Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 06, No. 01, March 2020.

A REVIEW ON THE SIGNIFICANCE OF MACHINE LEARNING FOR DATA ANALYSIS IN BIG DATA

Vishnu Vandana Kolisetty and Dharmendra Singh Rajput*

(Received: 2-Aug.-2019, Revised: 26-Oct.-2019, Accepted: 16-Nov.-2019)

ABSTRACT

Big data revolution is changing the lifestyle in terms of working and thinking environments through facilitating improvement in vision finding and decision-making. But, big data science's technical dilemma is that there is no knowledge that can administer and analyze large amounts of actively increasing data and pull out valuable information. As data around the world grows rapidly and its distribution with real-time processing continues, traditional tools for automated machine learning have become inadequate. However, conventional machine learning (ML) approaches have been extended to meet the needs of other applications, but with increased information or large data knowledge bases, there are significant challenges for ML algorithms for big data analysis. This paper aims to facilitate understanding the importance of ML in the analysis of large data. It contributes to understanding the implications and challenges in big data computational complexity, classification imperfection and data heterogeneity. It discusses the capability to mine value from large-scale data for decision-making and predictive analysis through data transformation and knowledge extraction. It will suggest the impact of big data on real-time data analysis and discuss the extent to which machine learning can be used to analyze large data through machine learning in big data analysis. It will also suggest the meaning and opportunity from the point of view of encouraging feature research development in the field of ML using big data.

KEYWORDS

Big data, Machine learning, Data analysis, Big data implications, Big data challenges.

1. INTRODUCTION

In today's information world, the volume of data is bursting at an extraordinary velocity with advances in "web technologies," "social media," "mobile devices" and "sensors". Due to the multiplicity of Big Data (BD), we had to rethink the implementation of automated learning algorithms in addition to data processing framework. Choosing the right tool for an individual working situation is mostly difficult, because different types of solutions may be needed while increasing the complexity of the data itself, along with that the requirements of an automated project learning may be different.

BD has tremendous potential for commercial significance in a diversity of areas, such as "healthcare," "transportation," "e-business," "power supervision" and "economic services" [1]-[3]. But, when faced with this huge amount of data, the traditional approach suffers to perform data analysis. Research performed by "ABI (Advance Business Intelligence) Research" [4] approximates that over 30 billion interconnected devices will be there for information need. These real systems can generate enormous quantity of data from numerous resources, making it complex and difficult to perform data management, processing and analysis. It is a difficult problem for several industries and organizations to incorporate today's "healthcare companies," "IT departments," "government agencies" and "research institutions". To solve such kind of problem, a separate area was created for BD science and new trends are needed for research and education efforts [5] for rapid and successful development.

BD analysis utilization and performance of Machine Learning (ML) depend on the algorithms as well as on the setting of the applied dataset that requires a lot of time-consuming operations. In fact, some systems cannot guarantee good performance without adjusting the module. BD solutions are of high performance in a short time by providing new scientific innovations that can be integrated with ML systems for decision making. In various studies, ML is believed to be an influential tool for handling BD. As presented in [3], it is similar to the relationship between BD and the ML association among the sources and individual learning. From this perspective, individuals are able to learn from the sources to deal with innovative problems. Similarly, they are able to solve new problems through learning from BD. More information on BD processing using ML can be found in [6]-[7].

V. V. Kolisetty, Research Scholar in VIT Vellore TN, India, Email: kvishnu.vandana2016@vitstudent.ac.in
*D. S. Rajput (Corresponding Author), Associate Professor, VIT Vellore TN, India, Email: dharmendrasingh@vit.ac.in

Most of the past research works described in [8]-[9] suggest that it is difficult to perform classification of BD, as it is distributed among diverse categories of data and extracting constructive knowledge from large and composite datasets is not an easy task. BD classification demands a technique that is able to manage setbacks reasoned because of the BD attributes of "volume," "velocity" and "variety" [5]. It also needs a few calculation models and procedures to efficiently categorize data utilizing suitable ML algorithms, as discussed in various proposals [5]-[9].

Current technology development includes the latest distributed file systems and ML approaches. One such technique is "Hadoop" [10], which facilitates ML deployment utilizing exterior libraries, such as the "scikit learning library", to handle BD. Many of the ML techniques in the library mostly rely on classification algorithms which might not be appropriate for BD processing. Nevertheless, several techniques, such as "decision tree learning" and "deep learning," are appropriate for BD classification and can help develop better-supervised learning skills over the coming development periods.

The rest of the paper is organized in the following sections. Section 2 discusses big data implications. Section 3 presents data transformation and knowledge extraction. Section 4 discusses machine learning in big data analysis. Section 5 shows the importance of ML's advantage in big data. Finally, Section 6 presents the conclusion of the paper.

2. BIG DATA IMPLICATIONS

The concept of BD is initially defined as high "volume," "velocity" and "variety," but later "veracity" [11] and "value" [12]-[13] have been added. The definition needs novel processing models to facilitate visibility detection, advanced decision-making and data processing. However, "value" is characterized as the needed results to handle BD [14] and not as one of the specified BD properties. The potential of BD is highlighted by definition; however, its achievement depends on the improvement of traditional approaches or the development of new methods capable of handling this data.

2.1 Challenges

The method of supervising and utilizing a large volume of data for proposing algorithms for active and proficient methods of large data can create distinctive challenges. The challenges and modern techniques currently included in BD analysis were reviewed by Chen and Zhang [15]. Jin et al. [16] addressed the importance and opportunities of the BD concept. They also presented the challenges encountered in terms of data, order and computational complexity and suggested possible solutions to these challenges.

2.1.1 Computational Complexity

One of the major challenges faced in BD computation complexity is due to a straightforward increase in data volume. As a result, when it develops into a large size, the utilization of trivial systems is expensive and even the current ML algorithms also show a significant time complexity based on various data sample features. In case of utilizing ML algorithms like "support vector machine (SVM)", complexity is faced during the training phase of "$O(m^3)$" time and "$O(m^2)$" in the space of complexity [17], where m is the iterations needed for the training samples. Thus, the impact of m will significantly influence the time and memory requirement for training BD, rendering the process impractical.

Causes of challenges are mostly classified as: "classification," "scalability" and "analysis" based on the task to perform. In terms of technological challenges, these are classified as: "computation," "communication" and "storage". Also, with increased data size, the performance of algorithms becomes additionally reliant on the structure used to store and transfer data. As a result, the data size does not only affect performance, but it also leads to the need to revise the general architecture used to implement and develop these algorithms. Thus, with all these algorithms, as the data size increases, the time required to perform the calculations can increase dramatically and the algorithm can become unusable for very large datasets.

2.1.2 Classification Imperfection

The classification process implements methods to collect input data, understand data, transform data and understand the BD environment based on hardware necessities and acceptance criteria. Ultimately,

the success of BD classification requires an understanding of modeling and algorithms. However, certain parameters affecting the classification of BD cause problems in the development of learning and classification imperfection models.

Classification imperfections are not limited to BD and have been the subject of research for more than 10 years [18]. According to experiments performed by Japkowicz and Stephen [18], the difficulty of the problem of imbalance depends on the complexity of the task, the degree of inconsistency in the classes and the total data size of the training. They recommended that the class is likely to be represented by a reasonable number of samples in a large dataset. However, an evaluation of the actual BD set is required to confirm these observations. In such a case, the complexity of BD operations is expected to increase, which can have serious consequences due to class discrepancies.

The larger the dataset, the more often it is broken, assuming that the data is evenly distributed among all classes [19]. This causes that the classification is incomplete. The performance of the ML algorithm will adversely be influenced when the dataset contains data for a class that has a variety of possible occurrences. This problem is particularly noticeable when various classes are characterized based on several samples and few are represented as extremely small numbers. As a result, in the BD context, the probability of class imbalance is high due to the size of the data. Also, because of complexity of data, the potential impact of class imbalance on the ML approach is significant.

2.1.3 Data Heterogeneity

BD analysis involves incorporating various data from multiple sources. Such data can vary depending on the data type, format, model and meaning. In practice, most real data analysis problems are caused by heterogeneous data [1], [20], different in type, structure and distribution due to the massive quantity of data composed from various sources with no class label information. For instance, in an emotion exploration activity, the data can be included as "text," "images" and "videos" collected from different social media sources. To extract knowledge from such large and unlabeled data, advanced autonomous learning technologies must have various models which are able to perform efficient integration and learning with minimum time and process complexity.

In statistics, heterogeneity defines the differences between statistical features in different datasets. These problems exist with BD as well as in small datasets, but the datasets usually contain parts from several sources. This statistical heterogeneity splits the familiar ML hypothesis that statistical features are related in an entire dataset.

In real-time applications, learning from heterogeneous sources is associated with significant challenges due to data dimensionality, multipart relationships, several structures having various objects and diverse distribution. In most cases, label learning through supervision for heterogeneous data is not presented or is time-consuming. In this case, the guidelines for heterogeneous information integration are missing and most learning methods fail to perform accurately. So, identifying an unsupervised function that will be beneficial to the overall analysis is still an important and crucial research problem.

3. DATA TRANSFORMATION AND KNOWLEDGE EXTRACTION

ML often requires data pre-processing and cleaning steps to configure the data for a particular model. However, in the case of data from different sources, the formats of the data may be different. In the context of data analysis, "data," "information" and "knowledge" are three foremost observations to be exploited. It is possible to perceive data analysis, which is able to transform and integrate data into information and can be used for visualization or decision making, as shown in Figure 1.

In an effort to optimize BD for data extraction and transformation, it is necessary to try to modify the data to become analysable by ML. This amendment process is in the pre-processing phase of the data. It also undertakes the challenges to remove dirty and noisy data through the cleaning process. In this area, there is no significant development in respect to BD and it has been an active research focus in

various domains.

The three essential aspects of data influencing ML are: large quantity, dimension and various samples. Hence, two-perceptive data for learning with BD is handled by limiting the dimension and selecting the instance. Reducing dimensionality aims to set a high-resolution space on a smaller area of dimensions

without much information loss. Dimension reduction mainly solves the problem of dimension curse and enhancement in processing.

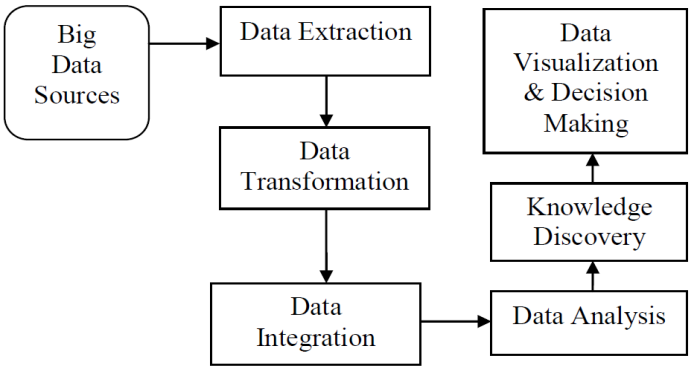

Figure 1. Big data transformation and analysis process.

The selection of instance refers to methods that select subsets, similar to the entire dataset. It is intended to reduce the dimension with large-scale datasets through data reduction and more specifically to select the required instances. The subgroup is then utilized to create conclusions regarding the entire dataset. The selection methods for various events include "random selection," "selection on genetic algorithms," "progressive sampling using domain knowledge" and "cluster sampling" [21], [46].

Data integration and management are critical issues in BD distribution. These are among the original activities utilized to advance the quality of distributed data in independent data sources. A traditional data collection system is a system that integrates the limited resources and usually has complex and time-consuming functions. As discussed in [22], data integration systems need to address uncertainties about semantic assignments between data sources and the intermediate schema in order to effectively index the keywords of the data access queries. This means that appointments are detected by understanding the meaning behind the tagging features of the elements of the schema elements, but many challenges are faced to understand the features that are reliable [23] and this is associated with BD integration [24].

3.1 Author Data Transformation

Data transformation converts data or information from one structure into another, generally from the structure of a source system to the necessitated structure of a required target data structure. Mostly, the standard procedures are engaged in converting text data files. However, during data conversion, for a while, a program is converted from one workstation execution program into another, so that the program can be executed on a diverse environment. The common motivation for this data relocation is the introduction of the latest system that is fundamentally dissimilar from the earlier systems.

In practice, data transformation engages the utilization of an exceptional program that reads the original base language of the data, determines the language in which it must be converted into a new program or the data that the system can utilize and then continues with data transformation.

Data Transformation engages two basic stages:

- ***Data mapping***: Assignment of components to capture all transformations that occur at the source base or from system to destination. This is organized for more complex systems when there are multifaceted transformations, such as multiple individuals or multiple regulations for transformation.
- ***Code generation***: Creating the original transformation program. As a result, the specification of the data map is utilized to produce carry-out programs for running on systems.

3.2 Data Analysis

Data analysis and data mining from a division of "Business Intelligence (BI)" that includes "data warehousing," "database management systems" and "online analytical processing (OLAP)". Data

mining is a specific data analytic strategy that targets predictive statistical modeling and information discovery, relatively entire expressive reasons, while BI covers data analysis aiming primarily on BI.

Deeper data analysis is able to reveal many of the most important features of data, which helps predict future data features. This allows to explore the development of patterns from a set of data to a BD set. Statistical and engineering features are key analytical bases that assist us to recognize the development of patterns. One area focused on BD classification is the development of the technology sector, where the fundamental elements of analysis should be clearly understood. Some numerical evaluations contributing to these goals are: "counting," "mean," "variance," "covariance" and "correlation" [25].

The methodologies to process data for data analysis need to follow these steps:

- ***Data requirements*:** Data is required as the input of analytics, which is particularly dependent on the needs of the analytics or clients' usage. The common individual on which data accumulated is identified as the testing unit. In particular, a demographic variable (e.g., height and weight) is obtained. Data can be statistical or definite.
- ***Data collection*:** Data is accumulated from various sources corresponding to data analysts at an organization. Data can also be gathered through sensors in the surrounding, such as "traffic cameras," "satellites," "recording devices," …etc. It can also be gathered through "interviews," "downloads from web sources" or "interpreting documents".
- ***Data processing***: The data primarily acquired must be processed or organized for analysis. For example, this might include data placed in tables and columns, such as spreadsheets or statistical software in tables and columns for further analysis.
- ***Data cleaning*:** This will process and classify data which might be imperfect, duplicate or enclosing errors. Data is accessed and stored showing the need for data cleansing from problems. Cleaning data is the process of avoiding and correcting such inaccuracies. In general, it consists of "record matching," "recognizing data incompleteness," "eminence of existing data," "transcription" and "column segmentation".

Moreover, as we have already noted in the context of BD, the challenges of data classifying and cleaning are becoming more common and more difficult. Therefore, it is difficult to identify such problems and separate them to represent a complete group. In the case of large inconsistency between data rows, the process of data selection is not able to guarantee accurate class selection solutions.

4. MACHINE LEARNING IN BIG DATA ANALYSIS

ML is a division of artificial intelligence, which consists of two phases: "training" and "testing" [3]. The primary phase proposes a learning mechanism based on some of the known characteristics of datasets. The second stage aims to make predictions of unidentified characteristics through the knowledge gained in the primary phase.

In this view, "training" and "testing" are also called "learning" and "prediction". In fact, the task of ML is to use a learning algorithm to build a model that is also applied to make predictions. Therefore, this activity is generally called predictive modeling. The phases of ML from data acquisition to constructing a predictive model are shown in Figure 2.

In recent literature, several researchers illustrated ML challenges with BD [26]-[28], while others examined them in terms of a particular methodology [26]. According to [28], ML algorithms are able to develop in numerous kinds of learning, such as "Decision Tree Learning," "Rule Learning," "Instance-based Learning," "Bayesian Learning," "Perception Learning" and "Collective Learning". All these learning algorithms reflect the nature of the promotion.

In ML, there are several algorithms for constructing the model, where the word algorithm points to the learning algorithm. In this scenario, the model is treated as information modified from training data. The testing phase aims to transform information into knowledge. The learning algorithm utilizes a given set of data to learn, validate and test the model. It discovers the best value for the parameter to validate and evaluate the enhancement.

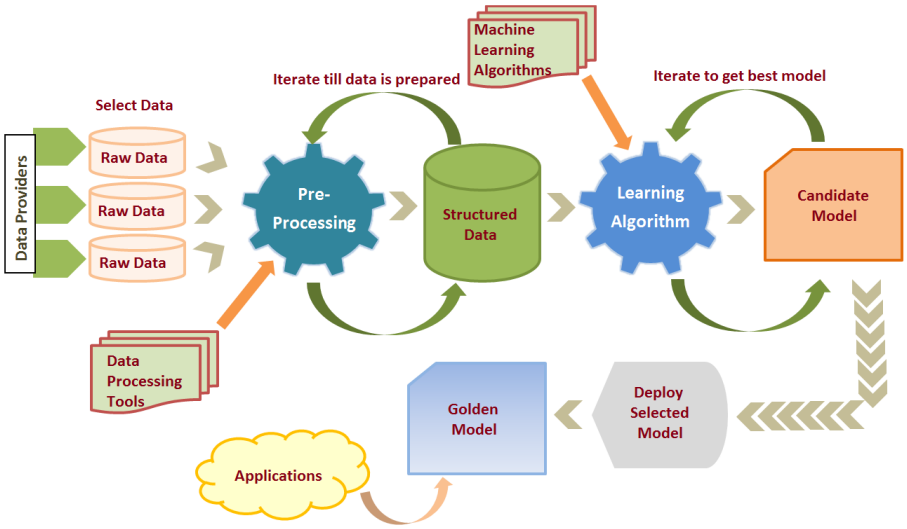

Figure 2. Machine learning phases of processing [44].

4.1 Supervised and Un-supervised Learning

Supervised Learning (SL) proposes the methods of studying with the trainer, since in all the cases, the training clusters are categorized to predict the outcomes accurately. In other words, the proposed learning is usually inspired by learners' learning under the control of supervised trainers. In doing so, the purpose of this kind of learning is to build a model by learning through accurate data and making other predictions and unrelated cases in terms of the expected attribute value. Therefore, SL can be part of the "classification" and "regression" functions for final prediction and statistical prediction, respectively [47]-[48].

In SL, classes are known and class boundaries are well defined in a given set of learning data and learning is carried out using these classes. Classification problems can be solved precisely depending on the knowledge transformations revealed above. A flowchart of supervised ML approach is shown in Figure 3.

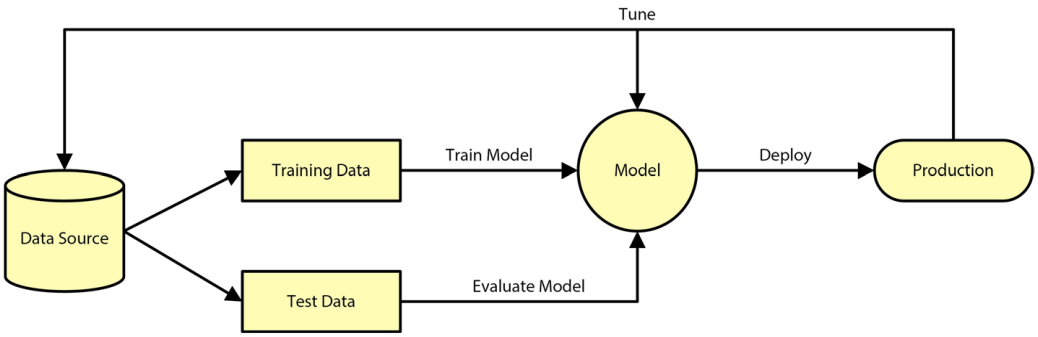

Figure 3. A flowchart of supervised ML.

Let's assume a dataset is specified and its data domain is D is R^c, which implies that the occurrences in the dataset are based on the c properties and create a "c-dimensional vector space". If it is supposed that there are "n classes," the function of knowledge can be given using Equation (1).

$$f: R^c \Rightarrow \{0, 1, 2, \dots, n\} \quad (1)$$

In Equation (1), the series from *"{0,1,2, . . . ,n}"* includes the groups of knowledge which allocate the distinct values of labels *"0,1,2, . . . ,n"* to dissimilar classes. This mathematical purpose assists to describe the classification criteria that are appropriate for data classification. A number of classification procedures have been recommended in the ML document and a few of the well recognized methods are "SVMs" [29], [52], "decision trees" [30], "random forests" [31] and "in-depth learning" [32].

Un-Supervised Learning (USL), on the other hand, means learning without learning. This is because the learning results are not clear. In other words, learning without supervision is naturally inspired by learning. In fact, the purpose of this type of learning is to discover previously unknown dataset patterns through association and cluster insertion. The first aims to identify the relationship between the objects and attributes and the second aims to cluster the items based on their similarity.

In USL, suppose that class boundaries are unknown; so, the class labels themselves have been learned as well and classes are defined accordingly. Thus, the class boundaries are statistical and not clearly described; known as "clustering". In the clustering problem [33], it is assumed that the dataset can be created, but not categorized. As a result, it can only generate approximate rules to help categorize new data that does not contain labels. Clustering forms a guideline that facilitates labelling the selected data points and assigning labels to the new data points. As an outcome, the data can simply be collected without being classified. Therefore, clustering problems are expressed using estimation rules [49].

Clustering difficulties can also be mathematically solved based on the knowledge of data transformation, as discussed previously. Let's suppose a domain "*D"* with the set of data records, having *c* depending features, can be represented as "R^c" and forms feature vectors with a c-dimension space. To construct a cluster for the k classes, a knowledge-based function can be derived as given in Equation (2).

$$f: R^c \Rightarrow \{0, 1, 2, \dots, k\} \tag{2}$$

The series of knowledge set is illustrated for *k* labels as "*{0,1,2, . . . , k},"* each label having different features. Based on these most associated features of labels, a suitable class is assigned to have accurate clusters. Few clustering algorithms in ML generally used are "*k*-Means clustering," "Gaussian mixture clustering" and "hierarchical clustering" [34].

4.2 Big Data Analysis

Business Intelligence is an application that can benefit from BD techniques. BD analyses also have systematic consequences in today's uses; hence, it is suitable to recognize them utilizing the features of the classes, the characteristics of the parameters and the characteristics of the observations; three important ideas of BD. A full understanding of the features of the classes, the characteristics of the parameters and the properties of the observations can support in addressing these problems.

Assuncao et al. [35] reviewed the development methodology and environment for performing BD analysis on the cloud platform. They categorized the BD analytics solutions "based on past customer activity -description models, "based on available data- forecast models" and "prescriptive models for supporting decision-making processes".

Personalization of acceptance and non-cooperative attempts can lead to difficulties in the BD area. Every acceptance will contribute to the BD and influence the uniqueness of the other orthogonal acceptances, thus determining acceptance problems using a three-dimensional space. This recommends that the classification of categories with BD development is very complex and unpredictable. Thus, an increase in the class forms depends on the scheme, irrespective of user knowledge and experience. Thus, BD classification becomes unpredictable and it is difficult to apply ML models and algorithms efficiently.

Similarly, the acceptance of the features contributes to BD complexity. It builds a classification utilizing the patterns to reduce complexity with growing data dimensions. These are considered as main factors that solve the scalability problem of the BD paradigm and its confirmation contributes to the complications in the data management, processing and analysis. Its expansion will increase data size and make processing difficult with current technologies in the near feature.

4.3 ML Modeling and Algorithms Approaches in Big Data

ML has different learning paradigms; however, not all these types of research are appropriate. Modeling and algorithms are defined based on the characteristics of "domain distribution," "batch learning" and "online learning" depending on the availability of data-level labeling and supervision and USL. The two foremost elements that help accomplish ML goals are aimed through learning models and learning algorithms utilizing different pattern recognition tools. Some of the tools utilized in BD for data processing are described in Table 1.

Table 1. Comparison of various BD tools.

BD Tools	**Description**	**Advantages**	**Disadvantages**
Apache Hadoop [62]	• It is one of the most prominent and used tools in the BD industry with a huge capacity for large-scale processing of data. • It processes large datasets through programming models, such as "MapReduce". • It is a 100% open-source framework and executes on product hardware in current data centers.	• It offers a robust ecosystem that best suits the analytical requirements of the developer. • It conveys elasticity and faster data processing. • Highly-scalable and highly-available service to rest in a cluster of computers. • The main strength of Hadoop is HDFS, which is capable of holding all types of data - video, images, JSN, XML and plain text in the same file system.	• Sometimes, disk space concerns are possible to be met due to its "3x" data redundancy. • I/O operations have to be optimized for better performance.
Apache Spark [63]	• It is the industry's next hype in BD tools. • It is an alternative to the MapReduce of Hadoop. • It is an added point for data analysts to handle definite kinds of data to accomplish quicker results.	• It is easy to execute on a particular local system to facilitate progress and testing. • It can run 100 times faster than a map of Hadoop. • It is easy to work with HDFS in addition to other data stores; for instance with "OpenStack Swift" or "Apache Cassandra". • The main point of this open-source BD tool is that it fills the gap in "Apache Hadoop" with regard to data processing. • It is capable of managing batch data and real-time data together. • It processes data quicker than conventional disk processing techniques because of in-memory data processing.	• It has no support for real-time processing. • It has no file management system and is expensive. • It has problems with small files. • Its number of algorithms is very few and it shows latency. • Manual optimization. • Iterative processing.
Apache Storm [64]	• It is a distributed real-time framework to consistently process unbound data streams. • Its topology can be considered as a MapReduce work. • It's a free and open-source BD computation system. • It can interfere with "Hadoop's HDFS" with adapters as required, which is an additional feature that builds it as an open-source BD tool.	• Its framework supports any programming language. • Depends on the topology configuration, it allocates the workload to the scheduler nodes. • It recommends distributed, real-time, fault-tolerant processing systems with real-time computation potential.	• Difficult to learn, use and debug. • The use of native scheduler and Nimbus becomes a hindrance.
Cassandra [65]	• It handles a distributed kind of database to process a big group of data on servers. • It is one of the best BD tools that mainly processes structured datasets.	• Its design architecture does not function as the master-slave architecture and every node functions as an identical role. • It is able to manage various synchronized clients across the data center.	• Troubleshooting and maintenance require some extra effort. • The process of clustering requires improvement.

	• It provides a highly available service with a single point of failure. • Additionally, it has specific capabilities that no other related database and no NoSQL database can provide. • It supports replication across multiple data centers, providing lower latency for users.	• Its database is extensively utilized at a moment to give valuable administration of a huge quantity of data. • The data is routinely duplicated to several nodes for fault tolerance. • It is mainly valuable for applications that are not able to lose data still if the complete data center is stopped. • It provides agreements' and services' support, provided by other vendors.	• The row-level locking feature is unavailable.
RapidMiner [66]	• It is a software stage for information science performance and presents a combined situation. • It follows the model of a client's server, where the server is perhaps situated on a pre-basis, or in cloud communication. • It is developed in Java and presents a GUI for designing and performing workflows. • It is able to give 99% of progressive solutions.	• It's an open-source BD tool. • It is utilized for "data preparation," "machine learning" and "model deployment". • It provides a collection of products for setting up the latest data mining procedures and projecting analytics. • It stores streaming data in numerous databases. • It allows for various data management approaches with batch processing and GUI Interface.	• Improvision in online data services is needed.
Hive [67]	• It is an open-source tool for BD. • It helps the developer perform BD analysis in Hadoop. • It assists to quickly find data search and manage large datasets.	• It maintains the SQL type query language for communication and data modeling. • It lets you define functions with Java or Python. • It is created to handle and search only structural data. • It's based on "SQL-inspired language" that sets the consumer apart from the complexities of map reduction programming. • It presents a "Java Database Connectivity (JDBC)" interface for programme integration.	• It is not designed for Online Transaction Processing (OTP). • It can be used for Online Analytical Processing (OLAP). • It doesn't support updating and deleting, but it supports overwriting or data capture. • Basically, the Hive subqueries are not supported.

Depending on the characteristics of the divisions of the field, "regression," "classification" and "clustering" might determine the modeling features of ML by supervised and unsupervised algorithms of ML [36]-[37]. Domain segmentation can also be essential in determining the learning algorithms. Suppose that the field is categorized and group labels are introduced, so that a classification model can be set up and the acquisition of optimal parameters can be monitored. It is therefore referred to as SL and classifications are defined under the SL model. If the field is separable and the class labels are not assigned, it is referred to as USL and then assigned to the USL format.

4.3.1 Supervised Learning Models

Models of SL provide parameters to move the data field for a response group, thus helping to take the knowledge from the data. These learning models are generally combined into predictive models and

classification models. The "regression model" is a predictive variable that is appropriate for systems that generate continuous reactions. There are various regression models, including "standard regression," "ridge regression," "lasso regression" and "elastic-net regression" [38]. In this model, the factor creates an important function in reducing the error to the incline factor and the normalization factor.

The classification model is suitable for scenarios where individual results are created. There are many classification representations that can be grouped under "mathematically intensive," "hierarchical models" and "hierarchical models". Hierarchical models assist to classify separated group points associated with base classes utilizing a tree-like structure [50]. This model is well suited for modern requirements, including BD and distributed ML. It adopts together regression analysis and classification approach using trees that can be constructed with a series of decisions, called decision trees.

4.3.2 ML Supervised Learning Algorithms

SL algorithms assist in model training effectively to provide high-grade accuracy. In general, SL algorithms support the use of large datasets to retrieve optimal values for model parameters without over-installing the model. Therefore, it is important to carefully design the learning algorithm using a systematic approach. The ML field proposes three phases of designing an SL algorithm as, "training phase," "verification phase" and "testing phase".

Training algorithms mainly help adjust and optimize model parameters using categorized datasets. The training algorithm needs "quantitative measures" to effectively train the learning model by means of the distinctive marked dataset. In general, it includes several sub-processes, such as extracting the data field and creating the associated group, standardization and modeling. Model testing is a procedure to evaluate the enhancement of a model that has been trained with a training algorithm. Few such algorithms based on training are described below.

- ***Support Vector Machine:*** This method helps in resolving one of the BD classification issues in a classic ML technology. Specifically, it can help in multiple domain applications in the BD environment, but it is complex during computation. It is utilized in BD frameworks, like "RHadoop," based on SVM implementation with R-programming for analyzing distributed file systems. Even for "MapReduce" in Hadoop framework SVM [53], associated algorithms are deployed to improvise the functions.
- ***Decision Tree Learning:*** Decision trees use rule-based approaches to divide domains into several linear spaces and predict reactions. If the predicted reaction is repetitive, the decision tree is a "regression tree" and if the predicted reaction is individual, the tree is a "classification tree". In fact, "decision tree-based learning" management is described as a "rule-based binary tree" creation procedure, but it is easy to recognize if it is interpreted as a hierarchical field partitioning system. The data area is recursively partitioned into two sub-domains to obtain more information gain than in the partitioned node approach. Decision trees are able to be "trained," "verified" and "assessed" exploiting SL algorithms, so it is clear that they form an SL model and satisfy this definition.
- ***Random Forest Learning:*** This learning method utilizes the decision tree modeling approach [3]. This technique utilizes a decision tree model for parameterization, which includes "sampling techniques," "sub-space techniques" and "ensemble techniques" to optimize modeling, which is generally called bootstrap modelling and is substituted with a "random sampling method". Based on this, it supports to construct and choose a decision tree for random forest configuration. This decision tree can be either in the form of a "classification tree" or a "regression tree". Hence, it can be mutually useful for classification and regression issues.
- ***Deep Learning Models:*** Deep learning models in ML try to understand the relations embedded in learning representation. This is mostly expressed by the frequently used term "learning by features" [40]. This kind of algorithm takes its name from the reality that it utilizes data representation rather than precise data functions to execute jobs. It transforms data into abstract illustrations that facilitate learning. In a deep-learning structure, these presentations will later be used to perform ML tasks. Since the functions are discovered directly from the data, the parameters do not need to be configured. In the BD context, the ability to avoid technical features is an immense benefit because of the challenges correlated with this process.

Deep-learning algorithms able to confine different stages of abstraction. This type of learning is, therefore, the best clarification to the "image classification" and "recognition problem". "Boltzmann machines" [41] are related with the exception that they use a random rather than an inevitable process. Another example of these algorithms is "deep-belief networks" [42]. Because of the illustrated features, deep learning appears to be well suitable to handle several predefined challenges, such as "geometry features," "data heterogeneity," "nonlinearity" and "noisy data". However, these algorithms are not designed primarily for varying and volume data learning [43] and therefore prone to the data speed problem. While they are well-suited to handling large amounts of data with complex problems, they are not computationally efficient [45].

Najafabadi et al. [26] focused on deep learning, but pointed out the common disadvantages of ML with BD: "unstructured data formats," "fast data streaming," "multi-source data entry," "noisy" and "bad data," "high dimensions," "scalability of algorithms," "unbalanced input data" and "limited labeled data". Similarly, Sukumar [27] recognized three main prerequisites: "designing flexible" and "highly scalable structures," "understanding the properties of statistical data" before applying algorithms and ultimately developing the capacity to work with large datasets. In Najafabadi et al. [26] and Sukumar [27], investigations reconsidered ML characteristics with BD, but did not do effort to link every acknowledged challenge.

Qiu et al. [28] developed various learning methods and presented various works of BD. Although they performed an immense job on current issues to identify possible solutions to the lack of classification as well as on approaches to solve the challenges and deepen the relationship between the hard-informed decision-making model and the learning outcomes which are most suitable for a particular task or a specific scenario. Thus, the focus of our work is to establish a link between solutions and challenges. A comparative analysis of these proposals and their limitations is presented in Table 2.

Table 2. Comparison of proposal enhancements and limitations.

Author	**Approach**	**Datasets Used**	**Enhancement**	**Limitation**
L. Xiang et al. [1]	Two-stage unsupervised multiple kernel extreme learning machine	UCI machine learning repository	Flexible algorithm for fast unsupervised heterogeneous data learning	High computational overhead
H. Liu et al. [6]	Predictive modelling, Decision tree, Bayesian and Instance-based learning	UCI and biomedical repositories	Building accurate, efficient and interpretable computational models	High-variability data showing high variance in terms of accuracy performance
I. W. Tsang et al. [17]	SVM and a core vector machine (CVM) algorithm.	KDDCUP-99 intrusion detection data	Optimal solutions for efficient classification with the use of core sets	High expense because of time and space complexities
M. Ghanavati et al. [19]	Integrated method for learning large imbalanced datasets	Water pipeline datasets	Effective for the well-learned datasets.	Ineffective for big and highly imbalanced data
C. Zhu et al. [20]	Heterogeneous metric learning with hierarchical couplings	30 datasets from different domains	Solution for complex categorical data with hierarchical coupling relationships and heterogeneities	Limited to specific data characteristics and domain knowledge

H. A. Mahmoud et al. [22]	Probabilistic model based on Naive Bayes classification	2323 schemas from 5 different domains from Google's web	The performance of the clustering algorithm shows increases in precision and recall	No comparison is shown with the existing clustering algorithms
N. Ayat et al. [24]	IFD (Integration based on Functional Dependencies) with a probabilistic data model	Dataset of the university domain	Significant performance gain in terms of recall and precision compared to the baseline approaches	No measure has been shown to enhance the integration of uncertain data
J. Read et al. [43]	Deep-learning techniques	Real-world datasets	Improvement in the accuracy of popular existing data-stream methods	No clear explanation of higher-dimensional datasets in terms of feature reduction and classification of labels

4.4 Limitations of Big Data Analytics

BD brings some big hopes. However, this is not a tool with unlimited features, making the most of the analysis means underestimating the limitations of using data capabilities [54]. The following are some of the major limitations of experienced users and first-time data explorer.

- *Data Misinterpretation*: Data can reveal the user's behavior. However, it cannot also advise why users think or behave in their ways. But, misinterpretation of data is able to misguide dealers in their business attempting to capture utilizing the market progressive information. In addition, depending exclusively on data to formulate possibility may guide companies to take actions based on wrong relevance. The actuality is that identifying the predicted correlation and attempting to respond to the correct problem in support of the data is a different job from gathering and interpretation the data.
- *Security Limitations*: BD is also facing limitations due to security issues. Companies that collect data have a significant responsibility to protect data. The consequences of data breaches may include litigation, fines and loss of reputation. Security issues can greatly inhibit your ability to process data. For example, analyzing data by other organizations can be complicated, since the data might be concealed with a firewall or private cloud server. This creates a lot of trouble for sharing and transmitting data to be analyzed and worked on in a reliable manner.
- *Outlier Effect*: The third major constraint with BD is that outliers are common. Once the data is processed and analyzed, the user's failure or a new upgrade to the popular search engine will produce some biased results. The reality is that technology is not yet able to collect data completely accurately. However, Google's own algorithms and the inability to correctly predict search behavior made the project one of the company's most compelling failures to date.

5. FEATURE SIGNIFICANCE OF ML IN BIG DATA

ML utilizes an algorithm to discover hidden knowledge without explicit programming. In ML, it is important to understand repetitive components, where the models tend to adapt independently when exposed to BD. So, with the advent of new computer technologies, ML has significantly advanced from the past. Recently, ML algorithms have been able to consistently perform complex computations to integrate and analyze BD, which has not been available for a long time. A few well-known examples are illustrated below.

- The concentrate of ML with BD able to be found in "Google's self-driving car".
- ML applications utilizing BD are able to find various "recommendations" and "online business systems", such as Amazon, Netflix online, …etc.
- In-text data processing in various social media information, like Facebook, Twitter, …etc.
- ML can process BD to predict fraud detection in various financial and security systems.

The most commonly used ML methods include "SL," "USL," "class supervised learning" and "reinforcement learning". However, SL-based method utilization is nearly 70%, whereas USL utilization is about 10-20%.

- The significance of the SL algorithm is to be utilized where the required result is well-known. This algorithm is used with a set of inputs and a corresponding set of outputs. The algorithm compares the actual output with the correct output in feature analysis.
- Unsupervised learning is utilized for data without historical label. The algorithm must know that it is displayed to give the correct result and semi-SL is utilized for labeled and unmarked data, such as "classification," "regression" and "prediction".
- USL is utilized in opposition to data with no past labels. This algorithm must predict the correct result without knowing the data labels.

5.1 Future Research Directions of Big Data Analysis

Today, BD analysis is getting more and more attention, but there are still many research problems to be solved in various domains.

- *Storage and Retrieval*: Multidimensional data has to be integrated with the analysis on BD; so, you can explore arrays depending on in-memory illustration models [55]. The incorporation of multidimensional data representations on BD involves the use of multidimensional extensions to enhance the query language HiveQL. With the rapid development of smartphones, images, audios and videos are produced at an alarming rate. However, the storage, retrieval and processing of this unstructured data require extensive research in various dimensions [56].
- *BD Computations*: In addition to the current BD paradigms, such as "MapReduce" [51], other paradigms are relevant, such as "YarcData (BD Graph Analytics)" and "high-performance computing cluster (HPCC)" systems need to be investigated [57].
- *Visualization of High-Dimensional Data*: Visualization facilitates assessment analysis at each action of data analysis. It concerns the remaining fraction of the "data warehousing" and "OLAP" research. For high-dimensional data, a various range of visualization tools is being developed [58].
- *Real-time Processing Algorithms*: Due to the frequency at which data and forecasts are produced, the various real-time algorithms might not be able to realize the processing time complexity and delay.
- *Social Perspective's Dimension*: It's essential to recognize that various technologies are able to produce quicker results, but assessment makers must utilize them intelligently [59]. These outcomes may possibly have some social and cultural influences. There is no doubt that large-scale search data will assist in generating improved tools and facilities, as well as privacy intrusion and intrusive marketing. Data analysis assists even in understanding online behavior, local community and political movements [60]-[61].

6. CONCLUSION

BD analysis is the process of examining large and diverse datasets. Learning from large, unstructured data offers significant opportunities for many sectors. However, most of these routines are not sufficiently computational, practical or scalable. This paper presents a review of the need for research aimed at proposing new techniques that can be used to analyze BD. The concept of ML is increasingly adopted in current and future trends in BD implementation. This paper presents the challenges faced by various ML tools to provide an adaptable framework that fits the BD field of analysis. Analytical units can be combined with an ML engine to overcome data processing conditions. BD analytics and ML

implementation support each other and can be powerful tools for understanding and predicting business behavior based on customer input information. With increasing use of ML concepts in research and business, the requirement of new methods to assist learning tasks has become increasingly essential in future research works to achieve significant improvements in ML approaches for BD analysis.

REFERENCES

[1] L. Xiang, G. Zhao, Q. Li, W. Hao and F. Li, "TUMK-ELM: A Fast Unsupervised Heterogeneous Data Learning Approach," IEEE Access, vol. 6, pp. 35305-35315, 2018.

[2] W. Raghupathi and V. Raghupathi, "Big Data Analytics in Healthcare: Promise and Potential," Health Information Science Systems, vol. 2, no. 1, pp. 1-10, 2014.

[3] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, "Efficient Machine Learning for Big Data: A Review," Big Data Research, vol. 2, no. 3, pp. 87-93, Sep. 2015.

[4] ABI Research, "Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020," [Online], Available: https://www.abiresearch.com/press/more-than-30-billion-devices-will-wirelessly-conne/, 2013.

[5] P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise-class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011.

[6] H. Liu, A. Gegov and M. Cocea, "Unified Framework for Control of Machine Learning Tasks Towards Effective and Efficient Processing of Big Data," Springer Data Science and Big Data: An Environment of Computational Intelligence, pp. 123–140, 2017.

[7] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[8] S. Suthaharan. "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning," ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, pp. 70–73, 2014.

[9] H. Tong, "Data Classification: Algorithms and Applications," Taylor and Francis Group, pp. 275–286, 2015.

[10] K. Shvachko, H. Kuang, S. Radia and R. Chansler. "The Hadoop Distributed File System," Proc. of the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1–10, 2010.

[11] R. Narasimhan and T. Bhuvaneshwari, "Big Data - A Brief Study," International Journal of Science Eng. Research, vol. 5, no. 9, pp. 350-353, 2014.

[12] W. Fan and A. Bifet, "Mining Big Data: Current Status and Forecast to the Future," SIGKDD Explorations Newslett., vol. 14, no. 2, pp. 1-5, Dec. 2012.

[13] Y. Demchenko, P. Grosso, C. De Laat and P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," Proc. of the International Conference on Collaboration of Technol. Systems (CTSs), pp. 48-55, 2013.

[14] M. Ali-ud-din Khan, M. F. Uddin, N. Gupta and N. Gupta, "Seven V's of Big Data Understanding: Big Data to Extract Value," Proc. Zone Conference Amer. Soc. Eng. Education, pp. 1-5, 2014.

[15] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data," Information Science, pp. 314-347, 2014.

[16] X. Jin, B.W. Wah, X. Cheng and Y. Wang, "Significance and Challenges of Big Data Research," Big Data Research, vol. 2, pp. 59-64, 2015.

[17] I. W. Tsang, J. T. Kwok and P.-M. Cheung, "Core Vector Machines: Fast SVM Training on Very Large Data Sets," Journal Machine Learning Research, vol. 6, pp. 363-392, 2005.

[18] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," Intell. Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.

[19] M. Ghanavati, R. K.Wong, F. Chen, Y.Wang and C.-S. Perng, "An Effective Integrated Method for Learning Big Imbalanced Data," Proc. of IEEE International Congr. on Big Data, pp. 691-698, 2014.

[20] C. Zhu, L. Cao, Q. Liu, J. Yin and V. Kumar, "Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings," IEEE Transaction Knowl. Data Eng., vol. 30, no. 7, pp. 1254-1267, Jul. 2018.

[21] H. Liu and H. Motoda, "Instance Selection and Construction for Data Mining," Springer, New York, vol. 608, 2013.

[22] H. A. Mahmoud and A. Aboulnaga, "Schema Clustering and Retrieval for Multi-domain Pay-as-you-go Data Integration Systems," Proc. of ACM SIGMOD International Conference on Management of Data, pp. 411-422, 2010.

[23] A. Kadadi, R. Agrawal, C. Nyamful and R. Atiq, "Challenges of Data Integration and Interoperability in Big Data," Proc. of IEEE International Conference on Big Data, pp. 38-40, USA, 2014.

[24] N. Ayat, H. Afsarmanesh, R. Akbarinia and P. Valduriez, "Uncertain Data Integration Using Functional Dependencies," Amsterdam: Informatics Institute, University of Amsterdam, 2012.

[25] D. A. Berry and B.W. Lindgren, Statistics: Theory and Methods, 2nd Edition, International Thomson Publishing Company, 1996.

[26] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R.Wald and E. Muharemagic, "Deep Learning Applications and Challenges in Big Data Analytics," Journal of Big Data, vol. 2, no. 1, pp. 1-21, 2015.

[27] S. R. Sukumar, "Machine Learning in the Big Data Era: Are We There Yet?," Proc. of the 20th ACM SIGKDD Conference on Knowl. Discovery and Data Mining, Workshop Data Science, pp. 1-5, 2014.

[28] J. Qiu, Q.Wu, G. Ding, Y. Xu and S. Feng, "A Survey of Machine Learning for Big Data Processing," EURASIP Journal Adv. Signal Process., vol. 67, pp. 1-16, 2016.

[29] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt and B. Scholkopf. "Support Vector Machines," IEEE Intelligent Systems and Their Applications, vol. 13, no. 4, pp. 18–28, 1998.

[30] S. K. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey," Data Mining and Knowledge Discovery, Kluwer Academic Publishers, vol. 2, no. 4, pp. 345–389, 1998.

[31] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun and R. Fergus. "Regularization of Neural Networks Using Drop-connect," Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 1058–1066, 2013.

[32] X. -W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, vol. 2, pp. 514-525, 2014.

[33] A. K. Jain, "Data Clustering: 50 Years Beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[34] T.-H. T. Nguyen and V.-N. Huynh, "A k-Means-like Algorithm for Clustering Categorical Data Using an Information Theoretic-based Dissimilarity Measure," Proceedings of the 9th International Symposium on Foundations of Information and Knowledge Systems (FoIKS), vol. 9616, pp. 115-130, 2016.

[35] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M. A. S. Netto and R. Buyya, "Big Data Computing and Clouds: Trends and Future Directions," Journal of Parallel Distributed Computing, vol. 79, pp. 3-15, 2015.

[36] S. B. Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249–268, 2007.

[37] O. Okun and G. Valentini, "Supervised and Unsupervised Ensemble Methods and Their Applications," Studies in Computational Intelligence Series, vol. 126, 2008.

[38] H. Zou and T. Hastie. "Regularization and Variable Selection *via* the Elastic Net," Journal of the Royal Society Series, vol. 67, no. 2, pp. 301–320, 2005.

[39] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[40] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Transaction Pattern Analysis Mach. Intell., vol. 35, no. 8, pp. 1798-1828, 2013.

[41] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann Machines," Proc. of the International Conference Artif. Intell. Statist., pp. 448-455, 2009.

[42] G. Hinton, "Deep Belief Nets," Encyclopaedia of Machine Learning, pp. 267-269, 2010.

[43] J. Read, F. Perez-Cruz and A. Bifet, "Deep Learning in Partially-labeled Data Streams," Proc. of the 30th Annu. ACM Symp. Appl. Computer, pp. 954-959, 2015.

[44] IMARTICUS, "What Is Machine Learning and Does It Matter?," [Online], Available: "https://imarticus.org/what-is-machine-learning-and-does-it-matter/".

[45] S. M. Basha and D. S. Rajput, "A Roadmap towards Implementing Parallel Aspect Level Sentiment Analysis" Multimedia Tools and Applications, Springer, vol 78, no. 1, pp 1-30, Jan. 2019.

[46] D. S. Rajput, R. S. Thakur and G. S. Thakur, "A Computational Model for Knowledge Extraction in Uncertain Textual Data Using Karnaugh Map Technique," International Journal of Computing Science and Mathematics, InderScience, vol. 7, no. 2, pp. 166-176, 2016.

[47] S. M. Basha and D. S.Rajput, "A Supervised Aspect Level Sentiment Model to Predict Overall Sentiment on Twitter Documents," International Journal of Metadata, Semantics and Ontologies, InderScience, vol. 13, no. 1, pp. 33-41, 2018.

[48] S. M. Basha, D. S. Rajput and V. Vandhan, "Impact of Gradient Ascent and Boosting Algorithm in Classification," International Journal of Intelligent Engineering and Systems, vol. 11, no. 1, pp. 41-49, 2018.

[49] D. S. Rajput, "Review on Recent Developments in Frequent Item Set Based Document Clustering, Its Research Trends and Applications," International Journal of Data Analysis Techniques and Strategies, InderScience, vol. 11, no. 2, pp. 176-195, 2019.

[50] S. M. Basha and D. S. Rajput "Parsing Based Sarcasm Detection from Literal Language in Tweets," Recent Patents on Computer Science, vol. 11, no. 1, pp. 62-69, 2018.

[51] F. Ö. Catak and M. E. Balaban, "A Map Reduce Based Distributed SVM Algorithm for Binary Classification," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 24, pp. 863-873, 2016.

[52] L. Demidova, E. Nikulchev and Yu. Sokolova, "The SVM Classifier Based on the Modified Particle Swarm Optimization," International Journal of Advanced Computer Science and Applications, vol. 7, no. 2, pp. 16-24, 2016.

[53] J. Tian, H. Rong and T. Zhao, "Hybrid Safety Analysis Method Based on SVM and RST: An Application to Carrier Landing of Aircraft," School of Reliability and Systems Engineering, vol. 80, pp. 56-65, Dec. 2015.

[54] L. Wang, G. Wang and C. A. Alexander, "Natural Language Processing Systems and Big Data Analytics," International Journal of Computational Systems Engineering, vol. 2, no. 2, pp. 76–84, 2015.

[55] A. Cuzzocrea, I.-Y. Song and K. C. Davis, "Analytics over Large-scale Multidimensional Data: The Big Data Revolution", Proceedings of the ACM 14th International workshop on Data Warehousing and OLAP, pp. 101-104, 2011.

[56] V. Agneeswaran, "Big-data - Theoretical, Engineering and Analytics Perspective," Big Data Analytics, Springer, vol. 7678, pp. 8-15, 2012.

[57] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey," Mobile Networks and Applications, Springer, vol. 19, no. 2, pp. 171-209, 2014.

[58] H. Li and X. Lu, "Challenges and Trends of Big Data Analytics", Proc. of the 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 566-567, 2014.

[59] N. Khan, I. Yaqoob, I. A. T. Hashem et al., "Big Data: Survey, Technologies, Opportunities and Challenges," The Scientific World Journal, vol. 2014, Article ID 712826, pp. 1-18, 2014.

[60] D. Jothimani, A. K. Bhadani and R. Shankar, "Towards Understanding the Cynicism of Social Networking Sites: An Operations Management Perspective," Procedia - Social and Behavioural Sciences, vol. 189, pp. 117–132, 2015.

[61] M. Blount, M. Ebling, J. Eklund, A. James, C. McGregor, N. Percival, K. Smith and D. Sow, "Real-time Analysis for Intensive Care: Development and Deployment of the Artemis Analytic System," IEEE Engineering in Medicine and Biology Magazine, vol. 29, no. 2, pp. 110-118, 2010.

[62] Apache Hadoop, "Hadoop Releases, Apache Software Foundation," [Online], Available: https://hadoop.apache.org/.

[63] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets," USENIX Workshop on Hot Topics in Cloud Computing (HotCloud).

[64] Apache Storm, "Apache Storm, Apache Software Foundation," [Online], Available: http://storm.apache.org/.

[65] Apache Cassandra, "Apache Cassandra, The Apache Software Foundation," [Online], Available: http://cassandra.apache.org.

[66] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, Taylor and Francis Group, A Chapman & Hall Book, 2013.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 06, No. 01, March 2020.

[67] Apache Hive, "Apache Hive, Apache Software Foundation," [Online], Available: http://hadoop.apache.org/hive.

ملخص البحث:

لقد أحدثت ثورة البيانات الضخمة تغييرات في نمط الحياة من حيث بيئات العمل والتفكير، من خلال تسهيل التحسينات في إيجاد الرؤى واتخاذ القرارات. غير أن المعضلة الفنية لعلم البيانات الضخمة تتمثل في عدم وجود المعرفة الكافية لإدارة وتحليل كميات هائلة من البيانات المتزايدة لاستخلاص معلومات قيّمة. وفي ظل النمو السريع للبيانات حول العالم واستمرار توزيعها باستخدام المعالجة في الزمن الحقيقي، فإن الأدوات التقليدية المتعلقة بتعلم الآلة لم تعد كافية. ومع ذلك، فإن الطرق التقليدية لتعلم الآلة قد جرى توسيعها لتلبية احتياجات تطبيقات أخرى، ولكن مع ازدياد المعلومات او قواعد المعرفة الخاصة بالبيانات الضخمة، فثمة تحديات تواجه خوارزميات تعلم الآلة بالنسبة لتحليل البيانات الضخمة.

تحاول هذه الدراسة تسهيل فهم أهمية تعلم الآلة في تحليل البيانات الضخمة. كما تسهم في فهم ما تتضمنه حسابات البيانات الضخمة من تعقيدات وبيان أبرز التحديات المتعلقة بتحليل تلك البيانات، الى جانب ما يرتبط بذلك من نقصٍ في دقة التصنيف وفي تجانس البيانات. وهي تناقش إمكانية استخلاص لامعنى من البيانات هائلة الحجم من أجل اتخاذ القرارات ومن أجل التحليل التنبؤي عبر تحويل البيانات واستخلاص المعرفة. وتبحث هذه الدراسة في أثر البيانات الضخمة في تحليل البيانات في الزمن الحقيقي ومدى إمكانية استخدام تعلم الآلة في تحليل البيانات الضخمة. ويؤمل أيضاً أن تكون هذه الدراسة منصّة انطلاق لدراسات وبحوث مستقبلية في مجال استخدام تعلم الآلة في تحليل البيانات الضخمة.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).