Journal on Big Data
DOI: 10.32604/jbd.2022.028363
Article

A Survey of Machine Learning for Big Data Processing

Reem Almutiri*, Sarah Alhabeeb, Sarah Alhumud and Rehan Ullah Khan

Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
*Corresponding Author: Reem Almutiri. Email: 431214358@qu.edu.sa
Received: 09 February 2022; Accepted: 05 July 2022

Abstract: Today's world is a data-driven one, with data being produced in vast amounts as a result of the rapid growth of technology that permeates every aspect of our lives. New data processing techniques must be developed and refined over time to gain meaningful insights from this vast continuous volume of produced data in various forms. Machine learning technologies provide promising solutions and potential methods for processing large quantities of data and gaining value from it. This study conducts a literature review on the application of machine learning techniques in big data processing. It provides a general overview of machine learning algorithms and techniques, a brief introduction to big data, and a discussion of related works that have used machine learning techniques in a variety of sectors to process big amounts of data. The study also discusses the challenges and issues associated with the usage of machine learning for big data.

Keywords: Machine learning; big data; processing; algorithms

1 Introduction

In the last few years, data has begun to rise exponentially until the volume of data has increased at an unprecedented rate, culminating in the expansion of "Web technologies, social media, and mobile devices" leading to big data. Twitter, for example, used to process over 70 million tweets per day, generating over 8 TB per day [1].

Social networking sites, hotel data, weather data, online shopping sites, banking, and other sources of Big Data are only a few examples [2]. It is, however, useless unless it is thoroughly and thoroughly examined. Big Data Analytics is a method for analyzing large data sets in order to get useful insights that may be applied to a range of corporate applications or to improve people's lives in general [2].

We live in an era when an incredible amount of data is being generated from various previously unseen and unheard sources. Even though technology has been developed to capture, process, and evaluate these unexpected data, numerous challenges and concerns remain. Many studies are being conducted to better comprehend and gain valuable insights from Big Data. We now deal with Big Data in every sector of research, including Basic Sciences, Applied Sciences, Engineering, Social Sciences, Bio-Medical Sciences, and so on. All of these sectors deal with large datasets, and a lot of effort is

being made to better harness and analyze Big Data, employing domains such as Machine Learning (ML), which has a lot of promise in dealing with modern data difficulties [3].

Machine learning techniques have become increasingly popular in the recent decade in a variety of vast and complicated data-intensive sectors, such as medicine, astronomy, biology, and others, since they provide potential answers for mining the data's hidden information [4].

This paper is structured as follows: Section 2 provides an overview of machine learning and its techniques. Section 3 introduces big data concepts and applications. Section 4 briefs the related work. Section 5 discusses the issues of machine learning related to the processing of big data. Finally, Section 6 presents some conclusions.

2 Machine Learning

An overview of machine learning (ML) and its various approaches, algorithms, and applications is provided in this section.

Machine learning is "a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms" [4]. It has broad applications in artificial intelligence, cognitive science, optimal control, theories of information and optimization, statistics, and other domains of science, mathematics, and engineering [5]. Machine learning research is primarily focused on creating efficient, speedy learning algorithms that generate data forecasts [6]. Nowadays, data is growing exponentially due to the technological era that enabled everyone to produce raw data using their devices. The raw data is probable to be noisy, fractured, unstructured, and contradictory [7]. Preprocessing converts that data to a format conducive to learning by cleaning, modifying, extracting, and combining it [7]. When performing some activity using training datasets to improve performance metrics, this is referred to as a learning problem [5].

Machine learning is divided into three types: supervised learning, unsupervised learning, and reinforcement learning. Labeled training data with inputs and expected outputs are required for supervised learning [4]. In contrast, labeled training data are not required in unsupervised learning, and only the inputs have to be given without the intended outputs. It is used to find hidden information in data that hasn't been labeled, like cluster analysis [6]. By examining the similarity of the input samples, the unsupervised learning algorithm classifies the sample sets into distinct clusters [8]. Reinforcement learning (RL) enables the learning from the feedback you get from the interactions with the outside world [4]. It works based on trial and error, with the learner determining which strategy to use based on the results [9]. RL algorithms are particularly effective in learning control rules without a prior knowledge when having a large group of training data, but they have some disadvantages, one of which is the significant computing cost associated with finding the ideal solution [8].

In [4], a classification of machine learning technologies was accomplished, which summarized the learning types, data processing tasks, distinction norm, and algorithms utilized for each, as shown in Table 1.

2.1 Machine Learning Methods

This section describes a few learning methods that could be useful for handling the issues of big data. The distinguishing feature of these learning methods is that they concentrate on the concept of learning instead of focusing on one particular algorithm [4].

Table 1: Comparison of machine learning technologies [4]

Learning types	Data processing tasks	Distinction norm	Learning algorithms
Supervised learning	Classification/ Regression/Estimation	Computational classifiers	Support vector machine
		Statistical classifiers	Naïve Bayes
			Hidden Markov model
			Bayesian networks
		Connectionist classifiers	Neural networks
Unsupervised learning	Clustering/Prediction	Parametric	K-means
			Gaussian mixture model
		Nonparametric	Dirichlet process mixture model
			X-means
Reinforcement learning	Decision-making	Model-free	Q-learning
			R-learning
		Model-based	TD learning
			Sarsa learning

2.1.1 Representation Learning

Learning the valuable, and meaningful representations of data is the primary goal of representational learning [5]. A reasonable-sized learned representation is able to capture an enormous number of alternative input configurations, which can considerably enhance computing and statistical efficiency [10]. The success of machine learning techniques is heavily dependent on how well the data is represented [5].

2.1.2 Deep Learning

Deep learning is an area of machine learning that relies on algorithms influenced by the function and structure of the brain, referred to as Artificial Neural Networks [7]. Unlike most classical learning techniques, which rely on shallow-structured learning architectures, deep learning primarily employs supervised and/or unsupervised algorithms in deep architectures to automatically learn hierarchical representations [11]. Deep learning has the property of improving model performance by expanding the depth or representational capacity of the model by training it with additional examples [7]. It is more appropriate to use the architectures and algorithms of deep learning to work with the variety and volume of big data analytics [12]. One of the promises of deep learning is that feature learning and hierarchical feature extraction techniques can replace the manual selection of features [8].

2.1.3 Distributed and Parallel Learning

The difficulty of learning algorithms to process vast quantities of data in a reasonable amount of time is a bottleneck; in such a case, distributed learning appears promising because it is a technique

way to scale up learning algorithms [13]. In contrast to the classical learning paradigm, where data must be collected in a database for the purpose of central processing, distributed learning allows the learning to be performed in a distributed way [14]. Meta-learning, decision rules, distributed boosting, and stacking generalization are some of the most prominent distributed and parallel machine learning approaches [5]. The essential principle is that the distributed and parallel learning algorithms should prioritize difficult-to-learn situations [5].

2.1.4 Transfer Learning

In some circumstances, collecting training data is costly or complicated. Thus, transfer learning must be used to teach high-performance learners using data from many domains [7]. Transfer learning was being presented as a way to separate functions, domains, and distributions, allowing for the extraction of information from multiple source tasks and its application to a target task [15,16]. The transfer learning solutions evaluated are data size-independent and can be used in big data situations [7]. The benefit of transfer learning is that it can cleverly utilize previously acquired knowledge to solve new issues more quickly [4].

2.1.5 Active Learning

Learning from vast volumes of unlabeled data is time consuming and challenging. Active learning seeks to solve this problem by picking a subgroup of the most critical cases for labeling [17]. The active learner's goal is to achieve exactness by employing as few labeled cases as possible, hence lowering the cost of tracking down labeled data [18]. A favorable classification performance could be achieved with less labeled samples using query algorithms that are more efficient than those used in traditional passive learning [19].

2.1.6 Kernel-Based Learning

An advance in the design of efficient non-linear learning algorithms has made kernel-based learning a particularly powerful tool for increasing processing capacity in the recent decade [20]. In kernel-based machine learning, we consider a single kernel function, rather than examining a large number of characteristics, to determine the similarity of objects or images [5]. The kernel function is combined with images and labels to create an approach of learning and obtain the intended output as a classifier [5].

3 Big Data

Big data refers to datasets that are difficult to comprehend, captured, managed, or analyzed using traditional IT, software and hardware tools in a reasonable time. In other words, Big data is described as data with a volume, acquiring speed, or format that prevents analysis effectively using relational traditional methodologies, or data that can be efficiently processed utilizing significant horizontally zoom techniques [21].

3.1 Big Data Dimensions

Big data concept can be recognized more clear by understanding the various Vs related with it. These Vs are main dimensions (challenges) faced by big data management systems. These dimensions are defined as follows:

3.1.1 Volume (Size of the Data)

The enormous quantity of data produced per second, ranging from terabytes to zettabytes. It is needed to rethink storage and processing models in order to design appropriate tools to analyze it. Distributed systems are used in big data to store and analyze data across databases all over the world [6,22–24].

3.1.2 Velocity (Speed Where Data is Produced)

This term refers to the rate at which data is generated and processed in order to meet the demands. Traditional analytics are challenged by the increased dependency on real-time data since the data is too massive and continuously in motion [6,22–25].

3.1.3 Variety (Data is Presented in Different Data Formats)

Since data can come from a variety of sources and take many different forms, the main issue is data format incompatibility. Data is now available in several formats, including structured, semi-structured, unstructured, and even complicated structured data. Because of this variety of data formats, traditional analytical methods fail to manage big data. Designing efficient techniques is needed for preparing data for analysis which requires a significant amount of time and effort [6,22–25].

3.1.4 Veracity (Data Quality)

The quality of data captured varies significantly. It shows the data's biases, noise, abnormalities, and so forth. This will have an impact on the analysis's accuracy. Maintaining veracity will not accumulate defective data in the system. The value can be influenced by the veracity [6,22–25].

3.1.5 Variability

Variability was introduced by SAS as a new dimension of big data. The term "variability" indicates the variation of data flow rates. The velocity of big data is often irregular, with peaks and troughs periodically [25].

3.1.6 Validity

The terms "validity of data" and "veracity of data" are often used similarly. They are not the same concept, yet they are similar. Validity refers to the data's correctness and accuracy with regard to its intended use. To put it another way, data may not have any issues with veracity, but it may not be valid if it is not comprehended [25].

3.1.7 Volatility

When it comes to the big data volatility, It may be easily recalled the structured data retention policy used every day in organizations. It may be easily destroyed once the retention term has expired [25].

3.1.8 Value

Value was presented by Oracle as a defining characteristic of big data. The term "value" indicates the valuable knowledge gained from the data. It is known the data is significant on a deep level. However, the significance must be drilled [25].

In Big Data, there are eight V's available in general. Note that these V's are not fixed; they may change in the near future.

3.2 Machine Learning Tools for Big Data

The majority of existing tools are oriented toward the processing of stream, analysis interactively, and processing of the batch. Some tools that are currently used for the analysis of big data are reviewed in this section.

3.2.1 Apache MapReduce and Hadoop

Hadoop and MapReduce are not interchangeable terms; Hadoop is essentially MapReduce concept implementation [26]. MapReduce is a model that uses the divide-and-conquer technique to process large amounts of data. Hadoop is made up of two nodes: a master and a worker, whereas MapReduce executes two primary steps: Map and Reduce. The Master Node separates the incoming data into subproblems, which are then in the Map step assigned to worker nodes. The outputs of all the subproblems are then combined in the Reduce step by the master node [26].

3.2.2 Spark

It's a processing engine for in-memory data that's designed for advanced and fast analytics. It is used for increasing performance from the bottom-up scenario. Because of in-memory computing and other enhancements, Spark is 100× quicker than Hadoop in terms of performance, particularly for large-scale data processing. When data is saved on disk, Apache Spark is also fast. It now maintains the world's record for on-disk sorting on a large scale. Spark provides a general middleware layer that reimplements current learning tasks to execute on a big data environment. A middleware layer like this usually involves general operations and primitives which are beneficial for a variety of tasks of learning [7,26].

3.2.3 Storm

It is a software that allows for real-time distributed computing. It is simple to install and use. Any programming language can be used with it. It is fault-tolerant and scalable [26].

3.2.4 Apache Mink

Apache Mink is a processing engine for stream designed for distributed and high-performance computation. Even with late-arriving data, it performs accurately. It's simple to scale to thousands of nodes while maintaining excellent latency and throughput [26].

3.2.5 H2O

H2O is the fastest processing engine for in-memory data, and it's utilized for analyzing predictive of big data. It is distributed, scalable, and open source software that can operate across several nodes [26].

The supported language, execution model, associated machine learning tools, fault tolerance, and latency are all taken into account while evaluating these tools.

3.3 Applications of Big Data

Big data made its presence felt in various fields. It has been utilized in media, entertainment, communication, healthcare, government service, education, insurance, wholesale commerce, marketing, transportation, utilities, energy, natural resources, and manufacturing, among other fields.

3.3.1 Medical big Data Applications

Big data is used in medical care to efficiently store, process, query, and analyze medical data. The healthcare industry will be significantly impacted by medical big data applications. It can be clinical trial data analysis, analysis of disease patterns, patient care analysis and quality, and medication research and development, and so on.

Mount Sinai Medical Center in New York, for example, uses Ayasdi's big data tools to examine all Escherichia coli genetic sequences, including nearly a million DNA variations, in order to figure out why some bacteria types are antibiotic-resistant. To analyze data features, Ayasdi employs analysis of topological data, a new mathematical research approach [21].

Genomic data, electronic medical records, instruments for monitoring medical care, and sensory devices that can be worn are all sources of big data in health care [23,27].

3.3.2 Big Data Applications in Online Social Networks

Some big data applications in social networking services are network public opinion analysis, network intelligence collection and analysis, socialized marketing, and government decision-making support.

Some big data applications in social networking services are support for government decision-making, social marketing, examination of public opinion on the network, and collecting and analyzing network intelligence.

Instant chats, online social, micro blogging, and shared space, among other sources of big data for online social networking services, reflect various user behaviors.

The Santa Cruz Police Department in the United States experimented with predictive analysis using data. The department of police may discover crime modes and patterns, as well as estimate crime rates in large cities by analyzing social media [21].

3.3.3 Big Data Applications in Education

Department of education of United States uses big data to assess performance of learner. 'Click patterns' of students are tracked to see how much time they spend on each topic. The enactment of a trainer can be measured in terms of the number of students, the subject given, and locations, among other things [23,27].

3.3.4 Big Data Application in Enterprises

Big data can help businesses improve their manufacturing efficiency and competition in a many areas:

E-Commerce

To generate in-depth consumer profiles, businesses evaluate customer data as well as behavioral data. These profiles may be useful for creating content for a variety of target audiences, recommending material on request, and tracking the quality of content.

"Spotify" collects data of consumer behavior and analyzes it with big data Hadoop tools in order to provide accurate recommendations of music [23,27].

The Taobao Data Cube is a big data tool on the Taobao platform that allows vendors to keep track of the Taobao platform's macro industrial status, market conditions for their brands, and consumer behavior, among other things, and make production and inventory decisions appropriately [21].

Financial and Securities

Big data is used in the financial sector for analytics to aid in pre-trade decision-making, scoring and analysis, credit risk, predictive analytics, and sentiment measurement, among other things. Big data analytics is also used to look at demand enterprise risk management, mitigation of fraud, and anti-money laundering [23,27].

China Merchants Bank (CMB), for example, uses data analysis to discover activities like "Multi-times score accumulation" and "score exchange in shops" are successful at attracting high-quality clients [21].

Logistics

Logistic companies may have extensive experience with the use of Big data from the Internet of Things (IoT). UPS trucks are incorporated with GPS, wireless adapters, and sensors so that company's headquarters can monitor truck locations and avoid engine breakdowns. For now, this technology aids UPS in the supervision and management of its staff, as well as the optimization of delivery routes. UPS trucks' optimal delivery routes are determined on their previous driving experience. UPS drivers drove almost 48.28 million kilometers less in 2011 [21].

4 Related Work

Related studies on the processing of large data using machine learning techniques will be discussed in this section.

In [28], the researchers' recent works on industrial-scale Machine Learning solutions yielded a set of concepts and methodologies, which were addressed in this paper. These principles and strategies cover the entire spectrum of huge Machine Learning systems and architectures, with the goal of learning how to make them efficient, widely usable, and able to scale and grow [28].

A review study was conducted in [29] that focused on three types of interactions that have been made to connect the Machine Learning (ML) and nanoscience communities, including the usage of ML on large nanoscience data sets to analyze and extract new insights from them, the application of ML for accelerating material discovery, such as using active learning to lead experimental design, and finally the usage of the nanoscience of memristive devices in order to realize hardware customized for ML. They concluded with a discussion of the obstacles and prospects for nanoscience-machine learning cooperation in the future.

The impact of two pioneer dimensionality reduction techniques, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), on four machine learning algorithms (Decision Tree

Induction, Support Vector Machine (SVM), Naive Bayes Classifier, and Random Forest Classifier) was explored in [30]. Experiments demonstrated that when dealing with large-dimensional datasets, machine learning methods based on PCA outperform LDA. It was shown that when a dataset's dimensionality is low, ML techniques without dimensionality reduction produce better results. Also, without utilizing dimensionality reduction, Decision Tree and Random Forest classifiers surpass the other two techniques, as well as PCA and LDA [30].

Machine learning and big data are capable of making a lot effort in the fight against COVID-19, like as analyzing epidemiological models, creating interactive dashboards, and recommending the optimal vehicles for getting treatments of virus [31]. Ababneh et al. in [31] demonstrated that using reinforcement learning to analyze big data provides effective and tremendous results, despite a number of challenges and restrictions.

For COVID-19 epidemiological data analyzing and processing, Leung et al. in [32] reveal a big data and machine learning analytic tool. For effective big data analytics, the tool makes effective use of OLAP and taxonomy to generalize some specialized features into some general attributes. This tool uses common patterns seen in historical data to train a supervised learning model that predicts medical outputs (e.g., dead or recovered from COVID-19) for new dataset. The findings of the evaluation show that the tool is useful in giving a wealth of knowledge on the features of cases of COVID-19. This aids epidemiologists, researchers, and policymakers in gaining a greater perspective of the disease, that may motivate them to explore new ways for detecting, controlling, and combating it [32].

The researchers in [33] conducted a systematic analysis of COVID-19 prediction using big data and machine learning, taking into account all key fundamental characteristics (with the exception of the data sets of CT scan and X-Ray Images) and all other available, correlated literature from around the world. The accuracy of algorithms' predictions was demonstrated, with some algorithms creating inverted numbers and others providing accurate predictions with less errors. The study used two classification methods for COVID-19 cases in India from January 30 to May 30, 2020, as well as a state-by-state population index. Furthermore, the results for both strategies were substantially comparable. They came to the conclusion that anticipating COVID-19's future diegesis can improve medical decision-making, particularly when rapid intervention is required.

The function of big data analysis in healthcare was examined in [34], and many flaws in typical machine learning methods were identified. Over the next decade, machine learning and big data will collaborate to better every aspect of the healthcare sector.

Machine learning techniques are improving breast cancer prior detection from a large data of clinical and expression of the gene [26]. Gupta et al. in [26] conducted research in the area of breast cancer on the application of data analytic frameworks, tools, techniques, and machine learning methods, specifically in the areas of cancer recurrence, cancer survivorship, cancer detection and prediction. They found that artificial neural network (ANN) and support vector machine (SVM) are two of the most often utilized machine learning methods for detecting breast cancer. Apache Spark has been confirmed to work with a lot of frameworks of machine learning (ML).

Big data can predict an upcoming risk of diabetes based on a dataset and deliver treatment accordingly [35]. Saxena in [35] studied the prediction model of machine learning algorithms and awareness on diabetes. In comparison to other ML algorithms, the Random Forest (RF) and Support Vector Machine (SVM) algorithms provide better prediction results.

The researchers presented an evidence-based survey of big data and machine learning (ML) applications in environment and water management (EWM) in this paper [36], with a focus on deep

learning (DL). The goal of this survey was to look at the potential and benefits of data-driven research in EWM, provide a summary of key concepts and methods in big data and machine learning, provide a systematic review of current applications, and finally discuss key issues and challenges, as well as suggest future research directions.

CC was implemented with dynamic feature selection decomposition in [37]. The researcher used CC to propose a random feature-pooling technique for feature selection and evaluated the performance of six ML classifiers in seven different data sets. His research revealed that the feature selection method has no major impact on classifier performance. It also looked at the impact of feature selection on various datasets, including those with a large number of samples but few features, as well as those with a small number of samples but numerous features. The effectiveness of the suggested CC-based content is determined by a comparison of classifier performance in terms of accuracy, sensitivity, and specificity. Although a thorough issue analysis technique for feature selection may lead to the greatest performance, the researcher believes that a proper decomposition strategy alone is insufficient to fulfill the full potential of CC-based approaches. A proper optimizer to develop each subpopulation and a proper collaborative technology to build the whole solution file are also required for CC-based techniques.

CatBoost is a machine learning suite technology from the GBDT family. CatBoost has been utilized successfully in machine learning projects utilizing huge data since its launch in late 2018. In light of this, researchers in [38] have looked at CatBoost research as it applies to big data, gaining best practices from studies that show CatBoost to be superior to other technologies as well as those that show CatBoost to be inferior. CatBoost is also well suited for machine learning applications requiring categorical and heterogeneous data because it is a decision tree-based approach.

The researchers in [39] offered information on the use of big data and machine learning in agriculture, highlighted problems and adjustments, created architectures for these systems, and conducted a systematic literature review (SLR) that allowed them to examine 34 real-world agricultural examples. The findings showed that, thanks to cloud technologies, processing vast amounts of data is no longer a problem. Due to a lack of control over the data in its various stages, raw, semi-processed, and processed (value data), and information visualization systems, which support technical data that farmers do not typically understand, processing speed remains a difficulty.

In [40], the challenges and opportunities posed by sensing technologies in terms of helping animal keep ers to produce more meat and animal products are revealed. More specifically, this paper explores the role of sensors, big data, artificial information, and machine learning in helping animal breeders lower production costs, increase efficiencies, enhance animal welfare and raise more animals per hectare. It also explores the challenges and limitations of technology. The researchers cited the many applications of animal husbandry technology to understand its value in helping farmers improve animal health, increase profits and reduce environmental impact.

Reference [8] focused on the role of big data analytics and machine learning techniques in the area of smart building. It gave a thorough review of research papers on machine learning and big data applications, specifically for developing smart services and infrastructure.

Many research carried out computation-based experiments and a comparison study of many machine learning algorithms used to the problem of image classification. One of them is [41]. Its result is that the deep convolutional network has the best performance. It has less error rate compared with Logistic regression, Multilayer perceptron, and Stacked de-noising auto-encoders. The convolutional neural network running on GPU provided the most considerable time reduction. This is also because of NVidia's cudnn library, which is tuned for operations with convolutions [41].

In linear accelerator sources, individual terahertz pulses' spectral form and temporal delay can be properly predicted using straight-forward machine learning algorithms [42]. On heterogeneous processors like FPGAs and GPUs, these algorithms can process a massive amount of data to train deep and special Artificial Neural Networks (ANN) [42].

As a consequence of the urgent need for huge astronomical data, machine learning has become extensively employed to meet a variety of data processes, including data classification, prediction, and archiving [43]. Recurrent neural network (RNN) was widely observed to be particularly efficient for time series analysis due to its ability to mine interactions and connections between different time points [43].

One application of big data processing techniques and machine learning algorithms is detecting the attack on the network and anomalies. Reference [44] presented an approach for detecting the attack on the network and anomalies in cybersecurity datasets using a mixture of machine learning algorithms and big data processing techniques. Two separate datasets were used to verify the proposed approach for detecting the attack on the network and anomalies. The first set was created in a mobile IoT network and has 7,009,270 instances. The second set is the CICIDS2017 dataset containing over 500,000 elements and representing two kinds of attacks: port scanning and DDoS. The core of this approach is using principal components analysis for analyzing a preliminary dataset, reducing the challenge to a problem of object classification, and a mixture of machine learning techniques (k-nearest neighbors, support vector machines, linear regression, Gaussian nave Bayes, decision tree, and two-layer perceptron) to build and share classifiers.

5 Methodology

The methodology followed in this research is described in this section. The authors searched well-known databases such as IEEE, SpringerLink, Google Scholar, and others for related papers. "Machine Learning", "Big Data", "Algorithms", and "Processing" were among the search Keywords. The authors used two methods to extract data: 1. Research extraction and 2. Research screening. The authors eliminated numerous papers based on irrelevant titles throughout the research extraction process, leaving 51 papers. After that, the authors read each paper's abstract and conclusion. As a result, some papers that were outside the scope of the study were excluded. A thorough review of the aspects described in the literature in relation to machine learning used for the purpose of processing big data was undertaken.

6 Discussion

This section addresses the most important concerns of machine learning approaches for processing big data.

6.1 Issues\Challenges of Machine Learning for Big Data

We give a thorough scenario in Fig. 1 that includes a review of challenges related to machine learning techniques for big data from several angles. It includes (i) learning for massively scaled data, (ii) learning for various structured data, (iii) learning for high-speed streamed data, (iv) learning for uncertain and incomplete data, and (v) learning for low-value density data.

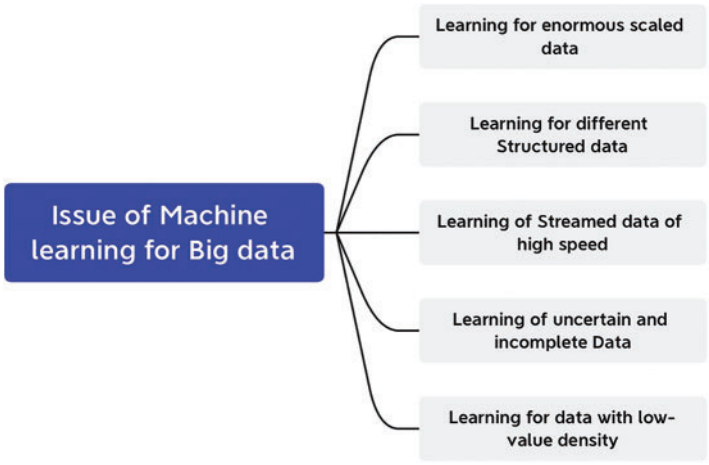

Figure 1: Learning methods for big data

6.1.1 Learning for Enormous Big Data

The amount of data we deal with is growing by the day, thanks to technological advancements. It was discovered in November 2017 that Google processes around 25 petabytes of data every day, and that firms would eventually cross petabytes of data [4]. The volume of data is clearly the major attribute of big data, which poses a significant issue [4,45]. Distributed and parallel frameworks computing should be preferred to tackle this difficulty [4].

6.1.2 Learning for Different Structured Data

There is a great variety of data nowadays. The three types of data that might result in heterogeneous, non-linear, and high-dimensional data are structured, unstructured, and semi-structured data [12]. Learning from this large data set is a huge challenge and leads to the increased complexity of the data. As a result, data integration will be required to overcome this obstacle [4].

6.1.3 Learning for High-Speed Streamed Data

There are a variety of activities that need the completion of work within a specific time frame. The speed of big data is one of its most important properties [45]. If a work is not completed within a specific amount of time, the processing results may lose their value, if not useless [4]. For example, stock market forecasting, earthquake prediction, and so on [4]. As a result, processing massive amounts of data on time quickly is a crucial and difficult task. An online learning strategy should be used to overcome the difficulties [4].

6.1.4 Learning of Uncertain and Incomplete Data

Previously, data was delivered to machine learning algorithms that was more accurate. Because the results were correct at the time. However, today's data is obscured by the fact that it is derived from a variety of sources that are both imprecise and incomplete. As a result, in large data analytics, obscurity is a significant issue for machine learning [4]. To underline the importance of addressing and managing the uncertainty and incompleteness of data quality, we list veracity as the fourth major issue for learning with big data [4]. In wireless networks, for instance, uncertain data is data created as

a result of noise, fading, shadowing, and other factors [4,46]. A distribution-based method should be employed to overcome this difficulty [4].

6.1.5 Learning of Low-Value Density Data

Machine learning is mostly used in big data analytics to extract meaningful information from massive amounts of data for commercial gain. The value of data is one of its most essential characteristics [4]. Finding meaningful value from massive a large volume of data with a low value density is extremely difficult. So it is a big challenge for machine learning in big data analytics [4]. Data Mining tools and database knowledge discovery should be employed to tackle this difficulty [4]. These technologies come into play because they provide prospective solutions for extracting critical information from large amounts of data. The authors of [24] looked at studies on data mining techniques.

Machine Learning's various issues in big data Analytics should be handled with care. Because there are so many machine learning solutions on the market, they all require a lot of data to train. Machine learning models require structured, relevant, and accurate historical data to be trained for them to be accurate. There may be other challenges, but this is not impossible.

7 Conclusion

Machine learning is vital for addressing the challenges posed by big data and revealing concealed patterns, information, and bits of knowledge from massive data, with the goal of transforming the capability into a genuine incentive for fundamental business leadership and logical investigation. This study showed the machine learning techniques role in processing of big data. It presented a general overview of the big data as well as machine learning algorithms and techniques. Also, it discussed the related works that processed big data using machine learning techniques in various fields. Finally, it discussed the challenges and issues that come with using machine learning for the purpose of processing big data.

Acknowledgement: The authors would like to thank the Deanship of Scientific Research, Qassim University, for funding the publication of this project.

Funding Statement: This work was supported by the Deanship of Scientific Research at Qassim University.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Rao, "TechCrunch is part of the yahoo family of brands," Techcrunch.com, 2010. [Online]. Available: https://techcrunch.com/2010/09/17/twitter-seeing-6-billion-api-calls-per-day-70k-per-second/ (Accessed: 12 December 2021).

[2] S. Mittal and O. Sangwan, "Big data analytics using machine learning techniques," in *9th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, Amity University, India, 2019.

[3] R. Bhatnagar, A. Hassanien, M. Tolba, M. Elhoseny and M. Mostafa, "Machine learning and big data processing: A technological perspective and review," *The International Conference on Advanced Machine Learning Technologies and Applications, Advances in Intelligent Systems and Computing*, Springer, Cham, vol. 723, pp. 468–478, 2018.

[4] J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, "A survey of machine learning for big data processing," *Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.
[5] S. Singh and U. Jaiswal, "Machine learning for big data: A new perspective," *International Journal of Applied Engineering Research, ISSN 0973–4562*, vol. 13, no. 5, pp. 2753–2762, 2018.
[6] S. Athmaja, M. Hanumanthappa and V. Kavitha, "A survey of machine learning algorithms for big data analytics," in *2017 Int. Conf. on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017.
[7] K. Divya, P. Bhargavi and S. Jyothi, "Machine learning algorithms in big data analytics," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 1, pp. 63–70, 2018.
[8] B. Qolomany, "Leveraging machine learning and big data for smart buildings: A comprehensive survey," *IEEE Access*, vol. 7, pp. 90316–0356, 2019.
[9] R. Senthil, B. Narayanan and K. Velmurugan, "A big data analytics literature survey using machine learning algorithms," *International Journal of Computer Science and Software Engineering*, vol. 9, no. 7, pp. 39–42, 2020.
[10] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[11] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2011.
[12] L. Wang and C. Alexander, "Machine learning in big data," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 1, no. 2, pp. 52–61, 2016.
[13] D. Peteiro-Barral and B. Guijarro-Berdiñas, "A survey of methods for distributed machine learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 1–11, 2012.
[14] H. Zheng, S. Kulkarni and H. Poor, "Attribute-distributed learning: Models, limits, and algorithms," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 386–398, 2011.
[15] E. Xiang, B. Cao, D. Hu and Q. Yang, "Bridging domains using worldwide knowledge for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 770–783, 2010.
[16] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[17] Y. Fu, B. Li, X. Zhu and C. Zhang, "Active learning without knowing individual instance labels: A pairwise label homogeneity query approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 808–822, 2014.
[18] B. Settles, "Active learning literature survey," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 2010.
[19] M. Crawford, D. Tuia and H. Yang, "Active learning: Any value for classification of remotely sensed data?," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 593–608, 2013.
[20] G. Ding, Q. Wu, Y. Yao, J. Wang and Y. Chen, "Kernel-based learning for statistical signal processing in cognitive radio networks," *Signal Processing Magazine IEEE*, vol. 30, no. 4, pp. 126–136, 2013.
[21] M. Chen, S. Mao and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014. https://doi.org/10.1007/s11036-013-0489-0.
[22] K. Taylor-Sakyi, "Big data: Understanding big data," 2016.
[23] A. Vinothini and S. Priya, "Survey of machine learning methods for big data applications," in *Int. Conf. on Computational Intelligence in Data Science*, Chennai, India, 2017.
[24] A. Tyagi and R. G, "Machine learning with big data," in *Int. Conf. on Sustainable Computing in Science, Technology and Management (SUSCOM)*, Jaipur, India, 2019.
[25] B. Padhi, S. Nayak, and B. Biswal, "Machine learning for big data processing: A literature review," *International Journal of Innovative Research in Technology*, vol. 5, no. 7, pp. 359–368, 2018.
[26] M. Gupta and B. Gupta, "Survey of breast cancer detection using machine learning techniques in big data," *Journal of Cases on Information Technology*, vol. 21, no. 3, pp. 80–92, 2019.
[27] Vaishali, Sonika, M. Narula and T. Dhingra, "Analysis of machine learning methodologies in big data applications," *International Journal of Recent Research Aspects*, vol. 5, no. 1, pp. 157–159, 2018.

[28] E. Xing, Q. Ho, P. Xie and D. Wei, "Strategies and principles of distributed machine learning on big data," *Engineering*, vol. 2, no. 2, pp. 179–195, 2016.
[29] K. Brown, S. Brittman, N. Maccaferri, D. Jariwala and U. Celano, "Machine learning in nanoscience: Big data at small scales," *Nano Letters*, vol. 20, no. 1, pp. 2–10, 2019.
[30] G. Reddy, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020. https://doi.org/10.1109/access.2020.2980942.
[31] M. Ababneh, A. Al-Jarrah and D. Karagozlu, "The role of big data and machine learning in COVID-19," *Broad Research in Artificial Intelligence and Neuroscience*, vol. 11, no. 21, pp. 1–20, 2020.
[32] C. Leung, Y. Chen, C. Hoi, S. Shang and A. Cuzzocrea, "Machine learning and OLAP on big COVID-19 data," in *IEEE Int. Conf. on Big Data*, Atlanta, GA, USA, 2020.
[33] S. Sharma and Y. Gupta, "Predictive analysis and survey of COVID-19 using machine learning and big data," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 1, pp. 175–195, 2021.
[34] A. Reddy and M. Nagendra, "A survey big data analysis in healthcare using machine learning approach," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 2, pp. 300–307, 2019.
[35] A. Saxena, "A survey on diabetic analysis on big data and machine learning," *Journal of Critical Reviews*, vol. 7, no. 17, pp. 3642–3648, 2020.
[36] A. Sun and B. Scanlon, "How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions," *Environmental Research Letters*, vol. 14, no. 7, pp. 073001, 2019.
[37] A. Rashid, M. Ahmed, L. Sikos and P. Haskell-Dowland, "Cooperative co-evolution for feature selection in Big data with random feature grouping," *Journal of Big Data*, vol. 7, no. 1, pp. 1–42, 2020.
[38] J. Hancock and T. Khoshgoftaar, "CatBoost for big data: An interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, pp. 1–45, 2020.
[39] A. Cravero and S. Sepúlveda, "Use and adaptations of machine learning in big data—applications in real cases in agriculture," *Electronics*, vol. 10, no. 5, pp. 552, 2021.
[40] S. Neethirajan, "The role of sensors, big data and machine learning in modern animal farming," *Sensing and Bio-Sensing Research*, vol. 29, pp. 100367, 2020.
[41] S. Boranbayev, A. Nurkas, Y. Tulebayev and B. Tashtai, "Method of processing big data," *Advances in Intelligent Systems and Computing*, vol. 738, pp. 757–758, 2018.
[42] M. Bawatna and B. Green, "Studies of big data processing at linear accelerator sources using machine learning," *Advances in Intelligent Systems and Computing*, vol. 1225, pp. 450–460, 2020.
[43] L. Yan and L. Xu, "Machine learning for astronomical big data processing," *IEEE Visual Communications and Image Processing*, St. Petersburg, FL, USA, 2017.
[44] I. Kotenko, I. Saenko and A. Branitskiy, "Machine learning and big data processing for cybersecurity data analysis," *Data Science in Cybersecurity and Cyberthreat Intelligence*, vol. 177, pp. 61–85, 2020.
[45] A. L'Heureux, K. Grolinger, H. Elyamany and M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.
[46] L. Zhou, S. Pan, J. Wang and A. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.