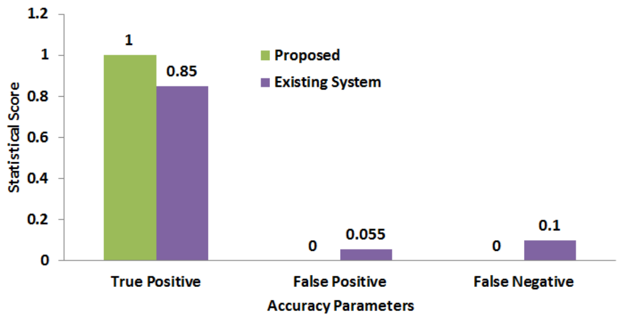

Figure 2. Comparative performances on accuracy

The comparative assessment of the accuracy is carried out using True Positive, False Positive, and False Negative for both proposed system as well as existing system. The study outcome shows that proposed system accomplishes better true positive from statistical score (probability) viewpoint, whereas the existing system offer slightly reduced accuracy performance as compared to proposed system. The prime reason behind this that the proposed system performs a sequential operation where output of first operation becomes input of second process. By this process, the proposed system offers better optimization without using any recursive function as well as with every increasing steps the problems get reduced. This procedure of filtering the problems is more iterative and less filtered and hence eixtsing system couldn't offer better accuracy.

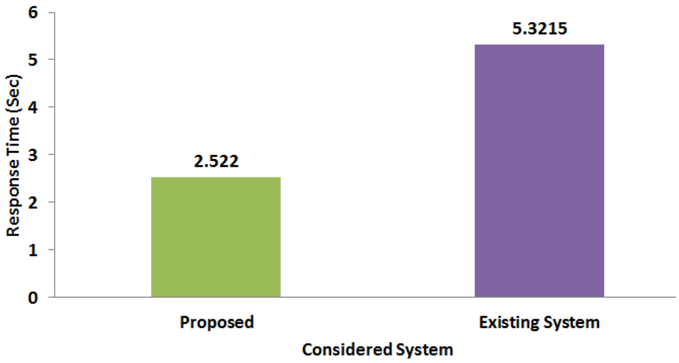

Figure 3. Comparative performances on response time

We also compute the amount of computational complexities associated with both proposed and existing system with respect to compuatational response time. Figure 3 clearly shows that proposed system offers better response time in comparison to existing system. However, because of too many dependencies on iteration, the existing system encounters the problem of faster convergence. Therefore, such iterative operations cannot be suitably applied over high-dimensional data thereby causing time lags to yield the outcome. On the other hand, the proposed system offers the results owing to faster convergence while making transition from one algorithm to other. Therefore, the proposed system offer better computational response time as well as lower memory consumption over normal machine to prove its cost effectiveness. The storage complexity is highly controlled as the complete framework offers results only over runtime processing of the algorithm and hence enough resource consumption is saved.

4. CONCLUSION

This paper emphasizes on the inherent problems associated with the big data analytics over cloud i.e. data variety, data uncertainty, and data speed. We have reviewed some of the recent literatures to find that there is a big tradeoff in the existing approaches between problems being addressed as claimed and real-time problems. The problems being addressed as claimed in literature only emphasizes on part of one problem where in real-sense there could be endless number of occurance of all the reported problems while performing real-time analysis over cloud environment. This paper describes one novel research model which is experimented over synthetic bigdata of healthcares sector. The technique introduces a novel framework where all the reported problems e.g. .data variety, data uncertainty, and data speed is addressed. At present, we find that it offer better accuracy and response time in comparison to existing optimization system. Our future work direction will be to further optimize the outcomes.

REFERENCES

[1] Mandal, Jyotsna Kumar, Handbook of Research on Natural Computing for Optimization Problems, IGI Global-Computer, 2016

[2] Ciprian Dobre, Fatos Xhafa, Pervasive Computing: Next Generation Platforms for Intelligent Data Collection, Morgan Kaufmann-Computers, 2016

[3] T Sutikno, D Stiawan, IMI Subroto, "Fortifying big data infrastructures to face security and privacy issues," *TELKOMNIKA Telecommunication Computing Electronics and Control*., vol. 12, no. 4, pp. 751-752, 2014.

[4] Anis Koubaa, Elhadi Shakshuki, Robots and Sensor Clouds, Springer, 2015

[5] Bhatt, Chintan M., Peddoju, S. K, Cloud Computing Systems and Applications in Healthcare, IGI Global, 2016

[6] S.P. Menon, N.P. Hegde, “Research on Classification Algorithm and its Impact on Web Mining”, International Journal of Computer Engineering and Technology, vol.4, Iss.4, pp.495-504, 2013,

[7] S.P. Menon, N.P. Hegde, “A Brief Insight into Computational Tools in Big Data,” International Journal of Innovation & Advancement in Computer Science, vol.4., 2015

[8] S.P. Menon, N.P. Hegde, “A Survey of Tools and Applications in Big Data,” IEEE 9th Interntaional Conference on Intelligent Systems and Controls, 2015

[9] S.P. Menon, N.P. Hegde, “The Critical Combined Role of Big Data Analytics in Health Care,” International Journal of Imaging Science and Engineering, 2015

[10] H. M. Chen, R. Kazman and S. Haziyev, "Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach," in *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 234-248, Sept. 1 2016.

[11] F. Dabek and J. J. Caban, "A Grammar-based Approach for Modeling User Interactions and Generating Suggestions During the Data Exploration Process," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 41-50, Jan. 2017.

[12] P. Fiadino, P. Casas, A. D’Alconzo, M. Schiavone and A. Baer, "Grasping Popular Applications in Cellular Networks With Big Data Analytics Platforms," in *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 681-695, Sept. 2016.

[13] C. Ordonez, Y. Zhang and W. Cabrera, "The Gamma Matrix to Summarize Dense and Sparse Data Sets for Big Data Analytics," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1905-1918, July 1 2016.

[14] A. Paul, A. Ahmad, M. M. Rathore and S. Jabbar, "Smartbuddy: Defining Human Behaviors Using Big Data Analytics in Social Internet of Things," in *IEEE Wireless Communications*, vol. 23, no. 5, pp. 68-74, October 2016.

[15] G. Sheng, X. Zhao, H. Zhang, Z. Lv and H. Song, "Mathematical Models for Simulating Coded Digital Communication: A Comprehensive Tutorial by Big Data Analytics in Cyber-Physical Systems," in *IEEE Access*, vol. 4, no. , pp. 9018-9026, 2016.

[16] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa and H. Song, "Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications," in *IEEE Access*, vol. 4, no. , pp. 6171-6180, 2016.

[17] X. Wang and Y. He, "Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies", IEEE Systems, Man, and Cybernetics Magazine, Vol. 2, No. 2, pp. 26-31, 2016

[18] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang, "–Omic and Electronic Health Record Big Data Analytics for Precision Medicine," in *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263-273, Feb. 2017.

[19] J. M. Lillo-Castellano *et al*., "Symmetrical Compression Distance for Arrhythmia Discrimination in Cloud-Based Big-Data Services," in *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1253-1263, July 2015.

[20] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson and A. Plaza, "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4634-4646, Oct. 2015. [21] Q. Lu, Z. Li, M. Kihl, L. Zhu and W. Zhang, "CF4BDA: A Conceptual Framework for Big Data Analytics Applications in the Cloud," in *IEEE Access*, vol. 3, no. , pp. 1944-1952, 2015.

[21] S.P. Menon, N.P. Hegde, “A Framework to handle Data Heterogeneity Contextual to Medical Big Data”, IEEE-International Conference on Computational Intelligence and Computing Research, 2015

[22] S.P. Menon, N.P. Hegde, “Predictive-based Data Analysis for Addressing Data Veracity Problems in Complex Medical Datak,” UNKNOWN

[23] D. Singh, D. Roy and C. K. Mohan, "DiP-SVM : Distribution Preserving Kernel Support Vector Machine for Big Data", in *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 79-90, March 1 2017.

[24] I. H. Chung; T. N. Sainath; B. Ramabhadran; M. Picheny; J. Gunnels; V. Austel; U. Chauhari; B. Kingsbury, "Parallel Deep Neural Network Training for Big Data on Blue Gene/Q," in IEEE Transactions on Parallel and Distributed Systems , vol.PP, no.99, pp.1-1