Journal of Big Data

METHODOLOGY Open Access

A systematic review on big data applications and scope for industrial processing and healthcare sectors

Kumar Rahul[1*], Rohitash Kumar Banyal[2] and Neeraj Arora[3]

*Correspondence:
Kumar Rahul
kumarrahul.niftem@gmail.com
[1]Department of Basic and Applied Science, NIFTEM, Sonipat 131028, India
[2]Department of Computer Science and Engineering, Rajasthan Technical University, Kota 324010, India
[3]School of Science and Technology, Vardhman Mahaveer Open University, Kota 324010, India

Abstract

Nowadays, big data is an emerging area of computer science. Data are generated through different sources such as social media, e-commerce, blogs, banking, healthcare, transactions, apps, websites, opinion platforms, etc. It is processed for effective utilization in different industries, including healthcare. These enormous generated data are essential for data analysis and processing for industrial needs. This paper reviews the work of various authors who have contributed to data collection, analyzing, processing, and viewing to explore the importance and possibilities of big data in industrial processing applications and healthcare sectors. It identifies different opportunities and challenges (data cleaning, missing values, and outlier analysis) along with applications and features of big data. This systematic review further proposed dirty data detection and cleaning and outlier detection models that can be used for many applications. The data cleaning and outlier detection models use the optimizations concept to solve the optimal centroid selection problem and suspected data.

Keywords Big data, Data cleaning, Outliers, Industrial processing, Healthcare

Introduction

Various authors define big data in multiple ways, but there is no adequate definition. However, big data is expressed in terms of V's (volume, velocity, and variety); but it includes many v's such as veracities, vagueness variability, vulnerability, volatility visualization, etc. Big data application areas include business organization, operation, production, marketing, information technology management, etc. All these sectors need extensive data management and processing at various levels required to be classified and filtered for every individual in the industry for effective utilization. Different algorithms used for multiple applications affect the future of data science. Big data management is described through the life cycle of study, data collection, documentation, integration, preparation, data analysis, publishing and sharing, data storage, and data reuse.

An industry including media, entertainment & communication uses big data for business benefits. Industries use and analyze customers' behavioral data to target audiences

and recommend products and services. Big data analytics is used to solve complex problems [1]. Healthcare big data presentations from 2015 to 2019 include different techniques such as statistical data analytical techniques, hidden Markova model, and machine learning techniques for bioinformatics given in [2]. Nowadays, healthcare data breaches appear through different medical field devices, becoming one of the industrial domain victims [3]. Healthcare information systems used effectively by various hospitals simultaneously need to implement meaningful information efficiently, and filtered data helps provide better medical services [4].

Big data improves industrial profit margin, production, and operations [5]. Big data plays a vital role in many applications and engineering domains, artificial intelligence-based, data analytics-based, etc. Grand View Research, Inc., U.S.A, Big Data will reach USD 72.38 billion by 2022 [6]. Healthcare big data management is essential for evaluating diagnosis through large datasets; thus, big data has a high role in the medical information system [2]. Some factors affect verintegrityuch as data provenance, uncertainty, and dirty and noisy data [7]. These algorithms are machine learning, deep learning, and natural language generations (NLG), directly or indirectly affecting applications such as healthcare, smart city, industry 4.0, etc. Deep learning is used for pattern designing to improve optimization and computations [8]. Data science benefits job hunting, product customization, cost optimization, etc.

The big data evaluation hierarchy moves like in the 1970s, statistical computing, 1980's massive data sets, 1990's data mining with statistical learning, 2000's business analytics, and 2010's big data analytics explored in [9]. However, firms are ready to implement big data analytics technologies that have been discussed, along with some challenges and preparedness in [10]. The system health monitoring and management (SHMM) concept is introduced, which contains active and passive data for forecasting, prediction, diagnosis, etc., in the healthcare system [9]. IoT and Cloud computing for healthcare are discussed in [11]. Similarly, big data applications have various benefits in the healthcare sector. Big Data technologies are supporting healthcare sectors and services in different ways. It covers advanced patient care, improved operational efficiency, disease detection, cost reduction, precise treatment, improved medical diagnosis, etc. Data quality, operations, and productivities are essential in the healthcare system.

Healthcare and other industries use multimedia techniques to enhance efficiency, coordination, and centric system to support end-users [12]. Healthcare analytics' objective includes patient safety, real-time monitoring, clinical decision support system, enhanced services, etc. [13]. Big data and its technological implementation in the industrial sector are essential. It can be understood through the value chain: data acquisition, analysis, curation, storage, and usage [14]. Data is used for various industrial operations obtained from different sources such as the internet, industrial incorporation, enterprise resource planning (ERP), customer relationship management (CRM), human resource management (HRM) software modules, the social media network, transactions, healthcare, geographical data, remote sensing, audio-video recording, etc. The remote sensing (RS) big data definition and satellite with volume and velocity are given in [15]. Big data management is required for scalable data management to support large applications [16]. Big data provides many emerging scientific research and value approaches to evaluate economic growth [17]. In e-commerce, most information and data are electronically collected, which may have different types of data evaluation and quality problems [18].

Big data is suitable for integrating other healing sources to form the life cycle of accessible health data [19]. The big data life cycle is executed as data generation, acquisition, storage, analytics, and visualization [20, 21]. The data acquisition system architecture for the agriculture domain is explained in [22]. In big data, data sources affect accuracy, inconsistencies, missing value, data cleaning approach, duplicity among datasets, and usability in many applications, including business analytics and the healthcare sector. Industrial processing identifies data analytics, business operation tools, report generations, etc., depending on big data's structure or unstructured form.

Big data for manufacturing industries can enhance the manufacturing process, customize the product development process, improve quality assurance, improve supply chain risk, etc. The authors have explained manufacturing processes' data analysis, which helps in-process monitoring and fault detection [23]. Users get data through the assistance system (process monitoring, anomaly detection, fault analysis) [23]. The different applications include chemical industries, complex processes, process control of harvesters, etc. [23].Big data can improve the performance of aging manufacturing by dealing with fault detections [24]. It is carried out in two phases, fault identification, and detection [24]. Big data exists in multiple industrial domain areas, such as production, quality evaluation, supply chain, etc., affecting operating costs within an organization [25].

The authors have used remote agents, intelligent mapping, and cleaning processes for automatic data-driven industrial data analysis for different operations comprising distributed intelligence, performance, predictive, and preventive procedures [25]. IoT-enabled industries are data-generating sources to solve featured engineering problems through deep learning mechanisms [26]. IoT technologies are used in different sectors. Agricultural IoT identifies data interoperability with the actual and predicted data that improve crop management's decision support system [22]. IoT agricultural processing system collects and transfers data through a server to an artificial neural network (ANN) for prediction in the agricultural domain. Big data applications exist in agricultural sectors where data acquisition is carried out through IoT-based technologies [22].

This paper focuses on the problems encountered while accessing big data resources. However, big data have many applications in healthcare, transportation, education, smart city, etc. Industries need actual data for their processing task, operations, and productions. When big data is used in various sectors, all sectors encounter a common problem, i.e., "Data Filtering." Moreover, factual information and knowledge for industrial operations cannot be easily achieved due to the large data set, and neither can it be merely implicit nor automatically extracted [27]. Differences between data production and data extraction are defined in [28]. Data production (a medical record), data reuse (coding of medical data), and data analytics are defined as an order in big data importance for healthcare industries [29]. Healthcare information system (HIS) provides information to the end-user through the execution of cloud computing, Electronic health record (EHR), security layer, big data analysis, and information, as mentioned in [30]. Thus, big data analytics is required to process in such a way to benefit industries for the following reasons: understanding the market, less time, product development, online feedback, decision-making process, etc.

The summary of the paper is described in different sections.

- Section 2 explained big data: preliminary,
- Section 3 focused on a review of big data analysis,
- Section 4 identified different big data applications,
- Section 5 explains big data in business organizations,
- Section 6 elaborated on the challenges and opportunities in big data and.
- Section 7 explores future research outlines and.
- Section 8 conclusions.

The outcome of this systematic review work is explained in the following forms:

- Design an optimization-based data cleaning model for dirty data detection and removal: The data cleaning model will detect and remove it from the system. This review paper identifies data cleaning as one of the major challenges in large application systems through studies based on [7, 19, 23, 27, 31–35]. Dirty data detection and removal can be performed through the optimization concept or nature-inspired algorithm (NIA). The detecting and corr Detecting and correcting include applications' inaccurate, corrupt, and irrelevant ratification and removal: The second most promising challenge found in terms of identifying missing value from work based on [7, 24, 27, 36, 37]. Data analysis encompasses preprocessing of data, feature extraction, reduction of features, clustering, filtering, and classifications.
- Design an outlier removal model: outliers differ from the existing standard data, deviating from the available dataset. However, several definitions of outliers exist, but no adequate explanation for removing the outliers in an extensive database system through clustering (An unsupervised model) could be the best outcome of this review. Outliers are an observation that appears different from a cluster's standard data. It affects the results and calculations.

Big data: preliminary

Big data is a collection and combination of very large data in different formats that are difficult to store, understand, and process for an application. Big data falls under the categories of data analysis and streaming. Big data defines with the help of V's, such as volume, variety, velocity, variety, veracity, vagueness, vocabulary, variability, venue, value, etc.

Volume

Big data generated through various sources, including healthcare, transportation, power grid, intelligent education energy, etc., must be classified as structured, semi-structured, and unstructured [38]. It represents the scale of data that originated from various sources. The data's size, amount, and volume are heterogeneous, and data volume is increasing exponentially. There is a 40% growth in total data generated every year versus a 5% only growth in IT expenditure. Around 90% of the world's digitized data was captured in the last few years, and it keeps going on [38]. Some challenges include the curse of modularity and dimensionality with a large volume of data in industrial applications where machine learning (ML) techniques are used to optimize and predict the outcome [7]. Healthcare data increase dramatically as per [39]. Machine learning is used to

evaluate data modeling for large-scale data generation access points of an application to minimize computational cost and less memory and make a valuable model for predicting accuracy [40]. The authors have explained big data for healthcare in parallel computing for volume, learning incrementally for velocity, and information fusion for a variety [41]. Its emphasizes various applications' input, analysis, and output [41].

Variety

It represents the different data types generated from various sources such as relational database systems, web text data, structured and semi-structured data, unstructured data, online data, social media data, healthcare sectors, etc. Processing and filtering data when required during processing in any application is complex. Data features, such as completeness and timeliness, are essential during accessibility in any system or application. For instance, knowing the weather condition of a particular city, the complete data set required, and timeliness reflects data availability for a specific duration. Variety is an attribute of big data associated with some challenges for industrial applications. Industrial applications are enabled with machine learning (ML) as a tool enclosed with some challenges: data locality, data heterogeneity (statistical, syntactic, or semantic), and dirty and noisy data. These data include different errors, different outliers, and missing values as well. However, data cleaning and outliers removal are critical challenges under big data analysis [7]. Various data also includes textual, images, audio, video, XML, JSON, and sensors data.

Velocity

It represents the speed at which data is generated, processed, analyzed, and stored at a distinct location in a big data system. Data are generated fast and need to be processed quickly as well. In an e-commerce business, promotions should be offered immediately to attract customers based on users' transaction details. Velocity also represents the rate at which it is generated and analyzed. Some essential factors affect industrial processing tasks through machine learning, including data availability, real-time processing or streaming, and random variable generation [7]. It is also defined as an "increasing data rate" within an organization. The velocity of data is essential for batch processing within industrial processing and other applications. Data streaming helps in machine learning-based processing in telecommunications and information technology-based industries. Many industrial applications require real-time data and processing task/unit update machine learning-based software that receives data stream as in moreover, many V's, such as visualization, veracity, value, virality, viscosity, and so on, concern thconcernsvalues contribution to big data analytics being changed from descriptive analysis to predictive analysis. Various data analytics categories define through text, visual, voice, network, and geospatial [1]. Textual case studies include social media, academic papers, website reviews, company documents, etc., whereas visual analytics includes surveillance systems, CCTV, drone trajectory heterogeneous cameras, etc. [1]. Big data provides designing analysis and execution management, one of the services under SaaS on the cloud [42]. The importance of many v's can be understood as:

Table 1 Importance of V's in Big data

Volume	Variety	Velocity	More v's
Megabytes	Structured Data : Tables	Batch	Visualization
Gigabytes	Semi-Structured Data : Records	Real-Time	Value
Terabytes	Un-Structured Data: Records	Stream	Veracity
	Partial data : Transactions	Near Real	Virality
	Spatial data: Transactions	Time and so on	Viscosity and so on
	Sensor Data: Transactions		
	Image and so on : Files, etc.-Records		

More v's

Table 1 gives an approach of different V's and its existing format. Value is the essential characteristic of big data, which provides meaningful and heterogeneous advantages to industries. Any amount is highly dependent on the execution of processes and data access. For example, a weather forecast requires any random data that would be enough to predict a forecast for a particular region. In contrast, continuous data is required for disease treatment, such as temperature measurement, etc. Similarly, customers' data are essential in banking services before proceeding with a loan [43]. Veracity refers to the quality of data used in the applications and systems. Since data used under any applications and system must be checked and analyzed. Veracity in big data (context, cross-validations) discussed in [44]. The authors stated data science and big data are equally important for healthcare [44]. Veracity tends to data processing strategies. Veracity reflects the trustworthiness of data, estimated in terms of accuracy.

Since noise is considered data, but no information can be extracted from it. Viscosity is another V that measures resistance to the volume of data; it improves the data streaming. Similarly, variability is defined as data differentiation in nature where different data sources exist, differentiated between important and noisy and faulty data. Visualization is an approach by which data relationships are shown in some form. Data visualization is a mechanism that suits presenting a few data sets (in the form of tabular, graphical, and circular) to deliver required and meaningful results with the support of tools and techniques.

Visualization supports through different stages (data identifications, acquisition, analysis, validation, and visualization). Data acquisition sources take place from multiple sensors [45]. Data visualization can be more effective in different ways. For instance, data is compelling and more credible, correct information is not overcomplicated, has graphical representations and style, is colorable e, has a visual hierarchy, and focuses on the point during presentations. Data visualization makes data more valuable. Virality defines the speed which iinteractionactionss among the network. It amounts to the dataset's many availabilities for data processing before its implementation in the applications. It also defines applying data sources from among different applications. The authors describe the importance and significance of data generations and utilizations of data in other dimensions, including service sectors, as mentioned in [46]. It also reflects how big data is suitable for the applications like banking and securities, communications, media and services, education, government, healthcare providers, insurance, manufacturing, natural resources, retail, transpordutiesn, ties wholesale trades, etc. [46]. Besides definitions of big data, many technologies, like machine learning algorithms, are useful for data analysis. It can be used to find a feasible solution for optimized and complex problems where data analysis problems can also be converted into optimization solutions [47].

From the data mining perspective, noisy, outliers, and incomplete datasets are essential in big data analytics in an application. In diabetes, the machine learning algorithm (MLA) is used along with data mining technologies [48]. However, in section IV, various applications of big data are described.

Few sectors, such as Banking (25%), Services (15%), and Manufacturing (15%,) are the three most active industries in making inquiries about big data to Gartner over the last twelve months. The adoption of big data in different sectors reached up to 53% [49]. IoT and big data projects have increased as industries focus on data for decision-making systems for positive impact and business improvements [50]. Crop yield improvement is another big data application area where production prediction enhances the agricultural field [51]. Big data applications, features, and challenges of different industries are mentioned in [52]. Big data is used effectively to improve smart farming and reduce wastage [53]. Big data analysis is also helpful in improving the food supply network within agricultural-based industries. Similarly, the authors mention the importance of data mining, data analytics, cloud computing, machine learning, etc., in the big data paper [54].

Systematic big data analysis review

Research articles have been selected from IEEE Access, Elsevier, ScienceDirect, Springer, Procedia, Scopus, SCI, SCIE, SSCI journal of repute, etc. Big data applications, resources, links, and availability have been taken from Gartner, fact sheet telehealth, etc. The author has identified the diversified nature of big data, so the focus of is the literature review is to find out the answer to the following questions:

1. What is the need and importance of big data applications?
2. How is big data analytics suitable for industrial processing and the healthcare sector?
3. The application domain's data life cycle identifies a verified problem like data filtering, outlier removal, and missing value.

The big data applications have been explained in Sect. 4. A comprehensive literature review was performed, keeping these above points in mind. In 2010, Apache HADOOP defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." Based on this, in May 2011, McKinsey & Company announced big data as the next frontier for innovation and productivity [18]. Big data is suitable for designing models, tools, and methods for different functions, including operational and clinical activities for the healthcare sector [19]. This big data analysis review comprises other research papers, including business processing and health. The territory of big data is a vast and diverse one. Since big data has high scope in different sectors, big data analysis and technology reviews concerning a particular domain become important. It has been observed that automation generates data duplication and data entry errors in the system. There is a need to identify common challenges among sectors or applications that can be further researched and improve accessibility in different applications. Moreover, it is essential to know how big data support various sectors in today's scenario, irrespective of challenges and issues in the usability of tools and techniques.

Since industrial processing task (including the manufacturing sector) encompasses through massive data set for quality evaluation of product and also used for future

prediction of outcome for identifying the nature of the customer, prediction of selling product to end-user, price of the product, the end-user habit of purchase, etc. Since data sizes are growing fast; thus, the manufacturing process becomes complicated and complex to execute practically [23]. Big data is suitable for process optimization and helps empower the decision process. Before giving the importance of any mechanism, we need to review the work done in this area with methodologies, features, and challenges. The research methodologies of selected research papers are based on the search mechanism "Big Data," Big Data in Healthcare," "Big Data in Industries," and so on. Thus, Table 2 shows the details of the work done by the various authors.

So, the above articles show various methodologies and features used in different domains and applications. From the above research papers [7, 19, 23, 27, 31–35], the challenges found in different terms. It includes data cleaning, missing value, and outlier detection under various applications. It affects in the following ways: interaction among users, sharing of data through a common platform, high accessibility among users, social support, health-related clinical information, and health policy. Big data research projects need a search engine for healthcare, financial transactions, weather data., index structure, and search mechanism [69]. In [19], the authors expressed data staging errors generated during integration, transformation, and migration. Since data generated for analytics must go through the data cleaning process, several errors such as whitespace removal and no. of zero's for identified numbers are addressed.

Different simulations environment were used, including parallel computing, modeling, metaheuristic approaches, artificial intelligence environment, knowledge handling, data analysis, reporting, machine learning, deep learning, principle component analysis, outlier removal, cleaning model, and random forest machine learning algorithm, etc. in the articles mentioned above. Data cleaning models and outlier analysis are the most important simulation methods of various applications. Different experimental parameter used such as accuracy, sensitivity, specificity, FPR, FNR, negative NPV, and FDR were used to compare with the existing models [7, 19, 23, 27, 31–35].

More in-depth machine learning-based experimental analysis was executed to resolve processing performance, where dependency reduces. Deep learning improves classification performance, design pattern formulation, and evaluation models [8]. It includes several steps, including preprocessing, feature selection, and text classification, and is evaluated and compared with global feature selection methods [8]. Data analytics simulation is used for healthcare data. Automation is required for data storage, processing, and handling [19]. Data analysis of manufacturing sector-based applications such as agricultural harvest and sorting plants have been discussed through data acquisition, distance-based, and regression-based approaches [23]. Data acquisition tools are used for experimental analysis for different sectors. Data preprocessing is essential to execute data mining tasks to provide meaningful information. Data preprocessing tasks include cleaning, normalization, transformation, missing value imputation, data integration, noise identification, etc. [27]. Because of the large volume of data and different data aggregators, it becomes difficult to maintain the data stream. Therefore, big data reduction methods are explained through network theory, compression, redundancy elimination, data preprocessing, dimension reduction, data mining, and machine learning as experimental approaches and methodologies to reduce and filter use valid [31]. While

Table 2 Review of Big data analytics in the industrial and healthcare sector

Authors	Methodologies	Features	Challenges
• A. L. Heureux and G. S. Member [7]	• Machine learning (ML) mechanism for Big data • Data analytics stages • Data manipulation techniques PCA, dimensionality reduction	• Manipulation for Big data • Processing manipulation • Data manipulation • Algorithm manipulation • Suitable for decision making	• Processing performances in a large volume of data • Dirty and noisy data in the varied nature of Big data • Real-time processing in velocity/speedy data generation • Data uncertainty in case of veracity behavior of data
• S. R. Sukumar,R. Natarajan, and R. K. Ferrell [19]	• Automation of data processing technologies • Healthcare analytical methods	• Sources of errors discussed • Data quality assurance	• Data quality issues • Automation in data handling, data processing, and data storage • Data quality rule engines • Customized software for data quality evaluation
• García et al. [27]	• Data preprocessing techniques • TF-IDF (Term Frequency-Inverse Document Frequency) • Discretization and Normalization	• The connection between Big data and data processing • Big data framework	• New technologies • Scaling data preprocessing techniques (missing value imputation, noise treatment) • Big data learning paradigm (semi-supervised, data stream, real-time processing)
• M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani [43]	• IoT Big data analytics • IoT streaming data analytics • Deep learning (DL) techniques for IoT data analytics	• IoT applications, DL approach • IoT characteristics • Summary of the DL model • Framework for designing deep learning (DL)	• Lack of precise deep learning (DL) method • Training data overload • Specific hardware required for a defined system
• T. Steckel et al., [23]	• Data acquisition for different industries • Anomaly detection (PCA-based, distance-based approach) • Regression-based anomaly detection • Outliers • Self-organizing map	• Application cases for the chemical industry, process control for an agricultural harvester • Failure detection • Anomalies detection • Optimization process	• Data acquisition has a problem in- • Data integration • Heterogeneous manufacturing process • Time synchronization
• P. Matta and A. Tayal [55]	• AHP (Analytical hierarchy process) and PCA (principal component analysis) based methodology • Correlation analysis • Clustering	• Big data for supplier selection problems in industries • Supplier evaluation for a manufacturing firm	• Lack of optimization model for industries • Highly un-structured • Time-consuming
• S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, and S. J. Childe [56]	• Big data analytical capability model (BDAC) • BDA talent capability (BDATLC) • BDA technology capability (BDATEC) • BDA management capability (BDAMAC) • Resource-Based Theory (RBT)	• BDAC-FPER (firm performance) relationship • BDAC and its three primary dimensions (technology, management, and talent capability) and 11 sub-dimensions • Data collection	• Lack of business process agility • Process-oriented dynamic capabilities • Analytics climate • Analytics privacy
• Fernández, S. del Río, N. V. Chawla, and F. Herrera, [57]	• Data preprocessing • Cost-sensitive learning • Big data classification using MapReduce	• Standard preprocessing techniques • Analysis of preprocessing techniques	• Imbalanced classification in big data problem • Design of novel algorithm for a different level of the partitioning of classification • Imbalance ratio between classes

Table 2 (continued)

Authors	Methodologies	Features	Challenges
• A. Waldherr, D. Maier, P. Miltner, and E. Günther [58]	• Filtering strategies • Classifying documents with a machine-learning algorithm • Extraction of the core network	• Web discourse in the era of Big data • Crawled webpage of USA and German	• Cleaning and reducing data during online discourses • Noise problem
• Giovanni Azzone [59]	• Completeness, • Timeliness, • Personalized policies, • Efficiency and effectiveness	• Public policies	• Data accessing and arithmetic computing procedures
• M. Habib, C. Sun, and L. Assad [31]	• Data generation and acquisition • Relationship between Cloud Computing and Big data • Relationship between IoT and Big data • Datacenter • Relationship between HADOOP and Big data.	• Big data storage, • Big data analysis and • Big data applications	• Data representations, • Redundancy reduction • Data compression, • Data life cycle management, • Analytical mechanism • Data confidentiality • Energy management • Expendability and scalability etc.
• X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang [60]	• Rule-based data cleaning technique • Data cleaning from a statistical perspective • Missing values	• Error detection • Error repairing • Business intelligence • Automation with tools	• Scalability • User engagement • Semi-structured and unstructured data • New applications for streaming data • Privacy and security concerns
• V. N. Gudivada, A. Apon, and J. Ding [32]	• Data quality life cycle • Data quality analytics • TIA Process (Transformation, Integration, Aggregation)	• Nature of data quality issues in the context of Big data • Data governance-driven framework • Data quality dimension	• Implementation of data quality lifecycle framework • A new algorithm is required to identify the original data element and source.
• X. Deng, P. Jiang, X. Peng, and C. Mi [33]	• Support tensor data description • Standard support vector data description (SSVDD) • Kernel support tensor data description (KSTDD) • Outlier detection algorithm	• Reduce high dimensional data	• It dealt with only tensor data directly.
• D. Guan et al. [61]	• Novel noise filtering mechanism called Enhanced soft majority voting by exploiting unlabeled data (ESMVU) • Multiple soft majority voting methods (MSMV)	• Effective use of unlabeled data • Improve noise filtering performance. • Noise handling • Worked for mislabeled data filtering	• Noise correction & comparison concerning Big data and its heterogeneous type
• D. Henry [62]	• Data cleaning methodology proposed on hashtags context (time, artificial and recent context)	• More general data cleaning tasks and preprocessing • Suitable for parallel computing	• It is required for text mining tasks such as text classification, sentiment analysis, opinion mining, or text clustering • It required work on a large no. of tweets.
• K. Kenda and D. Mladenić [63]	• Data cleaning algorithm • Kalman Filter • Streaming sensors data platform with data cleaning	• Meta classification method of prediction • Lower noise ratio	• Improvement required of Kalman filter parameter fine-tuning procedure, • Cleaning behavior • Usability of the algorithm • Fail to deal with a large number of sensors data

Table 2 (continued)

Authors	Methodologies	Features	Challenges
• C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, [64]	• Big data medicine • Big data in healthcare • EHR (Electronic health record)	• Data collection through the monitoring system • Clinical documentation	• Data aggregation • Unstructured data analyzing • Priority utilization of data • Data protection
• M. Yang, M. Kiang, and W. Shang [65]	• Automated adverse drug reaction (ADR) related posts filtering mechanism • Supervised classification approach	• Framework for tackling the problem of filtering big data from social media in general and • Consumer adverse drug reaction (ADR) messages identification in a specific application.	• Not suitable for unsupervised data • consumer ADR • Related messages are usually sparse and highly distributed • Reduction of high dimensionality required
• H. Asri, H., H. Al Moatassime, and T. Noel [66]	• Survey paper • Different product details including MCOT, HRS-I • e-HPA • ELCR • Realty mining	• Healthcare and big data • Realty mining and healthcare • Big data and realty mining • Impact of big data analytics in the healthcare industry (right living, proper care, right provider, promising innovation, the correct value, etc.)	• The Source of data acquisition is not synchronized • Data quality is an issue that is in the form of unstructured, nonstandard, improper • Lack of data scientists, resource availabilities, data analytics tools, • Constraints in data accessibility
• J. Wang, W. Zhang, Y. Shi, S. Duan, and J. Liu [67]	• Industrial data ingestion-integration • Repository • Data management • Industrial data analysis • Industrial data governance	• Highly distributed data source (large-scale devices data) • Production life cycle data • Business operation data • Manufacturing value chain • Collaboration data	• Production efficiency • Production quality • Minimize energy consumption • Cost minimization
• Y. Hu, K. Duan, Y. Zhang, M. S. Hossain, S. M. Mizanur Rahman, and A. Alelaiwi [68]	• Simultaneously Aided Diagnosis Mode (SADM) framework • Data preprocessing (data extraction, data cleaning, eliminating redundancy) • Machine learning algorithm (SVM)	• Focused on a disease like heart, diabetes, and cancer database of healthcare • Performance measurement with accuracy, precision, recall, and F1-measure	• Diagnosis efficiency improvement required • Deep learning (DL) required for diseases risk assessment
• Kaur, Pavleen, Kumar, Ravinder Kumar, Munish [34]	• IoT-based disease predictive system for heart, diabetes, and breast cancer patients • Random forest machine learning algorithm (RFML) technology used	• Dataset used of • heart, breast cancer, diabetes, thyroid, liver disorder, etc. • Results compared with k-NN, Linear SVM, Decision tree, MLP, random forest	• Accuracy can be increased further on an extensive database. • Data security is a big concern in IoT-based system • It can be applied to other applications like weather, forecasting, etc.
• S. Oueida, M. Aloqaily, and S. Ionescu [35]	• Maximum Reward Algorithm (MRA) - An optimization-based algorithm	• Enhances healthcare resources • Multimedia technologies are a booster for healthcare services • It improves efficiency and reliability from 50.1–77.2%	• Integration of multimedia technologies with mobile health care services and facilities is complex in some context • The heterogeneous network exists for multimedia technologies

dealing with large and voluminous data, data quality is enhanced by removing and identifying missing values, duplicate values, data heterogeneity, and data integration.

These are essential for big data and machine learning technologies [32]. Similarly, when dealing with data quality and cleaning, outlier removal becomes important in

Table 3 Enterprise software. (Source: (Gartner Oct 2012))

Year	Enterprise software spending for specified sub-markets	Forecast: Social Media Revenue, Worldwide, 2011–2016	Big Data IT Services Spending	Total
2011	2,565	76	24,407	27,047
2012	2,918	1,384	23,476	27,778
2013	3,516	1,812	28,578	33,906
2014	4,240	2,827	37,404	44,472
2015	5,207	3,615	36,189	45,010
2016	6,461	4,411	43,713	54,586

sensor-based data. For outlier detection, support high-order tensor data description proposed in highly dimension sensor data [33]. Since machine learning experimental analysis is imporessentialarious application incl, including healthcare, it is being applied to different test cases such as diabetics, heart diseases, and breast cancer. The experimental result compared with K-NN, support vector machine, decision tree, etc. [34]. So, the data cleaning and outlier removal concept from big data was identified as the baseline for the subjective technological comparison mentioned above 2. In [43], the authors designed a deep learning framework structure, discussed the importance of IoT big data, and found that IoT generates a different type of massive data. It accesses and processes deep learning models resulting in the abstraction and enhancement of services and describes the integration of mobile data, which has been generated at a large scale, is required. The authors expressed various challenges and opportunities in big data healthcare, describing human behavior, and data mining [39].

The authors showed how much statistics are essential for big data.It showed how big data relates to statistics and concluded that it could optimize dynamic decision problems, re-routing vehicles, live traffic management, skipping stops (if permitted), or allocating platforms [70]. Big data encompasses data generation, acquisition, storage, and access. There are processes to generate data for all these phases, acquire data at a few levels, and store it. In [71], the authors showed integration and customization of existing data and the knowledge of different experts to use big data for smart cities as one application area, which will be discussed in the next section. The authors have developed a framework of data and knowledge for smart cities drawing from the application-oriented perspective. Now industries are interested in big data's high potential and applications. In contrast, government agencies also announced launching big data-related projects for societies' welfare [71]. The authors addressed big data problems from a distributed perspective. They studied several significant data challenges, mainly on the machine learning (ML) algorithm following the MapReduce programming model, where fusion is the core of the system [72]. Development of distributed analytics models for big data, where filtered data should be given to MapReduce programming. According to [73], the authors discussed the role of statistics regarding some issues (heterogeneous sources, sub-population clusters, representation, significance, etc.) raised by big data in the new paradigm. It shows how big data relates to different areas of knowledge. Hence, the authors believe in providing high-quality and complete solutions to such issues. Experts from other regions of multidisciplinary teams will also be necessary to identify data learning as important for evaluating high quality.

Big data infrastructure supports data collection, data storage and transfer, and computational technologies (Parallel computing, Hadoop distributed file system (HDFS),

MapReduce, etc.) [74]. The Hadoop tools are also required as a primary platform for cycle time forecasting under production planning and data preprocessing. So cycle time forecasting goes through essential platform establishment (HADOOP) and data processing (including data extracting, formatting, and data cleaning) [75]. It is compared with the traditional database system under the enormous and increasing size of big data. According to Gartner's report (October 2012), total IT spending is driven by big data given below.

Table 3 represents yearwise growth in enterprise software and big data services spending in industries. Big data took 10.20% of CAGR from 2011 to 2016, as per the data given in the Gartner report in October 2012. The authors defined the internal structure of methodologies (HADOOP, Spark, Flink, etc.) proposed to conquer the imbalanced data problem in big data [36]. It discussed the imbalanced classification in the big data scenario—first, a detailed design of artificial data generation techniques to improve preprocessing approaches' behavior. Secondly, study the potential related to the fusion of models or the management of an ensemble system concerning the final reduced task. In 2018, the authors developed a framework of socio-economic sources that can process, integrate, and analyze the data from different sources to forecast economic and social changes. The authors showed a framework for a novel big data architecture system that accounts for the particularities of the digital era's good goods behavior analyses. It is proposed to implement big data architecture using internet data [76]. In continuation, the authors discussed filtering techniques for big data, outlier detection, signal extraction, and decomposition techniques related to big data [77]. It studied seasonal patterns and signal extraction and treatment of outliers and s, seasonal, and design required in the future. Given pros and cons, big data technologies are applicable and valuable in frequent pattern mining in horizontal and vertical data representationscons [78]. However, in 2017, the authors discussed association in GIS-based big data, geometrical associations in space and time, spatiotemporal correlations in statistics, and space-time relations in semantics [79]. In addition to the above, according to Gartner Reports, the 2021 Worldwide IT Spending Forecast (Millions of U.S. Dollars) are given in Table 4 [80]. This gives an impression of growth of data center and enterprises software etc.

Big data applications

Big data technologies include smart cities, hospitals, education, transport, social media, business operations, promotions, event organizations, e-commerce, etc. These sectors are the one that generates lots of data sources for big data to process and store. However, privacy and security are also more significant concerns under big data [81]. Different

Table 4 Worldwide IT Spending Forecast [80]

	2020 Spending	2020 Growth (%)	2021 Spending	2021 Growth (%)	2022 Spending	2022 Growth (%)
Data Center Systems	1,78,836	2.5	1,96,142	9.7	2,07,440	5.8
Enterprise Software	5,29,028	9.1	6,00,895	13.6	6,69,819	11.5
Devices	6,96,990	-1.5	8,01,970	15.1	8,20,756	2.3
IT Services	10,71,281	1.7	11,91,347	11.2	12,93,857	8.6
Communications Services	13,96,334	-1.5	14,51,284	3.9	14,82,324	2.1
Overall IT	38,72,470	0.9	42,41,638	9.5	44,74,197	5.5

characteristics (volume, variety, velocity, veracity, etc.) of big data, privacy, and security are defined as cybercriminals, unstructured data, physical risk, and security breaches [81]. The role of big data in healthcare and diagnosis is explained briefly in [82]. Large amounts of data are generated here through medical devices like MRI, CT, FMRI, etc. [82]. Big data application in banking is defined in loan risk analysis, anti-money laundering, trade analytics, predictive analytics, know your customers, etc. [52]. Video surveillance uses big data analytics for storage, retrieval, processing, access, and effectiveness [82]. Big data's significance relies on national development, industrial upgrades, scientific research, interdisciplinary research, and better prediction for the future [83]. Big data applies to other sectors, including banking and securities, insurance, manufacturing, natural resources, etc. (Fig. 1.)

Government policies

Big data are valuable and helpful for various schemes. It helps the government understand the clusters of objects that provide end-users facilities. Collecting data sets helps the government identify areas and methods to profit the nation's citizens. It includes data received from every transaction and other access to IT resources such as e-Mitra. Big data provides a platform for government officials to share and exchange information about the public's direct beneficiaries of schemes. However, there are still challenges in filtering and accessing the correct data and information, a massive problem in a survey of below poverty line (BPL) and other schemes. The primary beneficiaries are still away from getting benefits from the government. So, the big data cleaning and clustering approach to data filtration help citizens access government resources on time. Data cleaning is also applicable in crop selection analysis in agricultural applications, where big data is used effectively [84].

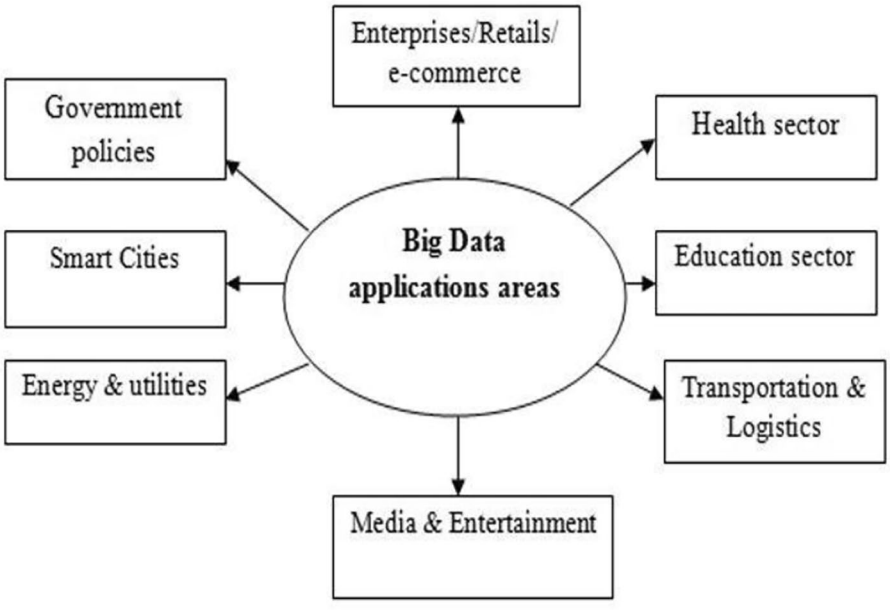

Fig. 1 Big data applications

Smart cities

Smart cities are the recent concern of big data applications where intelligent cities such as "smart health, smart logistics, smart education, smart transport, smart energy, etc." exist. However, these components require different data sets, acquisition methods, abstraction methods, classification and clustering of features, etc. People live a quality of life with an efficient and practical living model [71]. It comes to every citizen once a city's system is established in all respects [38].

Efficient classification algorithms are proposed in the form of multi-class classification, Naïve-Bayes classification, etc. [85]. Intelligent cities improve people's living standards by providing support for innovative health serviInnovativeSmart health services require real-time data for processing and sharing, sensors facilities, cloud services, the transformation from mobile to cloud, and sensors systecovercloud, etc., for various activities. Virtual machine (VM) migration technology transfers heterogeneous data to the cloud for access and store efficiency [86]. The authors have identified a virtual machine (VM) model with an ant colony optimization (ACO) approach for heterogeneous cloud computing systems (CCS) to enhance services in the innovative health system [86]. IoT-based big data is suitable for the decision-making system to form a smart city through MapReduce and HADOOP [87]. IoT is becoming an essential tool for ICT support in the smart cities framework designing process [88]. Smart city data analytics panel (SCDAP) was developed to incorporate functionalities, data model management, and information and communication technology (ICT) tools to the urbanization development of facilities for residential citizens [89].

Energy & Utilities

Energy & Utilities industries have started shifting from traditional systems like geographic information system (GIS) to cloud-based and big data (where many open-source technologies help achieve cleansing or filtered data from a large data pool). Industrial-generated sensor data are everywhere, and it is helpful for Energy &utility sector. This sensor data may be useful for predictive analysis modeling for the market, and spatial modeling design for analysis, including investigation, construction, transportation, and distribution. Big data open-source technologies can save time and money when a large volume of spatial, sensors, and GIS data are converted through HADOOP.

Education sector

Nowadays, information and communication technology (ICT) is an essential tool in the education sector, and every government has been promoting it since its inception. It improves effectiveness and efficiency at teaching and learning levels in education [38]. It also increases the outcome and productivity at various levels, and the government has been supporting implementing and enhancing assessments every time to benefit stakeholders and citizens. Innovative education policies bring an active learning environment for stakeholders, and through operational learning phases, society becomes educated and understands the value of growth in all respect [38]. Big data transfer data and knowledge through ICT tools to citizens, access resources, and engage in implementation projects. With the help of ICT and Big data technologies, a knowledge-sharing pool was generated, institutions collaborated, and industries-academia collaborations were performed, which helps build healthy societies and nations [38]. To understand the job

market advertised with "Big data," an article published in this work identified the content analysis for job advertisements [90].

E-commerce

It is one of the most significant big data applications, and because of this commercialization status, the values keep going on. In e-commerce, big data keep records of customers who purchase support to identify customers based on a health context. There are several ways in which big data may arise. In industries, the main objective is to earn more profits by providing services, goods, and materials in a competitive market and satisfying customers and other stakeholders [91]. E-commerce industries need big data analytics for various purposes, including decision making, market segmentation, business model, infrastructure, and transparency. Different types of big data are used in e-commerce, such as transaction or business activities, voice, video, etc. In contrast, business value defines e-commerce functions in detail [92]. As per [93], a CPG company generates 1,52,000 samples per second, resulting in 4 trillion data per year, which need to manage effectively and efficiently through the data processing stages.

The authors defined big data analytics as finding data values for industries and the economic system, human interactions with data analytics, process integration, and automation [94]. To make business experts understand drivers and the advanced analytical system implementation process, to upgrade the business process system, suitable for transparency, descriptive, and predictive, which are closely related to organizational functions. In 2016, 24 articles on big data in supply chain industries were published [94]. The authors have expressed different big data analytics applications in supply chain industries, including finance, manufacturing, and healthcare [94]. Business value is achieved through big data architectural concepts and information technology capabilities vision [95].

Transportation and logistics

This is another big data application area where the integration of e-commerce with transport or logistics plays a vital role in boosting industries for profit. But at the same time, many challenges occur due to large and inappropriate data generated by different sources of devices and equipment [96]. Data analytics tools require integrating components in the transport and logistics industry, but problems cannot be solved up to the mark.

Data analytics tools are used to inspect, filter, and transform (e.g., OpenRefine, Knime, R-programming, Orange, etc.) and to find meaningful results in big data [96]. Transport-related data is growing, including aircraft passengers worldwide, so processing becomes essential to provide features such as optimized infrastructure, better customer services, and prediction of components such as wrong road, traffic, passengers availability, etc. [96]. The data processing mechanism includes refining and collecting multiple data from different sources and aggregating them for processing [96]. Tourism is the other sector where big data is used to find prospective tourists based on the prus data.

Media and entertainment

This is another area of big data applications where media and entertainment agencies want to access big data resources in a profitable manner [97]. With large numbers

of digital audiences, media, and entertainment industries can utilize their big data resources more effectively [98]. The media and entertainment industries use big data to predict what audiences want, scheduling optimization, increase acquisition and retention, effective ad targeting, etc. [98]. Big data challenges in media and communications include leveraging mobile content, collection, analysis, pattern recognition, etc. [52]. Various big data applications in media and entertainment exist, including data journalism, dynamic semantic publishing, social media analysis, crisscross-sealing ducts, product development, and audio insight [14].

Big data in business organization

Big data is synonymous with business intelligence and business analytics tools (HADOOP, HPCC, STORM, QUOBLE, CASSANDRA, Spark, Rapid Miner, SAS, Knime, Orange, Weka, etc. NodeXL, Gephi, etc.), and data mining. It has shifted the business intelligence approach from reporting and decision support systems to predicting operations' benefits [91]. These big data tools (HADOOP, CASSANDRA, etc.) and cloud-based analytics brought advantages to businesses when the data volume was high. A hybrid kernel fuzzy c means algorithm (HKFCM) is proposed To maintain the voluminous data through clustering [99]. This method is passed through the setup and encryption stage, data storage, data reconstructions, and access stage [99].

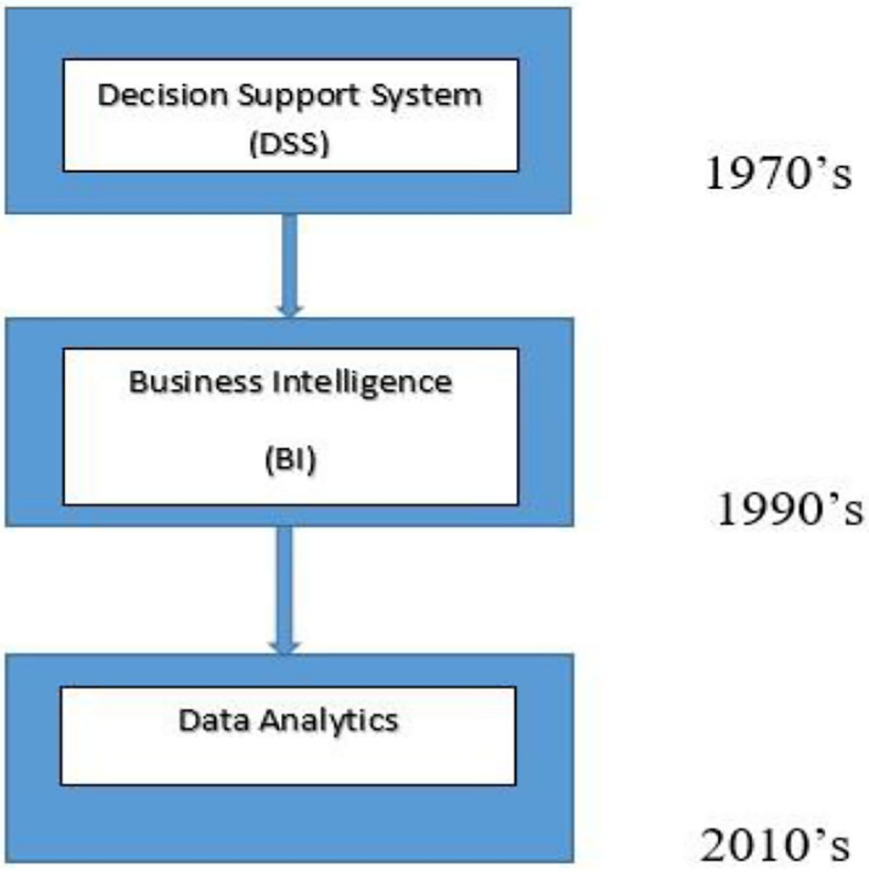

Fig. 2 Big data in business organization

These tools and in-memory analytics easily identify new data for business profit and help industries predict outcomes based on learning or trained data. Big data in business organization has grown from a decision support system to analytics and can be understood from the below diagram (Fig. 2).

Big data is essential and compares technologies, products, services, customers, suppliers, feedback, etc. of peers. Big data is a concept that implies access to and storing large volumes and a variety of data along with processes, where industries use data through several techniques and commercialize it [28]. The big data processing method is an effective in a large business organization where a large amount of data is available to process. Still, it includes data duplicity, noisy data, missing values among data, etc.

It contains many forms, and according to the query generated, it's been provided to the client. So, industries' benefits would be cost optimization, less time, new product development process, understanding market conditions based on customer demand, and online business control through sentiment analysis.

Big data in industrial processing

Big data are defined as voluminous data generated through different sources and mediums. It stores data beyond the feature of a relational database system—industrial extensive data rises-A whitepaper, including its importance in [93]. Big data analytics provides many supports like a decision support system, predictive modeling, futuristic activities, customers, business operations to industries, or enterprise systems [100]. Analytics-based enterprise information system (EIS) explains knowledge-based, analytics engine, big data analytics, and service analytics, which passes to GUI (graphical user interface) based approach to the end-user for facilitation [100]. Big data is a data analysis technique enabled recently in industries and other sectors mentioned above in Sect. 4. Because of its voluminous, cloud computing supports data storage of big industrial data (BID) and processes it for computing purposes. Various industries like telecommunication, manufacturing, retail, agriculture and farming, fitness, sports, consumer goods and trade, building, insurance, transportation, banking, financial services, tourism, and government benefit and implement big data technologies [101]. Manufacturing industries also need the resource management process long term profitability identification through a decision support system [102]. Manufacturing industries require big data analysis to optimize the resources and maximize the industrial benefits for their different operations. Innovative industrial process 4.0 uses other data generation sources for its industrial processing. The various industries (small scale, medium scale, large scale, and very large) use big data differently. Telecommunication industries use 87%, financial services industries use 76%, and healthcare uses 60% [103]. Big data analytics (BDA) is used effectively in sustainable smart manufacturing, where manufacturing resources, processes, and products are integrated for profitability enhancements [104]. Big data is suitable for industries in different forms, such as on-demand self-service, resource pooling, rapid elasticity, cost-effectiveness, etc. Various factors, such as human resources, technology resources, management support, adoption cost, security, privacy, complexity, regulations, etc., affect organizational adoption of big data technologies [105]. The adoption of big data depends on orcorporatetechnologies or environmental context [103]. Big data is also suitable for cost reduction, faster and better decision making, new products

and services, product recommendations, and fraud detection. Big data are transforming businesses through the advancement of robotics and automation system.

The sensor system is installed into manufacturing units and a machine to track the product. Business intelligence (BI) comprises applications and technologies suitable for collecting, analyzing, integrating, and presenting operational functions. The business intelligence objective is to provide an effective decision-making system [106]. For collaborative work and outcomes, BDA is used for organizational decision-making and co-innovations for many stakeholders [107]. However, different big data analytical approaches lead to business failure, where big data can bind to varying stages for justifying business failure [108]. Production industries have different types (numeric, audio, video, text, structured, semi-structured, and unstructured data) and various data sources (operational unit, sensors, simulations, CCTV, maintenance, store, online, etc.). Industrial operations improve the business objectivities through big data utilization, tools sophistication, and analytical skills [109]. These data are used for analysis for various purposes, such as safety, human resources efficiency, employee consistency, accuracy, etc. Big data technologies are more suitable for processing complex datasets effectively than traditional automation and data processing. Data processing removes adhocism under several processes in data acquisition. Big data analytical technologies extract hidden patterns, correlations, and so on in this enormous data. Today's industries need a business intelligence system to provide low-cost, and flexible service models. Industrial processes handle data through different stages, from collecting the correct data, storing effectively, analyzing, and reaching end-users. In these stages, sensors generate 4 trillion samples per [110].

Most IoT-based industries use Machine learning effectively to process a massive volume of data for decision-making and accurate forecasting [111]. IoT and industrial internet, cyber-physical systems (CPS), cloud-based technologies (CBT), data mining (DM), and artificial intelligence (AI) are the technological tools for smart manufacturing [104]. IoT can increase operational and recycling efficiency by 30% [104]. In [56], the author discussed improvement factors, including big data analytics management capability, technology capability, and talent capability for different industries such as financial, manufacturing, education, wholesale, and retail. In [23], the authors showed data analysis for the manufacturing process, including data acquisition to visualization form that suits end users. Various sources such as simulation, energy, GIS, sensors, strategy, and ERP systems generate enormous data sets analyzed effectively before their use. Data action shorting plant is discussed in [23]. Industry 4.0 requires implementations of modularity, security, interoperability, decentralization, real-time, service orientation, and visualization [112]. The potential benefits of big data for industries will vary from industry to industry.

According to the SAP annual report 2014, a prediction and decision were made to improve player performance at the world cup in Brazil [113]. Data generated through various devices, machines, cloud-based, business operations, systems, etc., in the modern industry, has reached up to 1000 Exabytes annually [114]. The primary objective of using big data in modern industrial applications is to acquire fault-free and cost-efficient running of the process [114]. Big data is suitable for various purposes, including prediction, minimization cost, and time, the effort for a different segment for modern

industries 4.0. Innovative products are developed by the intelligent and contemporary industries where massive data sets are used, and it is well suited for intelsmartsions.

According to [114], McKinsey Inc. suggests that manufacturing costs are reduced by 50% and up to a 7% reduction in working capital by using big data. Industry 4.0 mainly integrates automation, the Internet of Things (IoT), cloud computing, and wireless and concentric computing [112]. Industry 4.0 essential design principles and implementation, including different parameters. It includes modularity, visualization, security, interoperability, service orientation, and decentralization [112]. Similarly, industrial IoT(IIOT) generates large amounts of data because of the implementation of sensors, and IoT devices, which create difficulty for big data processing tasks due to low computations at IoT devices [115]. Design implementations of industry 4.0 are explained through modularity, security, interoperability, real-time service-oriented, and visualization [115]. Similarly, in [116], the authors explained different agriculture sensors such as optical, mechanical, electromechanical, dielectric soil moisture, etc. In [112], the authors stated the vision of the industry 4.0 manufacturing system needs big data analytics techniques and IoT for value creation. Big data analysis is executed through sensors, processing, communication, and storage [117]. The capabilities of big data analytics (BDA) in industrial manufacturing processes encompass a respective challenge (i.e., quality or process control, energy efficiency, diagnosis and maintenance, and risk analysis) followed by big data analytics techniques resulting in its values [118].

Data mining classification techniques(descriptive analytics-22%, inquisitive analytics techniques-30%, predictive analytics techniques-41%, and prescriptive analytics techniques-7%) are defined and classified as per the tabular form mentioned in [118]. Descriptive analytics is used to identify facts based on historical data. It requires clustering, correlation, etc. Predictive analytics predicts the future through statistical methods such as decision trees. At the same time, prescriptive analytics is used to derive the best possible outcomes [20].

Big data analytics (BDA) requires essential tools in a fertilizer plant, phosphoric plant, as mentioned in [118]. BDA in manufacturing sectors is data-driven and uses artificial intelligence,e (AI) and machine learning (ML) approaches to provide actionable results.

Big data in the healthcare sector

The health sector is a data-rich, high-volume sector. The health sector is another critical sector, and quality health services should be provided to everybody irrespective of their place, caste, and religion in society. In the healthcare sector, millions of sensors and devices generate patient-related data at different levels every day, which can predict the disease and be helpful to suggest citizens of the societies. Healthcare analytics is required for inpatient privacy, security, care, treatment, etc. [13]. The authors have also shown growth in healthcare analytics in the last eight years, from 2010 to 2018, where various journals and conferences paper were published [13]. In fact, in the healthcare industry, many devices in hospitals nationwide generate different data about the patient's health that can be sent to the cloud to support IoT services [91]. Integration of IoT for e-healthcare is discussed in [119]. Here, the importance of cloud computing integration with IoT was discussed through different tier explanations of other devices [119]. Healthcare information system (HISs) for diabetes analysis employs big data technologies where statistical assessment enhances accuracy and F measure [120]. The green

supply chain for the healthcare system is discussed with a focused study on environmental sustainability in [121].

Here, CAIE (Content analysis and information evaluation) and TAMS (Text analysis and mining system) are the two subsystems defined for extraction and pattern mining evaluations [121]. The significant data impact on the healthcare sector is explained through a patient-centric healthcare ecosystem where insurers, service providers, pharmaceuticals, practitioners, and patients are involved effectively [122]. Data sources such as government agencies, patient portals, research groups, public records, logs and notes, 3rd parties, pharmacies, medical claims, clinical, search engine databases, smartphones, etc., are found in big data healthcare cases [123]. The patient remote monitoring system is used to track patient disease status. It is helpful to identify the progress or deuteriations through high-performance computing infrastructure systems through retrieval, storage, and processing of data [124].

Different types of features of big data in healthcare are considered heterogeneity, incompleteness, timeliness, longevity, data privacy, ownership, etc. [125], Different types and sources of big data concepts and utilization in the healthcare sector are discussed in [125]. It includes big data in medicine, such as electronic health records (HER), electronic medical records (EMR), personal health records (PHR), medical images, vitals, human body samples, and clinical trials [125]. It includes big data concepts in public health, medical experiments, medical literature, hospital information system (HIS), and its evolution [125].In a similar line, the role of Nano-sensors and networks is suitable for a futuristic healthcare system where physical Nano-sensors, chemical Nano-sensors, and Bio Nano-sensors are used [126]. Different literature explosions up to 1800 have been recorded till 2017 in healthcare searched in PubMed as shown by the authors in [125].

Big data is changing the healthcare sector in various ways, including health tracking, reducing cost, assisting high-risk patients, preventing human errors, and developing in the healthcare sector. The healthcare sector is rapidly changed by transactional data, where most data are transformed into electronic health records (EHR). Medical treatment cost reduction, drug efficiency analysis, and preventive care improvement are mentioned in the given article [30]. In healthcare services, telehealth provides healthcare information services through mobile and computer devices [127]. The benefits of telehealth industries include detecting health risks, virtual care for senior living, attracting Medicare, etc. [127]. American health association (AHA)supports telehealth services to patients in online prescribing, medical malpractice, health professional license, privacy, and fraud [128]. It connects patients through videoconferencing and other electronic mediums and supports medical expansion coverage. Telehealth services have grown highly in hospitals from 2010 to 2017 [128]. Most hospitals (61.2%) implemented telehealth services until 2017 as per [128]. In 2015, it was 43.1% and grown up to 53.0% in 2016 [128].The scope of telehealth exists in various terminologies of service, software, and hardware. Services include remote monitoring, online interaction and communications, and consultations. Medical peripheral devices (MPD) include the facilities of ECG monitors, pulse oximeters, blood pressure, etc. [129].Different sources of big data generations in healthcare industries are hailed from medical claims, pharmacy claims, hospital EHR, academic researchers, etc. [130]. Big data challenges in healthcare include capturing data from multiple sources, cleaning, storage, security, visualization, data

integration, data aggregation, protecting patient privacy, etc. Data cleaning and tokenization is a technique of data mining [131].

EHR data are used for patient clinical treatment status. Data cleansing methodologies and processes are given in [132]. The most critical challenge is analyzing unstructured healthcare big data. In a structured data generation case, it is stored and used effectively through RDBMS software in a few instances. Unstructured data of patient-generated and stored, including different tests, scanned images, and progress notes of patients [19]. Big data technologies help improve patient care, enhance operational efficiency, and find a cure for dise,ases including clinical (readmission rate, overall health outcome, and other operational (general management decision making) aspects. Diseases prediction is carried out by machine learning technologies in the healthcare system through a convolutional neural network (CNN) based on risk prediction [133]. A stochastic prediction algorithm was developed to estimate and identify future health conditions based on previous and existing health data through correlation analysis [134]. It includes probabilistic data storage or collections made especially for cloud-based healthcare systems, correlation analysis, and an algorithm for predicting future health conditions [134]. Different analytical tools such as semantic analysis, predictive analysis, informatics analytics, descriptive analytics, prescriptive analytics, comparative analysis, and exploratory analytics are used in health big data analysis [135]. Healthcare sectors must be predictive and proactive, as mentioned in [136]. It must be individualized and focused rather than populations to provide better patient care [136]. To establish healthcare services, IoT, cloud services, applications and interfaces, sensors, and technologies work together to form a cyber-physical system (CBS) with unique components, including sensing, big data center, healthcare center, observation center, etc. [136],

Through IoT, 50 billion objects will be connected globally by 2020 for data generation and sharing [137]. Big data for IoT for modeling, data-intensive applications, services, and intelligent intel environments (smart cities, etc.) is examined and used for profitable outcomes for industrial goals [138]. The various task includes data preprocessing, data fusion, feature extraction, visualization, fuzzy logic, and supervised learning. Unsupervised learning, identified problems of missing value estimation (MVE), dimension reduction, feature selection, decision-making system, disease risk prediction, and tumor clustering is used in the healthcare and industrial sector for improvement as per [126]. Many policy actions, such as stakeholders, data sources, data analysis, etc., are identified in healthcare sectors. A hypothesis of mental health cases and Northern Ireland was explained through data science and machine learning technologies where much heterogeneous medical data set was used [139].

Big data in healthcare includes multiple factors such as public records, search engine data, smartphones, patient portals, generic databases, and electronic health records (EHR) [140]. The medical expert system and big data analytics (BDA) in healthcare are discussed in [141]. It explained the data analysis life cycle, multiple data generation sources, and CHESS(Comprehensive health electronics software system) [141]. Big data applications in healthcare include a few other parameters, including diagnostics, medical research, reduction of adverse medication effects, cost reduction, and population health [140]. Medical-related big data generations require consistency, data privacy, and security on platforms like cloud services [142]. One of the exciting areas of big data applications is the prediction of finding train delays through a tremendous amount

of train movement data in Italy [143]. The path value of big data in healthcare passes through data sources, analytical techniques, big data analytics capabilities (prediction, data mining, monitoring, etc.), deals, and challenges [144]. Various activities and operations include RFID, barcodes, sensors, monitors, consultation recordings and notes, patient instructions, social media discussions, blood test results, clinical information system, payment, etc. It can be executed with big data databases as given in [52]. The key elements such as integrating data, generating new knowledge, and transforming knowledge into practice make the healthcare sector leverage big data more broadly. Healthcare industries have unlike types of instrumentation data, diagnostics, unstructured data, and structured data. Big data technologies use unstructured data to store, correlate, process and evaluate in meaningful terms for patient benefits [145].

Apart from the issue mentioned above in healthcare issues, it includes some important and internal problems such as international level, national and regional level, hospital level, home and family level, a personal level as per given in [41]. Many metaheuristics algorithms can solve the healthcare sector's optimization and complex form problems [41]. It includes a genetic algorithm (GA), particle swarm optimization (PSO), firefly algorithm (FF), rider optimization algorithm (ROA), ant colony optimization (ACO), and so on [41].

Challenges and opportunities in big data

Although many authors give various challenges, few are important such as big data imbalance, big data management, big data cleaning, big data analytics, big data aggregation, big data accessibility, and so on [146]. These challenges are also described in big data management, imbalanced big data and system capacities, big data machine learning, and so on [122]. There is a dissimilar hindrance to industrial big data analysis where innovative manufacturing-based systems must be implemented. It consists of long-term investment in IT infrastructure, and it's difficult to move on automation for industries, industries regulations and constraints, protocol systems, low vision and commitment, risk association with the new technologies, etc. [25]. Diversified challenges are found in the forms of a futuristic healthcare system that includes data format standardization, unified data schema, data quality parameter, aggregation scale, analytics tool, etc. [126], There are various other challenges such as data representation, data compression, redundancy reduction, data life cycle management, data confidentiality, data analytical mechanism, scalability and expendability, cooperation in case of interdisciplinary, and so on [18]. Big data challenges also exist in data complexity, computational complexity, system complexity, and so on [83]. However, there are several challenges in finding meaningful data, and big data is data mining, data storing, data sharing, data privacy, data technologies, and talent. Data sharing supports limited data standardization, information barriers, and insufficient data integrations in big data applications [125]. In [32], the authors described a few challenges for big data as a compounding factor: dealing with missing data, duplicate data, data heterogeneity, semantic data generations, etc.

Some important and common challenges are available with big data in industrial processing and healthcare, including data storage, data quality, security and privacy, reliability and availability, big data filtering, big data processing tool, and big data analytics. Big data analytics and processing are challenging due to storage, networking, and limited computational access [112]. In addition, more challenges with big data create hurdles for

industries such as scalability, lack of professionals (data scientists), lack of appropriate data source with business objectives, security, and big data project cost evaluation and management. Various other challenges are also found in terms of data aggregation, data maintenance, data integration and interoperability, data analysis, and so on [135]. Some implementation level challenges in the healthcare sector are data management (20%) and technological issues (10.2%). Here, lack of technologies would not be in a position to produce effective results, evaluation process(7.1%), financial issue (3.1%), regulation(5-0.6%), awareness(0.5%) and support, political issue (0.4%), etc. [144]. So, the challenges and techniques through various research articles/papers/works in a consolidated form are mentioned below in Table 5. It represents the result analysis report of different healthcare and industrial processing works, including methodologies, features, and challenges mentioned in Table 2. This result analysis people 4) shows that data cleaning and outliers are important and have more weightage than other challenges.

From the literature review, the techniques below apply to the mostly application-based project, which is commercialized. In various literature review articles, it was found the performance measure in terms of accuracy, sensitivity, and precision [34, 125, 147–149]. Most of the article has worked with healthcare and industrial data taken from various SCI, SCIE, SSCI, ESCI, Web of Science, Scopus, etc., from 96 different journals and industrial processing-related articles, including 24 other journals. These technologies (as mentioned in Table 5), can also work together to improve the result for accuracy, sensitivity, precision, and specificity.

The proposed method for the data cleaning model and outlier removal has been given in subsections 6.2 and 6.3. Metaheuristics approaches have been preferred with machine learning to provide better healthcare and industrial data results. It has been observed; that metaheuristics are suitably applied in classification, clustering, feature selection, and association rules during experimental analysis. Table 5, shows that enhanced data cleaning and outlier detection models are required to meet noise/dirty free data for the growth of industrial performance analysis. Noisy and insufficient data will not be sufficient to conclude any applications.

Big data filtering

Data filtering encompasses several steps in many applications and provides better results, leading to the success of the application system in industries [77]. Big data is a critical approach for the industrial database system [77]. Since many sectors, such as telecom, e-commerce, banking, insurance, health, etc., rely on data generated by other authentic se such as banks, data filtration becomes essential in all respect [37]. It helps to ctorremove noise, incorrect, inappropriate, incomplete, and duplicated [37]. Data quality is affected by many factors, such as accessibility, completeness, accuracy, ease of understanding, value-added, etc. [150]. So these factors affect data quality in diverse ways [37].

However, the system's dissimilar error must be resolved before processing and analytical applicability evaluation in any system [37]. Data filtering includes measurement of error, inconsistent data(e.g., duplicate data, values out of band, error code, contradictive error, etc.), outlier, missing data, etc. [37].The challenge is how to perform and provide the complexity of big data nature (volume, variety, velocity) in a better way. It leads to an evaluation of the reliability analysis approach, which identifies the relationship and

Challenges / Techniques	*Data cleaning*	*Outliers*	*Missing values*	*Dirty data*	*Noisy data*	*Data quality*	*Data overload*	*Data integration*
Machine learning	✓	✓	✓	✓	✓	✓	✓	✓
Data analytics	✓	✓	✓	✓	✓	✓	✓	✓
Clustering	✓	✓	✓					
PCA	✓	✓	✓	✓	✓	✓	✓	✓
Analytical Hierarchy Process (AHP)	✓	✓		✓				
TF-IDF			✓		✓			
Discretization & Normalization								
Deep Learning							✓	✓
AHP								
Correlation analysis	✓			✓				
Automation	✓					✓	✓	✓
Big data classification	✓					✓	✓	✓
Big data analytical capability model (BDAC)	✓	✓						
Data preprocessing	✓	✓	✓	✓	✓			
Cost sensitive learning	✓	✓	✓	✓	✓			
Filtering strategies	✓	✓	✓	✓	✓			
Standard support vector data description (SSVDD)		✓						
Kernel support tensor data description (KSTDD)		✓						
Multiple soft majority voting methods (MSMV)		✓			✓			
Kalman filter	✓	✓			✓			
ICCFD_Miner techniques (Interest Constant Conditional Functional Dependency)	✓							
Systematic sampling method	✓	✓						
Wavelet-based multiresolution analysis technique (WMAT)	✓				✓			
Algebraic method	✓							
Distributed principle component pursuit (D-PCP)	✓							

Table 5 Result analysis report

complexity of the nature of big data [146]. At the same time, it is also essential to manage data collection, integration, and storage with few hardware support and software requirements [146].

Proposed data cleaning algorithm

From the literature review [60, 62, 63, 68, 75, 84], it was identified the problem of data cleaning in the form of text classification, sentiment analysis, text clustering, etc., along with cleaning behavior and algorithm usability. Moreover, the Kalman filter mechanism fails to support extensive sensor data, which is a part of data generation sources in industries and the healthcare sector. The recent research of the year 2018 also identified the

big data processing problem and data cleaning problem in big data processing. So far, different conventional data cleaning processes and methodologies exist suc,h as interest constant conditional functional dependency (ICCFD)_miner techniques, systematic sampling method, wavelet-based multiresolution analysis technique (WMAT), algebraic method, fuzzy Kohonen clustering network (FKCN) algorithm, Rete algorithm, distributed principal component pursuit (D-PCP), stacked denoising autoencoder (SDAE), etc. These methods include several challenges: data repairing, high processing time, cleaning integration, high computational cost, distributed real-time cleansing, etc.

So, to solve the data cleaning problem, a new and novel mechanism is proposed to identify and detect abnormal data. This data cleaning approach would be in two phases: dirty data detection (DDD) and dirty data cleaning (DDC). Dirty data detection goes through data normalization, hashing, clustering, and finding the inappropriate data, and the messy data cleaning phase passes through leveling, Huffman coding, and removal of erroneous data. However, centroid selection is an issue in the clustering process and the removal of suspected data. So, these two issues can be solved through the optimization concept. In this case, a metaheuristics algorithm would be beneficial. For finding the suspected data under dirty data cleaning (DDC), a combined approach of rider optimization algorithm (ROA) and firefly algorithm (FF) can be considered. The performance of the proposed mechanism will be compared with the grey wolf optimization (GWO) and particle swarm optimization (PSO) on certainspecificters comprising false-positive rate (FPR), false-negative rate (FNR), precision, accuracy, sensitivity, specificity, etc. The healthcare dataset such as diabetics, heart, and breast cancer is required from UCI Machine Learning Repository [151].

Proposed outlier removal mechanism

From the review [7, 19, 23, 27, 31–35], it's stated that outliers removal should be one of the critical challenges and opportunities in big data-based applications. Outlier removal is next under the data cleaning process. Outlier detections are parametric and non-parametric. Statistical methods fall under parametric, whereas non-parametric consist of density-based, clustering-based, and distance-based. The k-means clustering approach can be considered out of several outlier removal techniques because it does not require preliminary data information or knowledge. So far, different methods exist like k nearest neighbor (kNN), k-means with outlier removal (KMOR), generalized linear model (GLM), maximum correntropy estimator (MCE), maximum likelihood estimation (MLE), two-stages, and standard operating procedure (SOP). All outliers will be placed into a single cluster in the proposed outlier detection model so that it would be based on a k-means clustering process called enhanced k means outlier detection model (E-KMOR). However, k-means with optimal centroid would be an approach where centroid selection is carried out through an optimization mechanism.

So, to solve the optimization problem, a coma hybridoach of lion algorithm (LO) and particle swarm optimization (PSO) was proposed. The proposed enhanced k means outlier detection model (E-KMOR) would be compared with the existing generalized linear model (GLM) and K-means outlier removal model (KMOR) based on parameters such as accuracy, sensitivity, specificity, and so on. The weakness of this outlier detection model may fall under data volume. It needs large dataset values of any application. The proposed E-KMOR is an extension of the K-mean algorithm and does not apply to data

compressing. The future work may be towards implementations of EKMOR towards data sampling.

The big data processing tool

Many tools support big data compatible with cost efficiency, timing, accuracy, etc., to perform data analysis tasks under massive datasets [152]. Developing large-scale, accurate, adaptive, error-free, and fault-tolerant systems is complex [17]. Some essential big data processing tools are HADOOP, MapReduce, HPCC, Storm, Qubole, Cassandra, Statwing, CouchDB, Pentaho, Flink, Cloudera, etc. An experimental big data framework for machine learning and graph processing is mentioned in [153]. It explained MapReduce, HDFS, YARN (yet another resource negotiator), Apache Spark, etc. [153]. HADOOP supports community technologies and processes data where data is stored [122]. HADOOP can execute voluminous data in a few seconds with the support of parallel clusters and distributed file systems [122]. It has been designed to reduce timing complexity and space as well. It performs fast because of the Hadoop distributed file system (HDFS) and MapReduce. HDFS can store different types of data formats of large volumes.

The beauty of HDFS exists in heterogeneous hardware and software platform-independent [122]. MapReduce is used for data processing through job scheduling and resource management. One of the programming models under HADOP effectively reduces massive data processing complexity [122]. MapReduce performs execution in four stages: map function, key-value pair with the mapper, reduce function, and store key-value pair in the output file [122]. These tools help in multiple ways: framework design, a data computation system, a data management platform, an architectural structure for data processing, and statistical tests automatically for the massive volume of data [152]. Big data tools are used to develop and design healthcare frameworks through data integration tools, scalable searching, processing tool, machine learning tools, real-time and data stream processing tools, visual data analysis tools, etc. [122],

Big data analytics

For various application areas, big data execute transformative potential to harness such vast volumes of data. Advanced data analysis tools are required to measure the relationship among features and explore data at a lower level of access [146]. Big data applications need real-time analysis for practical evaluation through data processing models such as predictive analysis, NoSQL database, Search & knowledge discovery, stream analytics, distributed file sore, data virtualization, data integration, etc. Data analytic tools such as (Tableau public, OpenRefine, KNIME, RapidMiner, Google Fusion Table, NodeXL, Wolfram Alpha, etc.)are used under big data to meet the required services of industries [154]. Big data analytics enhances operational efficiency, decision- making, market demand prediction, etc.

Healthcare-based industries use artificial intelligence mechanisms such as machine learning (ML), agent-based (AB), heuristic-based (HB), cloud-based (CB), and hybrid-based (HB). These techniques are implemented through big data on various factors to enhance reliability, efficiency, accuracy, performance, availability, QoS, etc. [155]. The parameters (accuracy, sensitivity, specificity, FPR, FNR, NPV, etc.) would be used in the proposed work to compare the existing algorithm such as PSO, FF, GWO, etc. These

parameters are used in test cases of experimental work. Since various iterations are performed, so it becomes essential to use. The advantages and disadvantages of machine learning are given in [155].

Future research outline

Various challenges exist in big data that need to address. However, challenges may include new technologies, scaling data processing techniques in big data, big data learning paradigms, data sources, attributes, smart cities, etc.

It is clear from Sect. 6.4) that big data processing tools offer various services that need to be verified effectively for any system before using it. In the case of Spark, which performs better than HADOOP in processing [27]. Spark is the latest technology, and it will take time to develop concepts [27]. Similarly, Flink is another emerging technology platform that reduces the gap between stream data processing and batch processing.

Flink's core supports data allocation, message, and fault tolerance for dispersed computations over data streams. Whereas in the case of scaling data processing technology, the majority of tasks (such instance, reduction, missing values imputation, and noise treatment) need to resolve sincerely [27]. Apart from these, the optimization of big data techniques is equally important [156]. However, the big data learning paradigm includes unsupervised, semi-supervised, data stream, and non-standard, apart from classification and regression methodologies [27]. Different analyzing tools can be considered in classification, clustering, regression analysis, and association rules applicable after generations of large and voluminous data in healthcare sectors [125]. Different clustering techniques exist, such as hierarchical-based, partitioning-based, density-based, grid-based, and model-based. These clustering methods can also be used in healthcare and industries [157]. Data sources and attributes are essential in the case of data sources and features, and their variants in different forms are also challenging to manage [38]. Modern industries require a novel method and process improvement for big data analysis.

In industrial processing, cloud computing-based solutions are needed for ample data storage and analysis. For controlling and monitoring, generous data tools support must be focused on. Big data analysis is required for many industries' multiple and customized ERP solutions. Future work on big data analytics (BDA) in healthcare is necessary for approach development, a mechanism for data value maximization, risk improvement, etc. [144]. Some issues and challenges, such as handling extensive voluminous data, data processing, etc., are mentioned in [158].

Conclusion

Big data is a modern and vital area, where this paper focused on showing various authors' objectives, conclusions, and future work possibilities. This paper describes the author's contributions to big data and industries with multiple technologies. It shows many open issues, opportunities, and challenges that need to resolve in the future. Many challenges were discussed in the paper regarding analyzing, processing, and deployment in various industries. This paper focused on applications in modern enterprises and the healthcare sector. It was focused on machine learning and other technologies such as deep learning, classification, PCA, discretization, AHP, and data preprocessing. These methods worked on healthcare, other applications, and industrial data, where results improved for better services. However, effects can be enhanced regarding existing and studied

measured parameters, including accuracy, sensitivity, specificity, and precision. In the proposed work, a few other significant measurement parameters can be added, such as false-positive rate (FPR), false-negative rate (FNR), negative predictive value (NPV), and false discovery rate (FDR), and F1_Score, for higher and more effective results. These results would improve the timing complexity of the system to perform. Timing complexity improves once accuracy, sensitivity, and specificity enhance.

Big data analytics (BDA) in manufacturing identifies some crucial challenges which can be further taken up for improvement in terms of quality and process control system (QPCs), proactive diagnosis (PD), etc., before the implementation of BDA. Big data values such as performance improvement (PI), transparency, and decision support system (DSS) would be the objective for further improvement. Sustainable smart manufacturing, where data and services integrate to provide business outcomes, requires the effective integration of data sources. The significant outcome state requires a data cleaning model to detect and rectify anomalous data. Dirty data identification, removal, and outliers are essential issues in big data-based healthcare and industrial applications. These reviews show that data processing analysis, removal of outliers, and data cleansing methods are essential issues in the large dataset under any applications. Another significant result and conclusion of most of the research article state to develop an optimization model for outlier's identification and removal in large applications.

This paper indicates the three essential fields for further research to be carried out, i.e., big data filtering, big data processing, and big data analytics for an organization's technological improvements, organizational improvement, and research improvement. Several challenges can be carried out as research topics, including data cleaning, outlier removals over large datasets, missing value identification and removal and replacement, etc.

Out of these, this systematic review proposed a data cleaning model and outlier removal model. Those modelings are required to improve the timing complexity of the system on specific parameters mentioned above to serve better and provide more accurate data to applications. This paper deals with how industrial processing and healthcare sectors can utilize big data to enhance services in their respective areas to benefit organizations.

Abbreviations

ACO	Ant colony optimization
ADR	Automated adverse drug reaction
AHP	Analytical hierarchy process
ANN	Artificial neural network
BDAC	Big data analytical capability
BDATLC	Big data analytical talent capability
BDATEC	Big data analytical technology capability
BDAMAC	Big data analytical management capability
BDARBT	Big data analytical Resource-Based Theory
BID	Big industrial data
CAIE	Content analysis and information evaluation
CCS	Cloud computing systems
CHESS	Comprehensive health electronics software system
CPS	Cyber-physical systems
CRM	Customer relationship management
DDC	Dirty data cleaning
DDD	Dirty data detection
D-PCP	Distributed principal component pursuit
ESMVU	Enhanced soft majority voting by exploiting unlabeled data
ERP	Enterprise resource planning
FF	Firefly
FKCN	Fuzzy Kohonen clustering network
GIS	Geographic Information System

GLM	Generalized linear model
HDFS	HADOOP distributed file system
HIS	Healthcare information system
HRM	Human resource management
ICCFD	Interest constant conditional functional dependency
ICT	Information and communication technology
kNN	K Neural Network
KMOR	k-means with outlier removal
KSTDD	Kernel support tensor data description
MCE	Maximum correntropy estimator
MLE	Maximum likelihood estimation
MRA	Maximum Reward Algorithm
MSMV	Multiple soft majority voting
NIA	Nature-inspired algorithm
NLG	Natural language generations
PCA	Principal component analysis
PSO	Particle swarm optimization
RFML	Random forest machine learning algorithm
ROA	Rider optimization algorithm
SAAS	Software as a service
SADM	Simultaneously Aided Diagnosis Mode
SCDAP	Smart city data analytics panel
SDAE	Stacked denoising autoencoder
SHMM	System health monitoring and management
SOP	Standard operating procedure
SSVDD	Standard support vector data description
TAMS	Text analysis and mining system
VM	Virtual machine
WMAT	The wavelet-based multiresolution analysis technique

Acknowledgements
Authors would like to thank Department of Basic and Applied Science, NIFTEM Kundli, India, Department of Computer Science and Engineering, Rajasthan Technical University, Kota, and School of Science & Technology, Vardhman Mahaveer Open University, Kota, India for the utilizations of labs and resources.

Authors' contributions
Kumar Rahul explained big data: preliminary, identified different big data applications, elaborated on challenges and opportunities in big data and Rohitash Kumar Banyal focused on a review of big data analysis, explained big data in business organizations, and explores future research outlines and conclusions. Neeraj Arora collected big data tool implementation areas and worked on data processing tools.

Funding
There is no funding to declare.

Data Availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 July 2022 / Accepted: 2 August 2023
Published online: 27 August 2023

References

1. Amalina F, et al. Blending Big Data Analytics: Review on Challenges and a recent study. IEEE Access. 2020;8:3629–45. https://doi.org/10.1109/ACCESS.2019.2923270
2. Nazir S, et al. A comprehensive analysis of healthcare big data management, analytics and scientific programming. IEEE Access. 2020;8:95714–33. https://doi.org/10.1109/ACCESS.2020.2995572
3. Seh AH, et al. Healthcare Data Breaches: insights and implications. Healthcare. 2020;8(2):133. https://doi.org/10.3390/healthcare8020133
4. Islam M, Hasan M, Wang X, Germack H, Noor-E-Alam M. "Healthc. 2018;6(2):54. https://doi.org/10.3390/healthcare6020054. "A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining.

5. Geng D, Zhang C, Xia C, Xia X, Liu Q, Fu X. Big data-based improved data acquisition and storage system for designing industrial data platform. IEEE Access. 2019;7:44574–82. https://doi.org/10.1109/ACCESS.2019.2909060
6. "Technology _ Grand View Research_ Big Data Market Research. Report 2015 to 2022 by Grand View Research, Inc.".
7. Heureux AL, Member GS. Machine learning with Big Data : Challenges and Approaches. IEEE Access. 2017;5:7776–97. https://doi.org/10.1109/ACCESS.2017.2696365
8. Hussain S, et al. Implications of deep learning for the automation of design patterns organization. J Parallel Distrib Comput. 2018;117:256–66. https://doi.org/10.1016/j.jpdc.2017.06.022
9. Tsui KL, Zhao Y, Wang D. Big data opportunities: System health monitoring and management. IEEE Access. 2019;7:68853–67. https://doi.org/10.1109/ACCESS.2019.2917891
10. Ghasemaghaei M. Are firms ready to use big data analytics to create value? The role of structural and psychological readiness. Enterp Inf Syst. 2019;13(5):650–74. https://doi.org/10.1080/17517575.2019.1576228
11. Dang LM, Piran J, Han D, Min K, Moon H. "A Survey on Internet of Things and Cloud Computing for Healthcare," pp. 1–49, 2019, https://doi.org/10.3390/electronics8070768
12. Rathee G, Sharma A, Saini H, Kumar R, Iqbal R. A hybrid framework for multimedia data processing in IoT-healthcare using blockchain technology. Multimed Tools Appl. 2020;79:15–6. https://doi.org/10.1007/s11042-019-07835-3
13. Miah SJ, Gammack J, Hasan N. Methodologies for designing healthcare analytics solutions: a literature analysis. Health Inf J. 2019. https://doi.org/10.1177/1460458219895386
14. Kurumbalapitiya D. *Data acquisition*. 2005.
15. Ma Y, et al. Remote sensing big data computing: Challenges and opportunities. Futur Gener Comput Syst. 2015;51:47–60. https://doi.org/10.1016/j.future.2014.10.029
16. Agrawal D, Das S, Abbadi AE. Big data and cloud computing: current state and future opportunities. 14th Int Conf Extending Database Technol. 2011;530–3. https://doi.org/10.1145/1951365.1951432
17. Fan J, Han F, Liu H. "Challenges of Big Data analysis," *Natl. Sci. Rev*, vol. 1, no. 2, pp. 293–314, 2014, https://doi.org/10.1093/nsr/nwt032
18. Chen M, Mao S, Liu Y. Big data: a survey. Mob Networks Appl. 2014;19(2):171–209. https://doi.org/10.1007/s11036-013-0489-0
19. Sukumar SR, Natarajan R, Ferrell RK. Quality of Big Data in health care. Int J Health Care Qual Assur. 2015;28(6):621–34. https://doi.org/10.1108/IJHCQA-07-2014-0080
20. Rabhi L, Falih N, Afraites A, Bouikhalene B. "Big Data Approach and its applications in Various Fields: Review," *Procedia Comput. Sci*, vol. 155, no. 2018, pp. 599–605, 2019, https://doi.org/10.1016/j.procs.2019.08.084
21. Rahul K, Banyal RK, Goswami P. "Analysis and processing aspects of data in big data applications," vol. 0529, no. May, 2020, https://doi.org/10.1080/09720529.2020.1721869
22. Zhang C, Liu Z. Application of big data technology in agricultural internet of things. Int J Distrib Sens Networks. 2019;15(10). https://doi.org/10.1177/1550147719881610
23. Steckel T, et al. Big Data Analysis of Manufacturing processes. J Phys Conf Ser. 2015;659:012055. https://doi.org/10.1088/1742-6596/659/1/012055
24. Krishnan R, Samaranayake VA, Jagannathan S. A hierarchical Dimension Reduction Approach for Big Data with application to Fault Diagnostics. Big Data Res. 2019;18:100121. https://doi.org/10.1016/j.bdr.2019.100121
25. O'Donovan P, Leahy K, Bruton K, O'Sullivan DTJ. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. J Big Data. 2015;2(1):1–26. https://doi.org/10.1186/s40537-015-0034-z
26. Shah D, Wang J, He QP. "Feature engineering in big data analytics for IoT-enabled smart manufacturing – comparison between deep learning and statistical learning," vol. 141, 2020, https://doi.org/10.1016/j.compchemeng.2020.106970
27. García S, Ramírez-gallego S, Luengo J, Benítez JM, Herrera F. "Big data preprocessing : methods and prospects," pp. 1–22, 2016, https://doi.org/10.1186/s41044-016-0014-0
28. Bonde M, Bossen C, Danholt P. Data-work and friction: investigating the practices of repurposing healthcare data. Health Inf J. 2019;25(3):558–66. https://doi.org/10.1177/1460458219856462
29. Bossen C, Pine KH, Cabitza F, Ellingsen G, Piras EM. Data work in healthcare: an introduction. Health Inf J. 2019;25(3):465–74. https://doi.org/10.1177/1460458219864730
30. Kaur P, Sharma M, Mittal M. ScienceDirect Big Data and Machine Learning based Secure Healthcare Framework. Procedia Comput Sci. 2018;132:1049–59. https://doi.org/10.1016/j.procs.2018.05.020
31. Habib M, Sun C, Assad L. Big Data reduction methods : a Survey. Data Sci Eng. 2016;1(4):265–84. https://doi.org/10.1007/s41019-016-0022-0
32. Gudivada VN, Apon A, Ding J. "Data Quality Considerations for Big Data and Machine Learning : Going Beyond Data Quality Considerations for Big Data and Machine Learning : Going Beyond Data Cleaning and Transformations," no. July, 2017.
33. Deng X, Jiang P, Peng X, Mi C. Support high-order tensor data description for outlier detection in high-dimensional big sensor data. Futur Gener Comput Syst. 2018;81:177–87. https://doi.org/10.1016/j.future.2017.10.013
34. Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl. 2019;78:19905–16. https://doi.org/10.1007/s11042-019-7327-8
35. Oueida S, Aloqaily M, Ionescu S. A smart healthcare reward model for resource allocation in smart city. Multimed Tools Appl. 2018. https://doi.org/10.1007/s11042-018-6647-4
36. Fernández A, Nitesh R, Herrera F. An insight into imbalanced Big Data classification : outcomes and challenges. Complex Intell Syst. 2017;3(2):105–20. https://doi.org/10.1007/s40747-017-0037-9
37. Number D. "D3.1 Data filtering methods.&#8221.
38. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Internet Serv Appl. 2015;6(1):1–15. https://doi.org/10.1186/s13174-015-0041-5
39. Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. Proc 2015 Int Conf Cloud Comput Technol Appl CloudTech 2015. 2015. https://doi.org/10.1109/CloudTech.2015.7337020
40. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient machine learning for Big Data: a review. Big Data Res. 2015;2(3):87–93. https://doi.org/10.1016/j.bdr.2015.04.001
41. Tsai CW, Chiang MC, Ksentini A, Chen M. Metaheuristic algorithms for Healthcare: Open Issues and Challenges. Comput Electr Eng. 2016;53:421–34. https://doi.org/10.1016/j.compeleceng.2016.03.005

42. Elshawi R, Sakr S, Talia D, Trunfio P. " Big Data Res. 2018;14:1–11. https://doi.org/10.1016/j.bdr.2018.04.004. "Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service.
43. Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M. "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," *IEEE Commun. Surv. Tutorials*, vol. X, no. X, pp. 1–40, 2018, https://doi.org/10.1109/COMST.2018.2844341
44. Reimer AP, Madigan EA. Veracity in big data: how good is good enough. Health Inf J. 2019;25(4):1290–8. https://doi.org/10.1177/1460458217744369
45. Subbu KP, Vasilakos AV. "Big Data for Context Aware Computing – Perspectives and Challenges," *Big Data Res*, vol. 10, no. October, pp. 33–43, 2017, https://doi.org/10.1016/j.bdr.2017.10.002
46. "Big Data overview., Use cases, technology and opportunities. Presented at Everis by Wilson Lucas slide 23 of 25 on the 11th of April 2013.pdf.".
47. Furht B, Villanustre F. *Big Data Technologies and Applications*.
48. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104–16. https://doi.org/10.1016/j.csbj.2016.12.005
49. "Adoption-of-Big. -Data-2015-2017-and-By-Industry.&#8221.
50. "First Report on Facts and Figures: Updating the European Data Market Study Monitoring Tool," no. International Data Corporation (IDC) and The Lisbon Council, July. p. 167, 2018.
51. Vandana B, Kumar SS. "A novel approach using big data analytics to improve the crop yield in precision agriculture," *2018 3rd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2018 - Proc*, pp. 824–827, 2018, https://doi.org/10.1109/RTEICT42901.2018.9012549
52. "Top 10 Big. Data Applications Across Industries." [Online]. Available: https://www.simplilearn.com/tutorials/big-data-tutorial/big-data-applications
53. "Top 5. Industries Using Big Data Analytics To Enhance ROI _ Roosboard.&#8221.
54. Strang KD, Sun Z. Hidden big data analytics issues in the healthcare industry. Health Inf J. 2020;26(2):981–98. https://doi.org/10.1177/1460458219854603
55. Matta P, Tayal A. "Advances in Computing and Data Sciences," vol. 905, pp. 516–26, 2018, https://doi.org/10.1007/978-981-13-1810-8
56. Akter S, Wamba SF, Gunasekaran A, Dubey R, Childe SJ. How to improve firm performance using big data analytics capability and business strategy alignment? Int J Prod Econ. 2016;182:113–31. https://doi.org/10.1016/j.ijpe.2016.08.018
57. Fernández A, del Río S, Chawla NV, Herrera F. An insight into imbalanced Big Data classification: outcomes and challenges. Complex Intell Syst. 2017;3(2):105–20. https://doi.org/10.1007/s40747-017-0037-9
58. Waldherr A, Maier D, Miltner P, Günther E. Big Data, big noise: the challenge of finding issue networks on the web. Soc Sci Comput Rev. 2017;35(4):427–43. https://doi.org/10.1177/0894439316643050
59. Azzone G. Big data and public policies: Opportunities and challenges. Stat Probab Lett. 2018;136:116–20. https://doi.org/10.1016/j.spl.2018.02.022
60. Chu X, Ilyas IF, Krishnan S, Wang J. "Data Cleaning: Overview and Emerging Challenges," *SIGMOD '16 Proc. 2016 Int. Conf. Manag. Data*, pp. 2201–2206, 2016, https://doi.org/10.1145/2882903.2912574
61. Guan D, et al. Improving label noise filtering by exploiting Unlabeled Data. IEEE Access. 2018;6:11154–65. https://doi.org/10.1109/ACCESS.2018.2807779
62. Henry D. ScienceDirect Filter Filter hashtag hashtag context context through through an an original original data data cleaning cleaning method method. Procedia Comput Sci. 2018;130:464–71. https://doi.org/10.1016/j.procs.2018.04.050
63. Kenda K, Mladenić D. "Autonomous Sensor Data Cleaning in Stream Mining Setting," vol. 9, no. 2, pp. 69–79, 2018, https://doi.org/10.2478/bsrj-2018-0020
64. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and Opportunities of Big Data in Health Care: a systematic review. JMIR Med Informatics. 2016;4(4):e38. https://doi.org/10.2196/medinform.5359
65. Yang M, Kiang M, Shang W. Filtering big data from social media – building an early warning system for adverse drug reactions. J Biomed Inform. 2015;54:230–40. https://doi.org/10.1016/j.jbi.2015.01.011
66. Kumar S, Singh M. Big data analytics for healthcare industry: impact, applications, and tools. Big Data Min Anal. 2018;2(1):48–57. https://doi.org/10.26599/bdma.2018.9020031
67. Wang J, Zhang W, Shi Y, Duan S, Liu J. "Industrial Big Data Analytics: Challenges, Methodologies, and Applications," pp. 1–21, 2018, [Online]. Available: http://arxiv.org/abs/1807.01016
68. Hu Y, Duan K, Zhang Y, Hossain MS, Mizanur Rahman SM, Alelaiwi A. Simultaneously aided diagnosis model for outpatient departments via healthcare big data analytics. Multimed Tools Appl. 2018;77(3):3729–43. https://doi.org/10.1007/s11042-016-3719-1
69. "6 Reasons Why Big Data. Projects Need Search Engines _ Search Technologies.".
70. Dryden IL, Hodge DJ. Journeys in big data statistics. Stat Probab Lett. 2018;136:121–5. https://doi.org/10.1016/j.spl.2018.02.013
71. Lim C, Kim KJ, Maglio PP. Smart cities with big data: reference models, challenges, and considerations. Cities. 2018;1–14. https://doi.org/10.1016/j.cities.2018.04.011
72. Mírez-gallego S, Fernández A, García S, Chen M, Herrera F. "Big Data : Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce," *Inf. Fusion*, vol. 42, no. October 2017, pp. 51–61, 2018, https://doi.org/10.1016/j.inffus.2017.10.001
73. Torrecilla JL, Romo J. Data learning from big data. Stat Probab Lett. 2018;136:15–9. https://doi.org/10.1016/j.spl.2018.02.038
74. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and challenges of Big Data Computing in Health Sciences. Big Data Res. 2015;2(1):2–11. https://doi.org/10.1016/j.bdr.2015.02.002
75. Wang J, Yang J, Zhang J, Wang X, Chris W, Zhang. Big data driven cycle time parallel prediction for production planning in wafer manufacturing. Enterp Inf Syst. 2018;12(6):714–32. https://doi.org/10.1080/17517575.2018.1450998
76. Blazquez D, Domenech J. "Technological Forecasting & Social Change Big Data sources and methods for social and economic analyses," *Technol. Forecast. Soc. Chang*, vol. 130, no. March 2017, pp. 99–113, 2018, https://doi.org/10.1016/j.techfore.2017.07.027
77. Kapetanios G, Marcellino M, Papailias F. *Filtering techniques for big data and big data based uncertainty indexes*. 2017.

78. Kumar S, Mohbey KK. A review on big data based parallel and distributed approaches of pattern mining. J King Saud Univ - Comput Inf Sci no xxxx. 2019. https://doi.org/10.1016/j.jksuci.2019.09.006
79. Shu H. Geo-spatial Information Science Big data analytics : six techniques. Geo-spatial Inf Sci. 2016;5020:1–10. https://doi.org/10.1080/10095020.2016.1182307
80. C. STAMFORD, "Gartner Forecasts Worldwide IT Spending to Exceed $4 Trillion in 2022," *Gartner*. 2021, [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2022-04-06-gartner-forecasts-worldwide-it-spending-to-reach-4-point-four-trillion-in-2022
81. Sun Z, Strang KD, Pambel F. Privacy and security in the big data paradigm. J Comput Inf Syst. 2020;60(2):146–55. https://doi.org/10.1080/08874417.2017.1418631
82. Subudhi BN, Rout DK, Ghosh A. Big data analytics for video surveillance. Multimed Tools Appl. 2019;78(18):26129–62. https://doi.org/10.1007/s11042-019-07793-w
83. Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. Big Data Res. 2015;2(2):59–64. https://doi.org/10.1016/j.bdr.2015.01.006
84. Tseng F-H, Cho H-H, Wu H-T. "Applying Big Data for Intelligent Agriculture-Based Crop Selection Analysis," *IEEE Access*, vol. 7, no. August 2019, pp. 116965–116974, 2019, https://doi.org/10.1109/access.2019.2935564
85. Sun N, Sun B, Denny J, Lin, Wu MYC. "Lossless Pruned Naive Bayes for Big Data Classifications," *Big Data Res*, vol. 14, no. May, pp. 27–36, 2018, https://doi.org/10.1016/j.bdr.2018.05.007
86. Islam MM, Razzaque MA, Hassan MM, Ismail WN, Song B. Mobile Cloud-Based Big Healthcare Data Processing in Smart Cities. IEEE Access. 2017;5:11887–99. https://doi.org/10.1109/ACCESS.2017.2707439
87. Rathore MM, Paul A, Ahmad A, Jeon G. IoT-based big data: from smart city towards next generation super city planning. Int J Semant Web Inf Syst. 2017;13(1):28–47. https://doi.org/10.4018/IJSWIS.2017010103
88. Chianese A, Piccialli F. "Designing a smart museum: When cultural heritage joins IoT," *Proc. – 2014 8th Int. Conf. Next Gener. Mob. Appl. Serv. Technol. NGMAST 2014*, pp. 300–306, 2014, https://doi.org/10.1109/NGMAST.2014.21
89. Osman AMS. A novel big data analytics framework for smart cities. Futur Gener Comput Syst. 2019;91:620–33. https://doi.org/10.1016/j.future.2018.06.046
90. Gardiner A, Aasheim C, Rutner P, Williams S. Skill requirements in Big Data: a content analysis of job advertisements. J Comput Inf Syst. 2018;58(4):374–84. https://doi.org/10.1080/08874417.2017.1289354
91. Kim G-H, Trimi S, Chung J-H. Big-data applications in the government sector. Commun ACM. 2014;57(3):78–85. https://doi.org/10.1145/2500873
92. Akter S, Wamba SF. Big data analytics in E-commerce : a systematic review and agenda for future research. Electron Mark. 2016;173–94. https://doi.org/10.1007/s12525-016-0219-0
93. Platforms GEI. "The Rise of Industrial Big Data," *Whitepaper*, p. 6, 2012, [Online]. Available: http://www.geautomation.com/download/rise-industrial-big-data
94. Tiwari S, Wee HM, Daryanto Y. "Computers & Industrial Engineering Big data analytics in supply chain management between 2010 and 2016 : Insights to industries," *Comput. Ind. Eng*, vol. 115, no. October 2017, pp. 319–330, 2018, https://doi.org/10.1016/j.cie.2017.11.017
95. Wang Y, Hajli N. Exploring the path to big data analytics success in healthcare. J Bus Res. 2016. https://doi.org/10.1016/j.jbusres.2016.08.002
96. Clarke M. Big Data in Transport. Inst Eng Technol Sect Insights. 2016;1–70. https://doi.org/10.1057/9781137378972
97. "5 Ways. Big Data Plays a Major Role in the Media and Entertainment.&#8221.
98. "Big Data. in Media and Entertainment | Qubole.&#8221.
99. Verma OP, Jain N, Pal SK. Design and analysis of an optimal ECC algorithm with effective access control mechanism for big data. Multimed Tools Appl. 2020;79:15–6. https://doi.org/10.1007/s11042-019-7677-2
100. Sun Z, Strang K, Firmin S. Business analytics-based enterprise information systems. J Comput Inf Syst. 2017;57(2):169–78. https://doi.org/10.1080/08874417.2016.1183977
101. Ilin I, Klimin A, Shaban A. "Features of Big Data approach and new opportunities of BI-systems in marketing activities," *E3S Web Conf*, vol. 110, 2019, https://doi.org/10.1051/e3sconf/201911002054
102. Ismail A, Truong HL, Kastner W. Manufacturing process data analysis pipelines: a requirements analysis and survey. J Big Data. 2019;6(1):1–26. https://doi.org/10.1186/s40537-018-0162-3
103. Park JH, Kim YB. Factors activating Big Data Adoption by Korean Firms. J Comput Inf Syst. 2019;0:1–9. https://doi.org/10.1080/08874417.2019.1631133
104. Ren S, Zhang Y, Liu Y, Sakao T, Huisingh D, Almeida CMVB. A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: a framework, challenges and future research directions. J Clean Prod. 2018. https://doi.org/10.1016/j.jclepro.2018.11.025
105. Sun S, Cegielski CG, Jia L, Hall DJ. Understanding the factors affecting the Organizational Adoption of Big Data. J Comput Inf Syst. 2018;58(3):193–203. https://doi.org/10.1080/08874417.2016.1222891
106. Balachandran BM, Prasad S. Challenges and benefits of deploying Big Data Analytics in the Cloud for Business Intelligence. Procedia Comput Sci. 2017;112:1112–22. https://doi.org/10.1016/j.procs.2017.08.138
107. Lozada N, Arias-Pérez J, Perdomo-Charry G. Big data analytics capability and co-innovation: an empirical study. Heliyon. 2019;5(10). https://doi.org/10.1016/j.heliyon.2019.e02541
108. Amankwah-amoah J, Adomako S. Computers in industry big data analytics and business failures in data-Rich environments : an organizing framework. Comput Ind. 2019;105:204–12. https://doi.org/10.1016/j.compind.2018.12.015
109. Ghasemaghaei M. Improving Organizational Performance through the Use of Big Data. J Comput Inf Syst. 2018;00(00):1–14. https://doi.org/10.1080/08874417.2018.1496805
110. "Cover Story_. Industrial big data analytics_ The present and future - ISA.&#8221.
111. "Big Data In. Manufacturing - From Data Analytics to Machine Learning.&#8221.
112. Habib M, Yaqoob I, Salah K, Imran M, Jayaraman PP, Perera C. The role of big data analytics in industrial internet of things. Futur Gener Comput Syst. 2019. https://doi.org/10.1016/j.future.2019.04.020
113. Report A. "Run Simple," 2014.
114. Yin S, Kaynak O. "Big Data for Modern Industry :," *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, 2015, https://doi.org/10.1109/JPROC.2015.2388958

115. ur Rehman MH, Yaqoob I, Salah K, Imran M, Jayaraman PP, Perera C. The role of big data analytics in industrial internet of things. Futur Gener Comput Syst. 2019;99:247–59. https://doi.org/10.1016/j.future.2019.04.020
116. Elijah O, Rahman TA, Orikumhi I, Leow CY, Hindia MN. An overview of internet of things (IoT) and data analytics in Agriculture: benefits and Challenges. IEEE Internet Things J. 2018;5(5):3758–73. https://doi.org/10.1109/JIOT.2018.2844296
117. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics : a survey. J Big Data. 2015;1–32. https://doi.org/10.1186/s40537-015-0030-3
118. Belhadi A, Zkik K, Cherrafi A, Yusof SM, El fezazi S. "Understanding Big Data Analytics for Manufacturing Processes: Insights from Literature Review and Multiple Case Studies," *Comput. Ind. Eng*, vol. 137, no. October 2018, p. 106099, 2019, https://doi.org/10.1016/j.cie.2019.106099
119. Ahmed A, Latif R, Latif S, Abbas H, Khan FA. Malicious insiders attack in IoT based Multi-Cloud e-Healthcare environment: a systematic literature review. Multimed Tools Appl. 2018;77(17):21947–65. https://doi.org/10.1007/s11042-017-5540-x
120. Sivaparthipan CB, Karthikeyan N, Karthik S. Designing statistical assessment healthcare information system for diabetics analysis using big data. Multimed Tools Appl. 2020;79:13–4. https://doi.org/10.1007/s11042-018-6648-3
121. Balan S, Conlon S. Text analysis of green supply chain practices in healthcare. J Comput Inf Syst. 2018;58(1):30–8. https://doi.org/10.1080/08874417.2016.1180654
122. Oussous A, Benjelloun FZ, Ait Lahcen A, Belfkih S. Big Data technologies: a survey. J King Saud Univ - Comput Inf Sci. 2018;30(4):431–48. https://doi.org/10.1016/j.jksuci.2017.06.001
123. "Healthcare Big. Data and the Promise of Value-Based Care.&#8221.
124. Gachet Páez D, de Buenaga M, Rodríguez E, Puertas Sánz MT, Villalba, Muñoz R, Gil. Healthy and wellbeing activities' promotion using a Big Data approach. Health Inf J. 2018;24(2):125–35. https://doi.org/10.1177/1460458216660754
125. Hong L et al. "Big Data in Health Care : what is so different about was ist so anders am Neuroenhancement ?," vol. 1, no. 2, pp. 122–35, 2018.
126. Rizwan A, et al. A review on the role of Nano-Communication in Future Healthcare Systems: a Big Data Analytics Perspective. IEEE Access. 2018;6:41903–20. https://doi.org/10.1109/ACCESS.2018.2859340
127. "Intelligence Analysis_. Telehealth As Alternative Revenue Stream - Argentum.&#8221.
128. "Fact Sheet. : Telehealth | AHA." [Online]. Available: https://www.aha.org/factsheet/telehealth
129. "Telehealth _ Telemedicine Market. - Global Opportunity Analysis and Industry Forecast (2018–2023) _ Meticulous Market Research Pvt.&#8221.
130. Dhayne H, Haque R, Kilany R, Taher Y. In search of Big Medical Data Integration Solutions - A Comprehensive Survey. IEEE Access. 2019;7:91265–90. https://doi.org/10.1109/ACCESS.2019.2927491
131. Li J, Xu L, Tang L, Wang S, Li L. Big data in tourism research: a literature review. Tour Manag. 2018;68:301–23. https://doi.org/10.1016/j.tourman.2018.03.009
132. Ridzuan F, Wan Zainon WMN. A review on data cleansing methods for big data. Procedia Comput Sci. 2019;161:731–8. https://doi.org/10.1016/j.procs.2019.11.177
133. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access. 2017;5:8869–79. https://doi.org/10.1109/ACCESS.2017.2694446
134. Sahoo PK, Mohapatra SK, Wu SL. Analyzing Healthcare Big Data with Prediction for Future Health Condition. IEEE Access. 2016;4:9786–99. https://doi.org/10.1109/ACCESS.2016.2647619
135. Harerimana G, Jang B, Kim JW, Park HK. Health big data analytics: a technology survey. IEEE Access. 2018;6:65661–78. https://doi.org/10.1109/ACCESS.2018.2878254
136. Sakr S, Elgammal A. "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services," *Big Data Res*, vol. 4, no. May, pp. 44–58, 2016, https://doi.org/10.1016/j.bdr.2016.05.002
137. Cuomo S, De Michele P, Piccialli F, Galletti A, Jung JE. IoT-based collaborative reputation system for associating visitors and artworks in a cultural scenario. Expert Syst Appl. 2017;79:101–11. https://doi.org/10.1016/j.eswa.2017.02.034
138. Bessis N, Dobre C. *Preface*, vol. 546. 2014.
139. Cleland B, et al. Insights into antidepressant prescribing using Open Health Data. Big Data Res. 2018;12:41–8. https://doi.org/10.1016/j.bdr.2018.02.002
140. "Big Data in. Healthcare - the Challenges and the Promise – NEJM Catalyst." [Online]. Available: https://catalyst.nejm.org/big-data-healthcare/
141. Batarseh FA, Latif EA. "Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare," *Big Data Res*, vol. 4, no. October, pp. 13–24, 2016, https://doi.org/10.1016/j.bdr.2015.10.001
142. Lv Z, Qiao L. Analysis of healthcare big data. Futur Gener Comput Syst. 2020;109:103–10. https://doi.org/10.1016/j.future.2020.03.039
143. Oneto L, et al. Train Delay Prediction Systems: a Big Data Analytics Perspective. Big Data Res. 2018;11:54–64. https://doi.org/10.1016/j.bdr.2017.05.002
144. Galetsi P, Katsaliaki K, Kumar S, September. 112533, 2019, https://doi.org/10.1016/j.socscimed.2019.112533
145. Data B. "Big Data in Healthcare Sector – Revolutionizing the Management of Laborious Tasks," 2017.
146. Oussous A, Benjelloun F, Ait A, Belfkih S. Big Data technologies : a survey. J King Saud Univ - Comput Inf Sci. 2018;30(4):431–48. https://doi.org/10.1016/j.jksuci.2017.06.001
147. Hossain MS, Muhammad G. Healthcare Big Data Voice Pathology Assessment Framework. IEEE Access. 2016;4:7806–15. https://doi.org/10.1109/ACCESS.2016.2626316
148. Viceconti M, Hunter P, Hose R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. IEEE J Biomed Heal Informatics. 2015;19(4):1209–15. https://doi.org/10.1109/JBHI.2015.2406883
149. Mehta N, Pandit A. "International Journal of Medical Informatics Concurrence of big data analytics and healthcare : A systematic review," *Int. J. Med. Inform*, vol. 114, no. March, pp. 57–65, 2018, https://doi.org/10.1016/j.ijmedinf.2018.03.013
150. Li L. Data quality and data cleaning in database applications. no September. 2012;U639248:1.
151. "UCI Machine Learning Repository.&#8221.
152. "Top. 15 Big Data Tools in 2018.&#8221.
153. Inoubli W, Aridhi S, Mezni H, Maddouri M, Mephu Nguifo E. An experimental survey on big data frameworks. Futur Gener Comput Syst. 2018;86:546–64. https://doi.org/10.1016/j.future.2018.04.032
154. "10. Best Big Data Analytics Tools for 2018 – DataFlair.&#8221.

155. Pashazadeh A, Navimipour NJ. "Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review," *J. Biomed. Inform*, vol. 82, no. June 2017, pp. 47–62, 2018, https://doi.org/10.1016/j.jbi.2018.03.014
156. Yu JH, Zhou ZM. Components and development in Big Data system: a survey. J Electron Sci Technol. 2019;17(1):51–72. https://doi.org/10.11989/JEST.1674-862X.80926105
157. Garima H, Gulati, Singh PK. "Clustering techniques in data mining: A comparison," *2015 Int. Conf. Comput. Sustain. Glob. Dev. INDIACom* 2015, no. March, pp. 410–415, 2015.
158. Khan S, Shakil KA, Alam M. Cloud-based big data analytics—a survey of current research and future directions. Adv Intell Syst Comput. 2018;654:595–604. https://doi.org/10.1007/978-981-10-6620-7_57

Publisher's Note