Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766278

AN OPTIMISTIC REVIEW AND ANALYTICS METHODOLOGIES FOR BIG DATA TECHNIQUES

Ms.Suman Choudhary
Research Scholar, Department of Computer Science & Informatics, Maharishi Arvind University, Jaipur, Rajasthan, India

Mr.Manish Kumatr Jha
Assistant Professor, Department of Computer Science, Institute of Technology and Management, Gorakhpur, UP, India

Prof(Dr)Mahaveer Kumar Sain
Professor, Department of Computer Science & Informatics, Maharishi Arvind University, Jaipur, Rajasthan, India

Mr.Vipin Singh
Assistant Professor, Department of Computer Science, Maharishi Arvind Institute of Science and Management, Jaipur, Rajasthan, India

Abstract—Big data (BD) and data analytics (DA) are increasingly trying to make their way into public policymaking, as well as there are demands for systematic evaluations and research agendas concentrating on the effects of BD and analytics on policy formation. Because of the rapid expansion of such data, various methods for extracting essential information and values from big data sets must be developed. Additionally, decision-makers must've been able to derive some meaningful insight from such huge and continuously changing data, which varies from ordinary transactions to customer interactions, including data from social networking sites. It is possible to achieve a such vision through the application of Big Data Analytics, that is deployment of Advanced Analytics Methods to massive volumes of data. The dilemma today is how to generate a large infrastructure for effectively analyzing large data and how to construct a suitable mining technique to extract relevant information from huge data. This paper continues with a quick overview of BD & then moves on to topics about large data analytics. Some critical outstanding challenges and future research paths for BDA will also be highlighted. This paper also attempts to investigate some of disparate analytics methodologies and tools that may be applied to BD, &the value given by BDA applications in various decision domains.

Keywords—Big data, Neural network (NN), machine learning (ML) techniques, data analytics, data mining (DM), Deep Learning (DL)

I. INTRODUCTION

Big data as well as data analytics have been seen as enhancing knowledge as well as eventually leading to improved decision-making. According to statements like "the end of theory," if big data & data analytics are widely used, our assumptions of techniques' revolutionary impact are

clear. Among an equal chance for the public policy processes to benefit from big data and analytics, public actors have been reluctant to become involved, even though the industry has been leading the charge [1].

As our ability to create and gather data grows exponentially, the need to use BD and DA is becoming more and more critical as a consequence. The phrases "big data" & "data analytics" have led to an increase in research, industry, as well as government use during the last several years [2]. BD and analytics are becoming more important in public policy, and this can be observed in the scientific literature. We're also seeing more and more public organizations using BDA to tackle issues like the environmental crisis & pandemics. A lot of the recent scholarly dialogue has focused on how to use BD and analytics in the policymaking process. BD and DA in public policy, nevertheless, lack a comprehensive picture of the present status and there are clearly defined research gaps [3].

The digital age, with all of its possibilities & complexity, overwhelms sectors & marketplaces that are confronted with a massive quantity of probable data in every transaction. Being conscious of the worth of obtained data & reaping the benefits of hidden information help to build a new model in our age that redefines what it means to be powerful for businesses. The power of information propels businesses toward greater agility and the achievement of their objectives. Industries are forced to characterize, diagnose, anticipate, prescribe, and cognate latent growth prospects as a result of BDA, which helps them, generate commercial value [4]. BDA employs modern analytical approaches to generate knowledge from an ever-expanding quantity of data, which will have an effect on a managerial procedure by reducing the complexity of the process and thereby boosting its effectiveness. BDA requires unique and advanced algorithms that can handle and analyze real-time data to provide high-accuracy analytics, which is currently lacking. ML and DL algorithms assign their sophisticated approaches in this procedure based on the issue approach taken into consideration. [5]

II. BIG DATA AND BIG DATA ANALYTICS

The accumulation of huge volumes of unstructured data is an important outcome of the digital age. The manager's consideration must be given to managing such precious capital in diverse forms and colors depending on the demands of the companies. A wide range of social and educational issues may be affected by big data. Controlling raw data becomes more critical as the volume of data grows, particularly in technology-based enterprises. To deal with raw data properties of BD, such as variety, velocity as well as volume, new tools are required. This is why BDA has been advocated for "research, modeling, BA, including monitoring," to name a few uses. Based on supervised and unstructured data input provided by ML, a predictive analysis may be performed using BDA technologies. A reciprocal relationship exists between ML analytics & data input: more specific & reliable data input, more successful performed exceptionally. As a part of ML, DL is used to discover patterns in data that are otherwise hidden [6].

III. BIG DATA ANALYTICS

BD has been presented in the digital age, which is recognized for its huge volume, diversity, veracity, velocity, as well as high value due to the increasing pace of data generation. To deal with the complexity as well as massiveness of diverse forms of data, an organization must use new approaches and tools in analytical aspects. In other words, BDA is a sophisticated method for dealing with the complexity of BD by estimating a massive no. of information [7]. Analyzing large amounts of data may help companies become more innovative, productive, and competitive [8].

It has been characterized as a set of approaches that are used to identify patterns and offer insight into intriguing relationships in context by investigating, processing, discovering, and displaying the outcome [9]. The way to big data analytics' benefits is paved with the reduction of complexity and management of cognitive strain in our knowledge-based society. In addition, feature identification is a significant aspect of BDA success. As a consequence, it is necessary to identify the key elements that have an important effect on outcomes. The next phase is to find correlations among input and a dynamic, predetermined point [10]. E-commerce and worldwide connection have thrived as a consequence of the rapid development of BDA. To better serve their residents, governments are also making use of big data analytics. Huge DA, a specialized application of this area, maybe applied for managing& analyzing BD in a corporate environment. In addition, social media data may be effectively controlled via the use of BDA. The five characteristics of BD—volume, velocity, value, diversity, and veracity—can be managed in this manner, allowing for a better understanding of consumer behavior.

By gaining a more complete understanding of customer behavior using big data analytics, businesses may also improve their ability to devise and implement new strategies [11]. Product recommendation systems and the website design are enhanced by small and medium-sized businesses using analytics to mine their semi-structured large data [12]. Big data analytics, as mentioned in Ref. [13], gets benefits by employing technology and procedures on their huge data to enhance the performance of a corporation. Because the decision-making process is aided by insights derived from the analysis of different data, BDA is of significant value. [14] As a result, the decision-making process will become one that is based on data. Accordingly, big data insights have been separated into data organization and analytics. The previous refers to the technical provision for collecting, organizing, and storing data; the latter refers to the methodologies that are used for data analysis. A sub-process called big data analytics has therefore been identified with insight extraction. Analysis and interpretation of any digital information may be done using big data analytics [15]. Data storage, data management, DA, & data visualization are all part of the process. Using big data analytics, a corporation's financial and business approaches may be improved, and it can be a game-changer in boosting productivity.

IV. BIG DATA ANALYTICS AND DEEP LEARNING

Even though deep learning was presented in the 1940s [16], the year 2006 was the year when Hinton presented the layer-wise greedy-learning technique to conquer the inability of the neural network (NN) way of finding optimized points by trapping in optima local point, which was

impaired after size of training data was insufficient. Deep learning methods have been around since then. The essential concept of Hinton's suggested technique is that unsupervised learning should be performed before layer-by-layer training takes place. Deep learning algorithms, which take their cues from the hierarchical organization of the human brain, extract complicated hidden information with a high degree of abstraction. Deep learning algorithms perform well when faced with big capacities of unstructured data. This is due to the layered plan of deep learning algorithms. Deep learning aims to do this by deploying numerous transformation layers, with each layer resulting in the appearance of an output representation [18]. Deep learning has yielded a wealth of previously undiscovered information, which has been incorporated into BDA. As a consequence of its primary characteristic, which is the extraction of underlying features from massive volumes of data?

It was only recently that DL as a specific field of ML became popular due to a combination of factors such as the rise of chip processing, which also consequences in the creation of massive amounts of data, the decrease in computer hardware costs, & significant advancements in ML techniques.

4.1 Convolutional Neural Networks (CNN)

As input, an image may be fed into a CNN (ConvNet/CNN), which can then determine the relevance of various characteristics or distance measures (with biases that can be learned). Compared to other different classifiers, ConvNets require far less pre-processing. ConvNets, on the other hand, maybe trained to learn certain filters/characteristics without the need for hand-engineering. Although another NN performs better with image, speech, as well as audio signal inputs, CNN outperforms them. They are made up of the following three primary strata:

1) Convolutional Layer (CL)

This layer is a fundamental structural component of a CNN, & it is also the vast bulk of computation that takes place. Only rare components are required: input data, filter, as well as a feature map, amongst other things. We'll adopt that the input is a color image, which is made up of three-dimensional (three-dimensional) pixels. This implies that input will have 3 dimensions: a height, a width, as well as a depth, that match the RGB values in a picture, as seen below. Another component is the feature detector, which is sometimes called a kernel or a filter. It will traverse over activation functions of a picture, checking to see whether characteristicsare there. Convolution is the term used to describe this procedure.

2) Pooling Layer (PL)

Dimensionality reduction is accomplished by the use of pooling layers, which is also called downsampling. The no. of variables in input is reduced. The pooling operation works in a similar way to CL in that it sweeps filter over whole input, with the exception that this filter doesn’t contain any weights. However, the kernel performs an aggregation function on data contained inside the receptive area, resulting in values within the output array being populated. Pooling may be divided into two main categories:

- Max pooling: With each pass of the filter over input, it choosesa pixel with max value and sends it to the output array for processing. As an addition, as contrasted to the corresponding pooling, this strategy is more often utilized.
- Average pooling: As the filter passes across the input, it determines the average value inside the input vector, which it then sends to the output array as even filter moves across the input.

3) Fully Connected Layer(FCL)

The FCL's name accurately describes what it is. As previously mentioned, since the input signal is not strongly associated with the output nodes, pixel values of the input images are not likely related to the output layer in partially associated layers. Especially contrasted to FCLs, every node in the output layer creates a connection to a node in the previous layer.

By characteristics gathered from previous layers including their various filters, this layer conducts the duty of categorization on the data. Unlike convolutional as well as pooling layers, which often utilize ReLu functions to categorize inputs, FC layers generally use a softmax activation function to accurately categorize inputs, providing a possibility among zero and one.

4.2 Deep Neural Network

In recent years, breakthroughs in computing techniques have enabled the introduction of a deep framework in supervised data, which is referred to as a deep neural network (DNN) [19]. In artificial intelligence, it derives from shallow artificial NNs, which are connected to deep learning. The fact that the hierarchical design of DL might include nonlinear data in a set of layers prompts DNN to use the layered framework with complicated functions to cope with a density as well as a large number of layers. In the classification field, DNN is regarded as one of the most significant tools [20] because of its remarkable classifier performance in complicated classification situations. One of the most difficult concerns in deep neural networks is their training performance since in optimization tasks, they are tasked with minimizing an objective function with a large number of variables in a multidimensional search space. Finding & training an appropriate DNN method needs a significant amount of time and effort.

4.3 Recurrent Neural Network (RNN)

Essentially, it is a sort of NN in that output from the previous stage is used as input for the next step. However, in typical NNs, all inputs & outputs are completely autonomous of one another. Nonetheless, in other cases, including after predicting the next word of a sentence, preceding words are required, and so preceding words must be recollected. As a consequence, RNN was established, which, with aid of a Hidden Layer, was able to resolve the problem. The hidden state of an RNN is the most significant and most prominent aspect since it is responsible for retaining some data about sequences.

V. MACHINE LEARNING FOR BIG DATA MINING

Early research [21] shows that machine learning has great promise for data analysis. The "search" algorithm of the needed answer, as opposed to the data mining algorithm design for particular issues, is what makes machine learning techniques useful for so many diverse computational and analytical tasks. It is possible to apply most machine learning techniques if

the BDA issue can be framed as an optimization process, and this is the case with a large number of datasets. As an example, a machine learning method known as a genetic algorithm may be utilized to handle both the clustering and frequent pattern mining problems. Aside from tackling mining difficulties in the data analysis operation of KDD, machine learners have the opportunity to improve other elements of KDD, including such dimension reduction for input operators [24]. Standard mining techniques, statistical methodologies, preprocessing solutions, and perhaps even GUIs have been utilized in several representative applications and systems for BDA, according to recent research. The findings demonstrate that machine learning algorithms will play animportant role in BDA in the future. Traditional data mining techniques, such as those used for sequential or centralized computations, have a similar challenge when used for big data analytics. The most likely answer is to make them parallel-capable as well. However, parallel computing versions of genetic algorithms (e.g., population-based techniques) have been proven for many years and may effectively be employed for machine learning [25].

In addition to the conventional GA, the populations of the island modeling evolutionary algorithms, one of the concurrent GAs, can be divided into several subpopulations, that is not the case with both the conventional GA. To put it another way, the sub-populations may be distributed over many threads or computer nodes for parallel processing. Because of this, Kiran and Babu noted in [26] that architecture for distributed DM method still requirements to aggregating input from separate computers. The following is typical distributed DM technique architecture, as illustrated in Fig. 1: a mining method will be run on a computer node that has it's close by consistent data, then not the whole dataset.

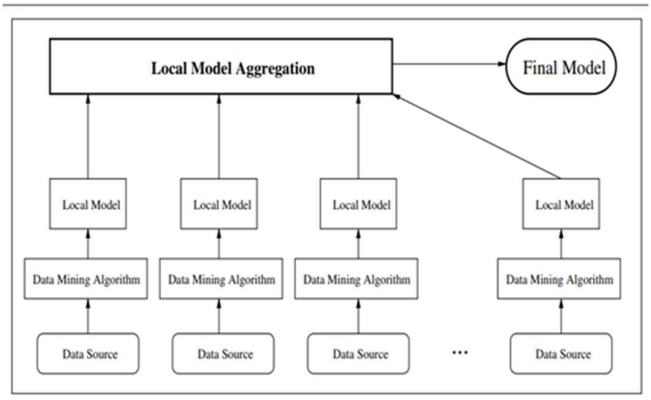

Fig. 1 A simple instance of distributed DM framework

An integrated global model must be built by combining and integrating the individual local models found by the mining algorithms on each computer node to reflect the total knowledge. When adopting a distributed computing system, Kiran and Babu [26] also noted that communications are the obstacle. There are several problems with applying machine learning methods to parallel computing systems, according to Bu et al. [27]. Iteration isn't supported in the first iterations of the map-reduce framework (i.e., recursion). Some recent efforts [28] have taken notice of this issue and attempted to remedy it, which is excellent news. In the same way that classic data mining methods benefit from using CUDA, a GPU, to speed up computation, a machine learning approach may benefit from using CUDA to speed up computation. SOM

with MBP was implemented using CUDA for the classification issue by Hasan et al. [29]. According to the simulation findings, GPU is quicker than CPU. As an example, the speedup achieved by SOM and MPB when implemented in GPUs is three times more than the speedup achieved by CPUs. The ant-based technique was also used for a grid computing platform in another work [30]. Because Deneubourg et al[31] .'s ant clustering technique has been included in the proposed mining technique, To better handle large datasets, Ku-Mahamud tweaked the ant clustering algorithm's ant behavior. In other words, the grid will be filled with ants in random order. As a result, a parallel computing platform may take advantage of the ant clustering process. The developments in ML research for BDA may be separated into two categories: one is an effort to createML systems operate on parallel platforms, including Radoop [32], Mahout [33], & PIMRU [27], while the other is an effort to restructure ML techniques to create them suited for equivalent computing or to parallel virtualized resources, including such NN models to GPU as well as ant-based methodologies for the grid. For big data analytics, both make it feasible to use machine learning methods, but there are still numerous research concerns to be addressed, such as how to communicate across multiple computer nodes and how much processing is required for most machine learning techniques.

VI. STORAGE TOOLS OF BIG DATA

Withthe advancement of computing knowledge, large amounts of data may be handled without the need of a supercomputer and at a low cost. Data may be preserved for transmission through a network. There are several storage management technologies and approaches available. Some of these are described more below:

6.1 Hadoop

Mike Cafarella as well as Doug Cutting initiated an effort to index approximately 1 billion pages for their search engine initiative. Google File System, sometimes called GFS, was launched by Google in 2003. Later in 2004, Google released the Map-Reduce architecture, which became the backbone of the Hadoop platform. In layman's terms, the key components of the Hadoop framework are Mapreduce as well as HDFC. In this part, we will go through the Hadoop component [34].

6.1.1 HDFS(Hadoop Distributed File System)

HDFS is a Java-based storage system that was built to span huge clusters of commodities machines and offers scalability as well as dependable data storage. There are 2 kinds of nodes in a cluster [35]. There is a namenode at the top of the tree that acts as the master node. There are 2kinds of nodes in a network: a master node and a slave node. With a default 64MB block size, HDFS arranges data into blocks. There are multiple copies of these files, allowing vast amounts of data to be processed at the same time.

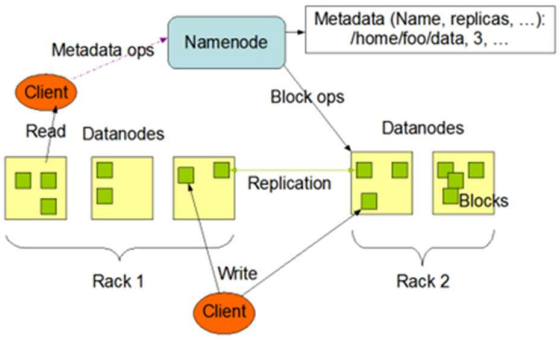

Fig 2. HDFS Architecture

HDFS stores massive amounts of data, & to do so, the files are distributed over numerous workstations. These files are kept in a redundancy way to protect the system from data loss in the case of a breakdown. HDFS also supports parallel processing. File system namespace, that normalizes client access to files, is managed by a single Name-Node, which sits on side of the highest server. A single [36] Data Node (DN) accompanies each cluster in this layout. The storage connected with the nodes on that they run is under the control of this. It is possible to store user data in files using HDFS's file system nomenclature. In HDFS, a file consists of one or more blocks that are linked together. DN is used to store these blocks. Closing or renaming directories and files are just a few of the many things the Name-Node may do for the file system. Maps blocks to DN, as well as other tasks.

6.1.2 MapReduce Frameworks

It is a Java-based distributed computing model known as Map Reduce. It's a way of doing things. To use the Map-Reduce technique, you must complete two steps: Map and Reduce. The phrase refers to 2 independent processes performed by Hadoop applications [37]. Work of this kind includes mapping one set of data to another and then breaking the new piece of data down into smaller units called tuples (key/value pairs). the output of a map, and the data tuples that are aggregated into a smaller set of data, are inputs to the reductions process. The decreasedeffort is always executed after map operation, as the term Map Reduce indicates.

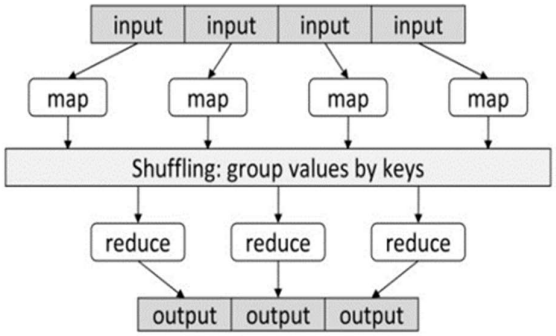

Fig. 3. MapReduce Architecture

VII. ISSUES AND CHALLENGES

Big data challenges may be roughly categorized into 3sets. The first is data difficulties, the second are collection and analysis issues, and also the third is data management challenges. A data problem has to do with a large amount of data. Processing issues are those experienced during data processing, whereas management issues are those faced while managing data, such as securing it. The properties of BD, e.g. its vast volume, diversity, and so on, provide several issues. Data collecting, pre-processing, data analysis, including data visualization are all process difficulties, while safety and confidentiality are management challenges. Figure 4 depicts the many sorts of obstacles related to the various stages of the big data analysis procedure.

Fig. 4. Challenges connected with the data, process, as well as management processes.

7.1. Data Challenges

Numerous descriptions of BD have been presented by investigators, as well as built on their understanding, they have come up with various additional features of BD. [38] Investigators explored 3V's of data characteristics (Volume, Variety, & Volume), [39] IBM proposed the 4th V as veracity, then [40] researchers studied 5th and 6th V's as variability as well as value. When the 10 V's of BD are considered [40], there are several notable issues related to data characteristics that should be mentioned. The following are some of the most important challenges:

7.1.1. Volume Challenges.

A vast quantity of data has been generated as a consequence of the extraordinary growth in data from primary and secondary data sources. This huge amount of data poses issues to data itself, like the inability of standard tools to store data for processing, necessitating the development of more inventive approaches to deal with data flood. [41].

7.1.2. Variety Challenges.

The challenges linked with varietyare due to its many forms. Large datasets may be organized, semi-structured, or unstructured. According to studies, 95 percent of data is in an unstructured format. As a result, putting data into a form that can be analyzed is a significant difficulty.

7.1.3. Velocity Challenges.

Velocity denotes the pace at which data is created by devices. Data processing may be done in 2 different behaviors: batch processing as well as real-time computing. Data is saved and then analyzed in batch processing, while real-time processing is ongoing. The present processing is necessary for online purchasing to offer value for consumers.

7.1.4. Veracity Challenges.

The veracity of data reveals the quality as well as the correctness of the data. It is concerned with data fabrication methods, inaccuracies, messiness, as well as misplaced evidence. Whenever a critical judgment must be made, it defines the credibility of data. User opinions on social networks may be characterized as good, negative, or neutral.

7.1.5. Value Challenges.

The most important property of large data is value. BD gives valuable data that must be pulled from massive databases. This presents a central challenge to information in terms of obtaining high-value information from a database in a cost-effective way and using it for predictive analytics, healthcare, and so on.

VIII. BIG DATA APPLICATIONS

By providing analytics as well as predictive methodologies, big data analytics assists businesses and entrepreneurs in making better-educated business decisions. The following are some examples of big data analytics applications:

8.1. Healthcare: Large volumes of data have been created as a result of electronic health records [42]. The majority of health records comprise quantitative data, qualitative data, including transactional data. By processing organized and unstructured data, big data analytics approaches augment established methodologies. Big data provides an observational foundation for clinical concerns. Big data may aid in the implementation of customized medicine efforts into clinical practice by demonstrating the ability to apply analytical skills. The system biology and health record may be combined in this way. Big data is employed in a variety of applications, including healthcare data solutions, anti-cancer treatment, monitoring patient vitals, hospital administration enhancement, industry growth expansion, fraud identification &anticipationto health insurance companies, and so forth.

8.2 Exploration of educational data as well as learning analytics: As during pandemic time, online [43] schooling has grown in popularity. Students' online actions generate a tremendous quantity of unutilized data. Big data approaches are becoming more important in the learning atmosphere. BD learning analytics approaches may be utilized for forecast problems, detecting attrition risk, data visualization, intelligent evaluation, instructional suggestions, estimating student competence, and detecting behavior, among other things.

8.3Operational security as well as risk management: Organizations[44] have begun to employ BD in proceduresecurity&RM. By giving statistics, big data helps enhance quality research and RM. As a result, management can make timely and required decisions.

8.4 Smart agriculture:BD may be utilized to control the process of precision agriculture [45]. Modern technology connects external big data sources, such as market data and meteorological data, to farmers, aiding in the creation of smart farms. BD is revolutionizing farming by increasing production, forecasting yields, RM, as well as ensuring food security. BD is such a strong tool that it may be used in a variety of sectors. Other uses include governance, social

media analytics, spam detection, contact center analytics, finance, marketing, and telecommunications.

IX. CONCLUSION AND FUTURE WORK

This study aims to provide an overview of BDA as well as its subfields, ML and DL approach. As previously said, BDA was created to solve the complexity of managing data while simultaneously creating and bringing information into businesses to boost efficiency. DNN, RNN, as well as CNN are presented as deep learning approaches in this paper, and classifications, clustering, including evolutionary strategies are discussed. Finally, contemporary research on BD value realization is distinguished by a small no. of experiential investigations & some repackaging of old concepts.

We think that big data analytics is very important in this age of information overload and that it may give unexpected insights plus advantages to decision-makers in a variety of fields. Big data analytics does have the ability to provide a foundation for improvements on the scientific, technical, & humanitarian levels if properly explored and implemented. We as researchers would be able to determine the degree to which big data value fulfills predictions associated with future scientific evidence, both for enterprises intending to effectively gain from big data as well as social benefit of the entire.

References

[1] Bożejko W et al. Parallel tabu search for the cyclic job shop scheduling problem. Computers & Industrial Engineering. 2018;113:512-524

[2] Kiziloz H, Dokeroglu T. A robust and cooperative parallel tabu search algorithm for the maximum vertex weight clique problem. Computers & Industrial Engineering. 2018;118:54-66

[3] Acharya U et al. Automated detection of coronary artery disease using different durations of ECG segments with a convolutional neural network. Knowledge-Based Systems. 2017;132:62-71

[4] Babu GP, Murty M. A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm. Pattern Recognition Letters. 1993;14(10):763-769

[5] Bonyadi MR, Michalewicz Z. Particle swarm optimization for single-objective continuous space problems: A review. Evolutionary Computation. 2017;25(1):1-54

[6] Caliskan A et al. Classification of high resolution hyperspectral remote sensing data using deep neural networks. Engineering Applications of Artificial Intelligence. 2018;67:14-23

[7] Cano A. A survey on graphic processing unit computing for large-scale data mining. WIREs Data Mining and Knowledge Discovery. 2017;8(1):e1232. DOI: 10.1002/ wide.1232

[8] Caraveo C et al. Optimization of fuzzy controller design using a new bee colony algorithm with fuzzy dynamic parameter adaptation. Applied Soft Computing. 2016;43:131-142

[9] Castillo O, Amador-Angulo L. A generalized type-2 fuzzy logic approach for dynamic parameter adaptation in bee colony optimization applied to fuzzy controller design. Information Sciences. 2018;460-461:476-496

[10] Chen J et al. The synergistic effects of IT-enabled resources on organizational capabilities and firmperformance. Information and Management. 2012;49(34):140-152
[11] van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups, Tech. Rep. 2013. [Online]. Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/.
[12] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online]. Available: HTTP:// www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.
[13] Press G. $16.1 billion big data market: 2014 predictions from IDC and IIA, Forbes, Tech. Rep. 2013. [Online]. Available: http://www.forbes.com/sites/ gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.
[14] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: http://www.idc. com/prodserv/FourPillars/bigData/index.jsp.
[15] Taft DK. Big data market to reach $46.34 billion by 2018, EWEEK, Tech. Rep. 2013. [Online]. Available: http://www. eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html.
[16] Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive analytics, ABI Research, Tech. Rep. 2013. [Online]. Available: https://www.abiresearch.com/press/ big-data-spending-to-reach-114-billion-in-2018-loo.
[17] Furrier J. Big data market $50 billion by 2017—HP vertica comes out #1—according to wikibon research, SiliconANGLE, Tech. Rep. 2012. [Online]. Available: http://siliconangle.com/blog/2012/02/15/ big-data-market-15-billion-by-2017-hp-vertica-comes-out-1-according-to-wikibon-research/.
[18] Kelly J, Vellante D, Floyer D. Big data market size and vendor revenues, Wikibon, Tech. Rep. 2014. [Online]. Available: http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues.
[19] Kelly J, Floyer D, Vellante D, Miniman S. Big data vendor revenue and market forecast 2012-2017, Wikibon, Tech. Rep. 2014. [Online]. Available: http://wikibon.org/wiki/v/ Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017.
[20] Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.
[21] Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart. 2012;36(4):1165–88.
[22] Kitchin R. The real-time city? big data and smart urbanism. Geo J. 2014;79(1):1–14.
[23] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.
[24] Han J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann Publishers Inc.; 2005.
[25] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proc ACM SIGMOD Int Conf Manag Data. 1993;22(2):207–16.
[26] kranthi Kiran B, Babu AV. A comparative study of issues in big data clustering algorithm with constraint-based genetic algorithm for associative clustering. Int J Innov Res Comp Commun Eng 2014; 2(8): 5423–5432.

[27] Bu Y, Borkar VR, Carey MJ, Rosen J, Polyzotis N, Condie T, Weimer M, Ramakrishnan R. Scaling datalog for machine learning on big data, CoRR, vol. abs/1203.0160, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/ corr1203.html#abs-1203-0160.

[28] Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, Czajkowski G. Pregel: A system for large-scale graph processing. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2010. pp 135–146

[29] Hasan S, Shamsuddin S, Lopes N. Soft computing methods for big data problems. In: Proceedings of the Symposium on GPU Computing and Applications, 2013. pp 235–247.

[30] Ku-Mahamud KR. Big data clustering using grid computing and ant-based algorithm. In: Proceedings of the International Conference on Computing and Informatics, 2013. pp 6–14.

[31] Deneubourg JL, Goss S, Franks N, Sendova-Franks A, Detrain C, Chrétien L. The dynamics of collective sorting robot-like ants and ant-like robots. In: Proceedings of the International Conference on Simulation of Adaptive Behavior on From Animals to Animats, 1990. pp 356–363

[32] Radoop [Online]. https://rapidminer.com/products/radoop/. Accessed 2 Feb 2015

[33] Apache Mahout, February 2, 2015. [Online]. Available: http://mahout.apache.org/.

[34] Abbass H, Newton C, Sarker R. Data mining: a heuristic approach. Hershey: IGI Global; 2002

[35] MapReduce,https://hadoop.apache.org/docs/r1.2.1 /hdfs design.html

[36] Aaltonen, A., Tempini, N., 2014. Everything counts in large amounts: a critical realist case study on data-based production. J. Inform. Technol. 29 (1), 97–110. http://dx.doi.org/10.1057/jit.2013.

[37] Abbasi, A., Sarker, S., Chiang, R.H.K., 2016. Big data research in information systems: toward an inclusive research agenda. J. Assoc. Inform. Syst. 17 (2), i– xxxii.

[38] Huang J et al. A clustering method based on an extreme learning machine. Neurocomputing. 2018;227:108-119

[39] Shah T, Rabhi F and Ray P 2015 IInvestigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health condi Cluster Comput. 18 351–67

[40] Schroeck M, Shockley R, Smart J, Romero Morales D, and Tufano P Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report IBM Inst. Bus. Value Said Bus. Sch. Univ. Oxford

[41] Khan N, Alsaqer M, Shah H, Badsha G, Abbasi A and Salehian S 2018 The 10 Vs, issues and challenges of big data Proceedings of the 2018 International Conference on Big Data and Education (New York, NY, USA: Association for Computing Machinery) pp 52–6

[42] Murdoch T B and Detsky A S 2013 The inevitable application of big data to health care J. Am. Med. Assoc. 309 1351–2

[43] Sin K and Muthu L 2015 "Application of big data in educational data mining and learning analytics – a literature review " ICTACT J. Soft Comput. 05 1035–49

[44] Hariri R H, Fredericks E M and Bowers KM 2019 Uncertainty in big data analytics: survey, opportunities, and challenges J. Big Data 6 44

[45] Wolfert S, Ge L, Verdouw C and Bogaardt M-J 2017 Big Data in smart farming – A review Agric. Syst. 153 69–80