**Employing Machine Learning Models to Process and Assess Lumbering Big Data**

**Nazeer Shaik1, Dr.P. Chitralingappa1, Dr.B. Harichandana1**

**1Department of Computer Science & Engineering, Srinivasa Ramanujan Institute of Technology, Anantapur.**

**Abstract**

We live in a data-driven world, with enormous amounts of data produced as a result of the rapid development of technology that permeates every aspect of our lives. It is necessary to develop and improve data processing techniques over time in order to extract useful information from such a huge amount of data produced in diverse forms. Machine learning technologies offer possibilities for handling massive quantities of data and making the most of them. This study examines the literature on the application of machine learning techniques to large data sets. The book provides an overview of ML algorithms and techniques, an introduction to big data, and a discussion of related studies utilizing machine learning techniques to process huge amounts of data across many industries. Machine learning for big data has a number of challenges and problems, which are also discussed in the paper.

**Indexed Terms:** ML, Big Data, Process, Models or Patterns

1. **Introduction:**

Data has been growing quickly over the last few years, reaching an unprecedented amount as "Web technologies, social media, and mobile devices" have expanded, resulting in big data. For instance, Twitter processed more than 70 million tweets daily, producing more than 8 TB [1].

Only a few examples of Big Data sources include websites of Social Networking, hotel data, weather data, online retailers, bank sectors, and other resources [2]. However, unless the information is carefully and in-depth studied, it is meaningless. Big Data Analytics is a technique for examining huge data sets in order to gain insightful information that may be used for a variety of business applications or to enhance people's lives generally [2].

In the modern period, a huge amount of data is being produced from numerous formerly unheard-of and unobserved sources. Despite the fact that technology has been developed to acquire, analyze, and assess these unexpected data, there are still many difficulties and worries. Numerous research is being carried out to better understand and extract useful insights from big data. All fields of research—basic sciences, applied sciences, engineering, social sciences, biomedical sciences, and others—now deal with big data. All of these industries work with enormous datasets, and significant effort is being made to better harness and assess Big Data by utilizing fields like ML, which has great potential for addressing contemporary data challenges [3].

Since they offer prospective solutions for mining the data's hidden information, machine learning techniques have grown in popularity over the past ten years in a range of huge and complex data-intensive industries, including medicine, astronomy, biology, and others [4].

The organization of this essay is as follows: A description of machine learning and its methods is given in Section 2. The ideas and applications of big data are introduced in Section 3. The pertinent work is summarised in Section 4. The topic of machine learning and large data processing is covered in Section 5 of the article. Finally, Section 6 offers some findings.

1. **ML**

In this section, a summary of ML and its different methods, programs, and applications is given.

It is a formalized field of study that focuses on the theory, effectiveness, and attributes of learning systems and algorithms [4]. There are many fields that can be applied to it, including artificial intelligence, cognitive science, optimal control, theories of information and optimization, statistics, and other sciences [5] The primary goal of machine learning research is to develop quick, effective learning algorithms that produce data forecasts [6]. Due to the technology era that made it possible for everybody to produce raw data via their gadgets, data is currently rising exponentially. The raw data is probably going to be jumbled, fragmented, unstructured, and conflicting [7]. Through cleaning, altering, extracting, and combining the data, pre-processing transform it into a format that is suitable for learning [7]. A learning problem is one in which a task is carried out using training datasets to enhance performance indicators [5].

ML is classified into three types: supervised learning, unsupervised learning, and reinforcement learning. For supervised learning, labeled training data with inputs and anticipated outputs are necessary [4]. Unsupervised learning, in contrast, simply requires the inputs to be provided; the anticipated outputs are not necessary. In data that hasn't been labeled, such as cluster analysis, it is utilized to uncover hidden information [6]. The unsupervised learning method divides the sample sets into discrete clusters by evaluating how similar the input samples are to one another [8]. RL enables you to learn from the input you receive from interactions with the outside world [4]. It functions by trial and error, with the learner selecting the best approach in light of the outcomes [9]. When given a lot of training data, RL algorithms are especially good at picking up control rules without any prior knowledge. However, they do have some drawbacks, one of which is the high cost of computing necessary to discover the best solution [8].

The Categorization of ML technologies was completed in [4], which summarised the learning kinds, data processing tasks, distinguishing standards, and algorithms used for each, as shown in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Types of Learning | Tasks for Data Processing | Distinction Norm | Learning Algorithms |
| Supervised Learning | Classification/Estimation/Regression | Computational Classifiers | SVM |
| Statistical Classifiers | Naïve Bayes  Hidden Markov Chain Model  Bayesian Networks |
| Un Supervised Learning | Clustering & Prediction | Connection Classifiers | Neural Networks |
| Parametric | K-Means  Gaussian Mixer Model |
| Non-Parametric | Dirichlet Process Mixture Model  Q-Learning |
| Reinforcement Learning | Decision Making | Model – Free  Model-Based | R-Learning  TD-Learning  Sarsa Learning |

**Table.1: ML Technologies on Comparison**

* 1. **ML Methods**

The learning strategies that are described in this section may be helpful for dealing with large data-related difficulties. These learning strategies are unique in that they emphasize learning as a concept rather than a specific algorithm [4].

**2.1.1. Learning through Representation**

The main aim of representational learning is to learn useful and significant data representations [5]. It is possible to capture a huge variety of potential input configurations using a reasonable-sized learned representation, which can significantly improve statistical and computer efficiency [10]. The quality of the data representation is crucial for machine learning approaches to be successful [5].

**2.1.2. Deep Learning**

Artificial Neural Networks, or "Neural Networks," are the basis of the ML technique known as "deep learning," which is influenced by the structure and function of the brain [7]. DL typically uses supervised and/or unsupervised algorithms in deep structures to mechanically learn hierarchical representations, as opposed to Learning techniques that are most commonly used, which depend on shallow-structured learning architectures [11]. By increasing the model's depth or representational capacity with additional training examples, deep learning has the ability to enhance model performance [7]. To deal with the different kinds and amounts of big data analytics, it is more acceptable to apply deep learning architectures and algorithms [12]. Deep learning has the potential to eliminate the need for human feature selection by using techniques like hierarchical feature extraction and feature learning [8].

**2.1.3. Parallel and Distributed Learning**

A bottleneck in learning algorithms is how to scale them up so they can handle enormous amounts of data in a reasonable amount of time. In this situation, distributed learning holds promise because it is a strategy for doing just that [13]. Distributed learning enables the learning to be carried out in a distributed manner as opposed to the conventional learning paradigm, which requires data to be amassed in a database for the reason of central processing [14]. Some of the most well-known distributed and parallel ML techniques are stacking generalization, decision rules, distributed boosting, and meta-learning [5]. The fundamental tenet is that challenging learning scenario should be given priority by distributed and parallel learning algorithms [5].

**2.1.4. Learning Transfer**

Data collection for training purposes might occasionally be expensive or difficult. Consequently, high-performance learners must be taught using data from other areas through the application of transfer learning [7]. Transfer learning was proposed as a method to divide functions, domains, and distributions, enabling the knowledge to be extracted from several source tasks and applied to a target task [15,16]. Data size is not a factor when using the transfer learning algorithms that have been reviewed [7]. Transfer learning's advantage is its capacity to deftly apply previously acquired knowledge to quickly resolve brand-new problems [4].

**2.1.5 Active Study**

It's time-consuming and difficult to learn from massive amounts of unlabeled data. This issue is addressed by active learning, which selects a subset of the most important cases for labelling [17]. The active learner's objective is to attain exactness by using the fewest number of labeled cases possible, hence reducing the cost of locating labeled data [18]. With query methods that are more effective than those employed in conventional passive learning, a favorable classification performance could be obtained with fewer labelled samples [19].

**2.1.6 Learning Using Kernels**

Kernel-based learning has become a mainly potent tool for boosting processing capacity in the last ten years thanks to improvements in the design of effective non-linear learning algorithms [20]. The similarity of objects or images is assessed using a single kernel function in kernel-based machine learning rather than by looking at a variety of features [5]. To develop a learning technique and produce the desired output, a classifier, the kernel function is paired with images and labels [5].

1. **Big Data**

Big data is a term used to define datasets that are challenging to interpret, manage, or analyze using conventional IT, software, and hardware techniques in a timely manner. Alternatively, big data is defined as data that can be effectively processed employing significant horizontal zoom techniques or data that has a volume, acquiring speed, or format that makes it impossible to analyze using relational standard methodologies [21].

**3.1. Dimensions of Big Data**

By comprehending the many Vs associated with it, the big data notion may be understood more clearly. Big data management solutions must deal with these Vs as their primary dimensions (challenges). These attributes are described as follows:

**3.1.1. Volume (Size of the Data)**

As data is created at a rate of terabytes to zettabytes every second, it requires rethinking storage and processing models in order to develop tools that can analyze it effectively. [6,22–24] Distributed systems are used to store and analyze data across databases located around the world.

**3.1.2. Velocity (Velocity at Which Data Are Produced)**

The rate at which data is created and processed in order to satisfy demands is referred to by this word. Because real-time data is so vast and constantly in motion, traditional analytics are challenged by their rising reliance on it [6,22–25].

**3.1.3. Variety Data is presented in a variety of data formats**

Data format incompatibility is a fundamental problem because data can come from many different sources and assume several dissimilar forms. In the modern era, data can be found in a variety of formats, including organized, semi-structured, unstructured, and even intricately structured data. Traditional analytical techniques cannot handle huge data because of the variety of data forms. It takes a lot of time and work to prepare data for analysis, thus it is necessary to develop approaches that are efficient [6,22–25].

3.1.4. Reliableness (Data Quality)

The collected data's quality varies widely. It displays the biases, noise, anomalies, and other characteristics of the data. This will affect how accurate the analysis is. Maintaining accuracy prevents the accumulation of inaccurate data in the system. Truthfulness may affect the value [6,22–25].

3.1.5. Variability

SAS added the concept of variability as a new dimension to big data. Data flow rates can vary, as the word "variability" denotes. Big data's velocity frequently fluctuates, with peaks and valleys at random intervals [25].

**3.1.6. Validity**

Many times, the words "validity of data" and "veracity of data" are used interchangeably. Despite not having the same meaning, they are similar. Validity describes how accurate and true the data is in relation to the purpose for which it is being used. To put it another way, even if data is accurate, it might not be valid if it is not understood [25].

**3.1.7. Volatilities**

The structured data retention strategy that businesses utilize on a daily basis can be quickly recalled while discussing the volatility of big data. Once the retention term has passed, it can be readily destroyed [25].

**3.1.8. Value**

Big data's defining quality, according to Oracle, is value. The word "value" refers to the worthwhile knowledge derived from the data. The data's profundity of significance is known. But it's important to focus on the relevance [25].

Eight **Vs** are generally available in big data. Keep take mind that these Vs are not set; they might alter soon.

**3.2. Tools for Big Data ML**

The majority of currently available tools are focused on batch processing, stream processing, and interactive analysis. In this part, a few tools that are currently in use for large data analysis are examined.

**3.2.1 Hadoop and Apache MapReduce**

MapReduce and Hadoop are not synonymous; Hadoop is essentially a MapReduce concept implementation [26]. To process enormous data, the MapReduce model employs the divide-and-conquer strategy. Hadoop consists of a master node and a worker node, whereas MapReduce carries out two main processes, Map and Reduce. The Master Node divides the arriving data into subproblems, which are then given to worker nodes in the Map stage. The master node then performs the Reduce step, which involves combining all of the subproblems' results [26].

**3.2.2 Spark**

It is a processing engine for in-memory data that is made for quick and sophisticated analyses. It is applied to situations when bottom-up performance improvement is desired. Spark is 100 times more performant than Hadoop, especially for handling massive amounts of data, thanks to in-memory computing and other improvements. Apache Spark is quick even when data is stored on a disc. It currently holds the world record for the largest-scale on-disk sorting. In order to run existing learning tasks in a big data context, Spark offers a general middleware layer. A middleware layer of this type typically includes normal operations and primitives that are helpful for a range of learning activities [7,26].

**3.2.3 Storm**

Real-time networked computing is made possible by this piece of software. It is easy to use and install. It supports the use of any programming language. It can handle faults and is scalable [26].

**3.2.4 The Apache Mink**

For distributed and high-performance computation, Apache Mink is a stream processing engine. It operates accurately even with incoming data that is late. A thousand nodes can easily be added while still achieving exceptional latency and throughput [26].

**3.2.5 H2O**

The fastest in-memory data processing engine, H2O, is used to analyze and predict huge data. It is scalable, open-source software that can run on multiple nodes and is distributed [26].

These technologies are evaluated by taking into account the available language, execution model, related ML tools, fault tolerance, and latency.

**3.3. Big Data Applications**

Big data became a factor in a number of industries. It has been used, among other things, in the media, entertainment, communication, medical sector, government sectors, education, insurance, wholesale trade, marketing, and transportation and utilities.

**3.3.1. Applications for big data in Medicine**

Big data is employed in healthcare to store, handle, query, and analyze medical data effectively. Applications for medical big data will have a substantial impact on the healthcare sector. Analyzing clinical trial data, examining illness patterns, evaluating the effectiveness of patient care, researching and developing new medications, and other topics are examples.

For instance, Mount Sinai Medical Centre in New York employs Ayasdi's big data capabilities to analyze every genetic sequence of Escherichia coli, including roughly a million DNA variants, in an effort to determine why some bacteria types are antibiotic-resistant. Ayasdi uses the analysis of topological data, a novel mathematical research strategy, to examine the aspects of the data [21].

Big data in healthcare can be derived from a variety of sources, including genomic data, electronic medical records, monitoring devices, and wearable sensory devices [23,27].

**3.3.2. Applications of Big Data in Online Social Networks**

In social networking services, big data applications include socialized marketing, network intelligence gathering and analysis, network public opinion analysis, and support for government decision-making.

The applications of big data in social networking services include the gathering and analysis of network intelligence, social marketing, and network analysis of user sentiment.

Big data from social networking services encompass more than microblogging, chat rooms, and instant messaging.

Based on an analysis of social media data, the Santa Cruz Police Department identified criminal modes and patterns, as well as calculated crime rates in major cities [21].

**3.3.3. Education-Related Big Data Applications**

Big data is used by the US Department of Education to evaluate student performance. To determine how much time students spend on each topic, their "click patterns" are monitored. A trainer's effectiveness can be assessed, among other things, by looking at the number of students, the subject taught, and the venues [23,27].

**3.3.4. Applications of Big Data in Organisations**

Big data may assist companies in numerous ways to increase their industrial productivity and competitiveness:

* **E-Commerce**

Businesses analyze behavioral data as well as customer data to create detailed consumer profiles. These profiles could be helpful for developing content for various target markets, recommending content when asked, and monitoring the caliber of content.

In order to make precise music suggestions, "Spotify" gathers information about user behavior and uses big data Hadoop techniques to analyze it [23,27].

The Taobao Data Cube is a big data tool on the Taobao platform that enables merchants to stay updated on the macro industrial status of the Taobao platform, the market circumstances for their brands, and customer behavior, among other things, and make the proper decisions on production and inventory [21].

* **Securities and Finance**

Big data analytics are utilized in the financial industry for a variety of purposes, including pre-trade decision-making, scoring and analysis, credit risk, predictive analytics, and sentiment measurement. Big data analytics is additionally employed to examine enterprise risk management needs, fraud mitigation, and anti-money laundering [23,27].

For instance, China Merchants Bank (CMB) finds that practices like "Multiple times score accumulation" and "score exchange in shops" are effective at drawing in high-quality customers through data analysis [21].

* **Logistics**

Companies that handle logistics may have a lot of expertise with big data from the Internet of Things (IoT). GPS, wireless adapters, and sensors are included in UPS trucks so that the company's headquarters can track truck positions and prevent engine failures. For the time being, this technology helps UPS with personnel guidance, management, and route optimization. Based on their prior driving expertise, UPS trucks are assigned the best delivery routes. In 2011, UPS drivers travelled over 48.28 million fewer kilometers [21].

1. **Related Work**

This section will highlight related works on the handling of huge data sets using machine learning methods.

Recent research on large-scale machine learning solutions by researchers ([28]) produced a set of concepts and approaches that were discussed in this paper. In order to learn how to build massive Machine Learning systems and architectures efficient, broadly applicable, scalable, and expandable, these concepts and tactics encompass the entire spectrum of those systems and architectures [28].

In order to accelerate the discovery of new materials, machine learning can be applied to large datasets of nanoscience data by analyzing and extracting new insights, for example by integrating active learning into experimental design, as well as by integrating memristive devices into hardware that can be used for machine learning. The discussion focused on the potential for future collaborations between nanoscience and machine learning as well as the challenges that remain.

On four ML algorithms (Decision Tree Induction, SVM, Naive Bayes Classifier, and Random Forest Classifier), the effects of two pioneering dimensionality reduction techniques, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), were investigated in [30]. Experiments showed that machine learning approaches based on PCA perform better than LDA when working with large-dimensional datasets. It was established that ML methods without dimensionality reduction get superior results when the dimensionality of a dataset is low. Random Forest and Decision Tree classifiers outperform the other two approaches, as well as PCA and LDA, without the use of dimensionality reduction [30].

In the fight against COVID-19, machine learning and big data can contribute significantly by assessing epidemiological models, developing interactive dashboards, and suggesting the best methods for obtaining virus therapies [31]. Ababneh et al. in [31] showed that, despite a lot of difficulties and limitations, employing reinforcement learning to assess huge data produces excellent and fantastic outcomes.

Leung et al. in [32] unveil a big data and ML analytical tool for processing and estimating COVID-19 epidemiological data. In order to generalize some specialized features into some general qualities for big data analytics, the tool takes excellent use of OLAP and taxonomy. In order to train a supervised learning model that predicts medical outputs (such as whether a patient will be dead or recover from COVID-19) for new datasets, this tool customs common patterns seen in prior data. The evaluation's findings indicate that the instrument is helpful in providing a wealth of knowledge on the characteristics of COVID-19 cases. This makes the disease more understandable for epidemiologists, researchers, and decision-makers, which may inspire them to look at fresh approaches for preventing, managing, and treating it [32].

In [33], the researchers used big data and machine learning to analyze COVID-19 predictions, except for CT scans and X-ray images, taking into account all significant fundamental characteristics and other literature from around the world that was correlated with each other. It was shown how accurately computers predicted outcomes by showing how some algorithms produced inverted numbers while others produced accurate forecasts with fewer errors. In addition to a state-by-state population index, the study used two classification techniques for COVID-19 cases in India from January 30 to May 30, 2020. Furthermore, the outcomes of the two approaches were largely comparable. They arrived at the conclusion that medical decision-making can be enhanced by anticipating COVID-19's future diegesis, especially in situations requiring quick intervention.

In [34], the role of big data analysis in healthcare was investigated, and various shortcomings in conventional machine learning techniques were found. Machine learning and big data will work together over the next ten years to improve every element of the healthcare industry.

Machine learning algorithms are increasing the early identification of breast cancer using extensive clinical and gene expression data [26]. In [26], Gupta et al. did research on the use of data analytic frameworks, tools, methodologies, and ML approaches in the context of breast cancer, particularly in the areas of cancer recurrence, cancer survivorship, cancer diagnosis, and cancer prediction. They discovered that the two most popular ML techniques for spotting breast cancer are artificial neural networks (ANN) and support vector machines (SVM). Numerous machine learning (ML) frameworks have Apache Spark's compatibility has been verified.

Based on a dataset, big data can anticipate an impending risk of diabetes and recommend the appropriate course of treatment [35]. In [35], Saxena examined the relationship between diabetes awareness and the prediction model of machine learning systems. The prediction results produced by the Random Forest (RF) and Support Vector Machine (SVM) algorithms are superior to those of other ML algorithms.

This study examined the application of machine learning and big data in environmental and water management (EWM), with a particular focus on DL. This survey provides a comprehensive overview of important ideas and techniques in big data and machine learning as well as a systematic review of existing applications, demonstrating the potential and advantages of data-driven research in EWM. As well as discussing important research issues, it also discusses future directions.

The effectiveness of six ML classifiers was evaluated using a random feature-pooling strategy in [37]. According to his research, the effectiveness of the classifier is not much impacted by the feature selection approach. The effect of feature selection on several datasets, including those with many samples but few features and those with few samples but many features, was also examined. By contrasting classifier performance in terms of accuracy, sensitivity, and specificity, it is possible to assess how effective the suggested CC-based content is. Although a detailed issue analysis method for feature selection may produce the best results, the researcher thinks that a good decomposition strategy alone is insufficient to realize the full potential of CC-based techniques. For CC-based approaches, the development of each subpopulation requires a suitable optimizer, and the development of the entire solution file requires a suitable collaborative technology.

A member of the GBDT family of machine learning technologies, CatBoost. Since its introduction in late 2018, CatBoost has been effectively used in machine learning projects employing massive amounts of data. In light of this, researchers in [38] have examined CatBoost research as it relates to big data, gathering best practices from studies that demonstrate both CatBoost's superiority over other technologies and its inferiority to them. Due to the fact that CatBoost is a decision tree-based method, it is also ideally suited for ML applications that need categorical and heterogeneous data.

Based on their findings in [39], the authors discussed the application of big data and machine learning in agriculture, identified issues and modifications, developed architectures for these systems, and conducted an in-depth systematic literature review (SLR), which examined 34 examples of these systems. As a result of cloud technologies, processing huge amounts of data is no longer a challenge. Data processing speed remains a challenge due to the lack of control over raw, semi processed, and processed (value data) data, as well as information visualization systems that support technical data that farmers rarely understand.

[40] reveals the difficulties and possibilities presented by sensing technologies in terms of assisting animal keepers to yield excess meat and the products of animals. This study focuses on the importance of sensors, big data, AI, and ML in assisting animal breeders in reducing production costs, boosting efficiency, improving animal welfare, and producing more animals per hectare. It also looks at the difficulties and constraints posed by technology. In order to comprehend the usefulness of animal husbandry technology in aiding farmers to enhance animal health, increase revenues, and minimize environmental effects, the researchers underlined the numerous uses of the field.

The focus of reference [8] was on the application of ML and big data analytics to the field of smart buildings. It provided a thorough analysis of research articles on big data and machine learning applications, especially for creating smart infrastructure and services.

Many studies have investigated machine learning techniques that are used to classify images using computation-based experiments and comparison studies. It is [41] one of them. As a result, the deep convolutional network performs the best. It has a lower error rate than stacked de-noising auto-encoders, multilayer perceptrons, and logistic regression. The greatest time savings came from the convolutional neural network running on the GPU. This is also a result of NVidia's cudnn library, which is optimized for convolutional operations [41].

Using simple machine learning algorithms, the spectral structure and temporal delay of individual terahertz pulses in linear accelerator sources may be correctly predicted [42]. These algorithms can process enormous amounts of data to train deep and specialized Artificial Neural Networks (ANN) on heterogeneous processors like FPGAs and GPUs [42].

Machine learning is now widely used to satisfy a variety of data operations, such as data classification, prediction, and archiving [43], as a result of the pressing requirement for enormous astronomical data. Due to its capacity to uncover connections and interactions across various time points, recurrent neural network (RNN) has been extensively noted to be very effective for time series research [43].

Finding anomalies and network attacks is one use for big data processing techniques and machine learning algorithms. Using a combination of machine learning algorithms and large data processing approaches, Reference [44] described a method for identifying network attacks and abnormalities in cybersecurity datasets. In order to validate the suggested method for identifying network attacks and abnormalities, two different datasets were used. There are 7,009,270 examples in the first collection, which was developed in a mobile IoT network. The second set consists of the CICIDS2017 dataset, which has over 500,000 entries and represents DDoS and port scanning attacks. As part of this method, k-nearest neighbors, support vector machines, linear regression, Gaussian nave Bayes, decision trees, and two-layer perceptrons are used to build and share classifiers. An initial dataset can be assessed using principal component analysis, reducing the challenge of object classification.

1. **Methodology**

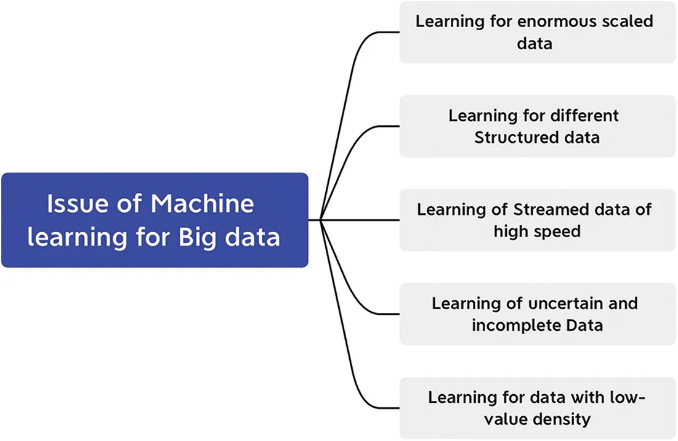
This section describes the research approach used. The authors looked for comparable publications in well-known databases including IEEE, SpringerLink, Google Scholar, and others. One of the search terms was "Machine Learning," along with "Big Data," "Algorithms," and "Processing." Data was gathered by the authors using two techniques: research extraction and research screening. During the study extraction procedure, the authors removed many papers based on pointless titles, leaving 51 studies. The authors then read the abstract and conclusion of each manuscript. As a result, certain documents that fell beyond the parameters of the investigation were disregarded. A thorough analysis of the aspects of machine learning utilized to process big data that were discussed in the literature was conducted.

1. **Discussion**

The most significant issues with big data processing methods utilizing machine learning are covered in this section.

**6.1 ML for Big Data Issues and Challenges**

In **Fig. 1,** we provide a complete scenario that includes an assessment of difficulties associated with ML methods for big data from many perspectives. It contains (i) learning for massively scaled data (ii) learning for different structured data (iii), learning for high-speed streamed data (iii), learning for uncertain and incomplete data (iv), and learning for low-value density data (v).



**Fig.1: Big Data Learning Methods**

**6.1.1 Preparing for Huge Big Data**

We deal with an increasing volume of data as a result of technological improvements. Google processes about 25 petabytes of data per day, and businesses will someday process petabytes of data, according to research published in November 2017 [4]. Big data's primary characteristic, which also presents a serious problem, is unquestionably its bulk [4,45]. In order to address this issue, distributed and parallel computing frameworks should be chosen [4].

**6.1.2. Learning for Various Structured Data, 6.1.2**

Data today comes in a wide range. Structured, unstructured, and semi-structured data are the three forms of data that could produce diverse, non-linear, and high-dimensional data [12]. It is quite difficult to learn from such a massive data set, which makes the data more complex. Data integration will therefore be necessary to get beyond this difficulty [4].

**6.1.3. Streaming High-Speed Data Learning**

There are many tasks that have to be finished in a certain amount of time. One of the most crucial characteristics of big data is its speed [45]. The processing outcomes may become worthless or lose their value if a work is not finished by a certain date [4]. Among these are the prediction of earthquakes and the stock market [4]. Processing enormous amounts of data fast and accurately is therefore an essential but challenging task. To get past the obstacles, a method for online learning should be employed [4].

**6.1.4. Learning from Uncertain and Incomplete Data**

Machine learning techniques were used to get data that was more accurate. since the outcomes were accurate at the time. Today's data is muddied, however, by the fact that it comes from a range of unreliable and partial sources. Because of this, machine learning faces serious challenges in huge data analytics [4]. We identify veracity as the fourth major problem for learning with big data [4] to emphasize the significance of addressing and managing the uncertainty and incompleteness of data quality. For instance, in wireless networks, uncertain data is information that is produced as a result of noise, fading, shadowing, and other variables [4,46]. To get around this issue, use a distribution-based approach [4].

**6.1.5. Data Learning with Low-Value Density**

In big data analytics, machine learning is primarily used to glean useful information from vast amounts of data for financial gain. One of the most crucial aspects of data is its value [4]. It is quite challenging to extract meaningful value from enormous, varied, and sparsely valued amounts of data. As a result, machine learning in big data analytics faces numerous difficulties [4]. To address this issue, database knowledge discovery and data mining technologies should be used [4]. The usage of these technologies is necessary since they offer potential remedies for obtaining crucial facts from vast amounts of data. Data mining studies were examined by the authors of [24].

Big data analytics should be cautious while addressing machine learning's numerous problems. Because there are so many machine learning products available, they all need a lot of training data to function well. To be accurate, machine learning models require to be trained on structured, pertinent, and accurate historical data. There might be other difficulties, but it's not impossible.

1. **Conclusion**

As a means of overcoming the challenges presented by big data and discovering hidden patterns, facts, and knowledge within vast amounts of data, machine learning is essential to turning this capability into a meaningful incentive for core business leadership and logical investigation. This study illustrated the function of machine learning algorithms in the handling of large amounts of data. Both big data and machine learning algorithms and methodologies were covered in a general overview. It also covered similar studies that applied machine learning methods to the processing of massive data across different domains. Finally, it covered the difficulties and problems that arise when utilizing machine learning to process large amounts of data.

**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, 20
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.