International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

A Review on Several Machine Learning Algorithms Used to Handle Big Data Classification Problem

Rajesh Pandey[1], Dr. Mamta Bansal[2], Yogesh Awasthi[3]

[1]Assistant Professor, School of Computational Sciences & Engineering, Shobhit Institute of Engineering & Technology, Meerut, (Deemed to- be- University), Meerut, Uttar Pradesh, India

[2]Professor, School of Computer Science & Engineering, Shobhit Institute of Engineering & Technology, Meerut, (Deemed to- be- University), Meerut, Uttar Pradesh, India

[3]Professor and Dean, College of Engineering and Applied Science, Africa University, Mutare, Zimbabwe

Abstract: ***In the highly technologically advanced and digitally transformative era where several new technologies are emerging and creating enormous amounts of data, big data is growing faster. A vast amount of digital data has been generated by several electronic devices, including smart phones, computers, sensors, smart kitchens, and domestic appliances, leading to a global shift in the acceptability of the internet. The data in each domain grew over the time period. Big data analytics may help different domain like banking, Finance, business and medical, but it takes exceptional skills to find a meaningful pattern in these data. From everyday transactions to consumer interactions and social network data, decision makers should be able to extract valuable information from such vast and rapidly evolving data. This huge volume of data is very complex and unclean which makes it difficult to analyze and extract meaningful information. The main problem for machine learning algorithms is this unbalanced and disorganized data. Our goal is to conduct a comprehensive analysis of various tools, methods, and machine learning classification algorithms for big data, evaluating them on the basis of cleanliness and big data complexity reduction. In addition, it examines recent developments and suggests a methodology to improve algorithmic efficiency and data preprocessing to address the complexities of real-world datasets.***

Keywords: Big data; Machine Learning; Imbalance data; Data mining

1. Introduction

In today's technologically advanced era, many new technologies are being developed and generating enormous amounts of data. Every aspect of contemporary civilization contributes to the enormous velocity of these data, which accelerates the big data. Due to the increasing global adoption of the internet, a huge volume of digital data has been produced by several electronic devices, including smart phones, computers, sensors, smart kitchens, and household gadgets. The data in each domain grew over the time period. Big data analytics can help industries like banks, credit cards, and medical, but it takes exceptional ability to find a valuable pattern in these data. The utilization of contemporary technologies releases vast amounts of data that are beneficial for numerous businesses. These massive amounts of historical transactional data that are produced as byproducts from various domains can be used for a variety of decision-making purposes, but only if they are correctly processed and used. It is necessary to modify the traditional, outdated methods of doing data analysis in order to incorporate big data. New techniques and technology are therefore urgently needed for data analytics in every field.

Figure 1 illustrates how data is fundamental to both data science and data mining. Data mining is a method for drawing out knowledge and important patterns from huge volume of data. Depending on the type of data, the pattern that was extracted from the enormous amount of data is useful for a variety of applications, including market research, fraud detection, disease diagnosis, customer retention, science investigation, etc. To extract pertinent information from the vast amount of data, data mining employs a machine learning algorithm. In order to create a model that can forecast future data, a machine learning (ML) technique uses a variety of algorithms to learn prediction rules from historical data

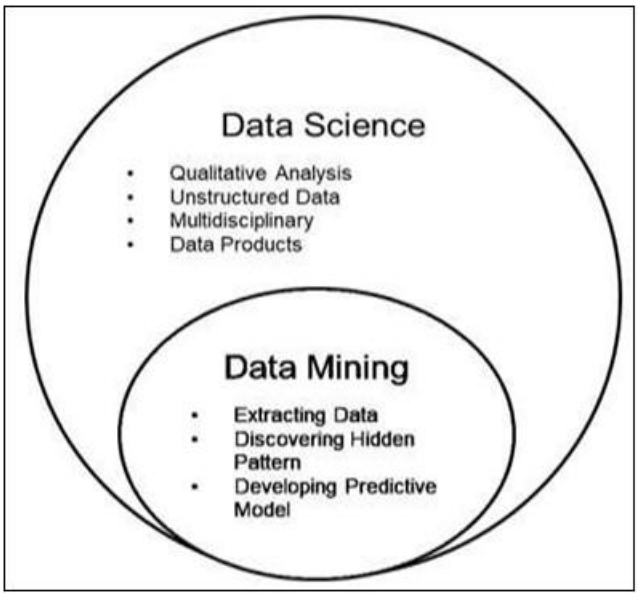

Figure 1: Role of data mining in data science

The process of altering a dataset by either over sampling or under sampling is known as sampling. In order to obtain a normal distribution for a dataset, under sampling decreases the majority class and oversampling increases the minority class. However, both sampling techniques have a number of drawbacks. For example, under sampling can lead to information loss, while over sampling can result in poor performance for high-dimensional datasets and an additional over fitting issue, which presents another difficulty for machine learning algorithms.

Machine Learning algorithms need very large set of clean data as a pre-requisite for training purpose but almost every dataset when gathered from different sources is unclean. Messy data and data imbalance is a real challenge for ML algorithms and should be taken care off during data preprocessing stages. Several solution is present for considering both the issues with certain limitations but there is a need of single solution capable of solving two major dataset complexity at a time for saving lots of effort and time - a technique capable of (a) cleaning the dataset (b) while balancing it. Therefore, in order to fulfill the need intension here is to design a technique capable of removing redundancy and outliers from a majority sample of a dataset, which cleans and at the same time reduces it and then increasing minority sample for balancing the dataset.

Messy data and data imbalance are the real challenges for ML algorithms and should be taken care off during data pre-processing stages. Big data complexity has received a lot of attention over the past decade.

2. Big Data: An Introduction

The data management in big data by 3 V's : volume, velocity and variety as shown in fig 2. He designed these three V's on the basis of his observation of significant change in the volume of data. In Laney's 3 V's, volume means massive quantity of data, velocity means increasing rate at which data is generated and variety means various format of data gathering from a diverse source [(Laney 2001)].

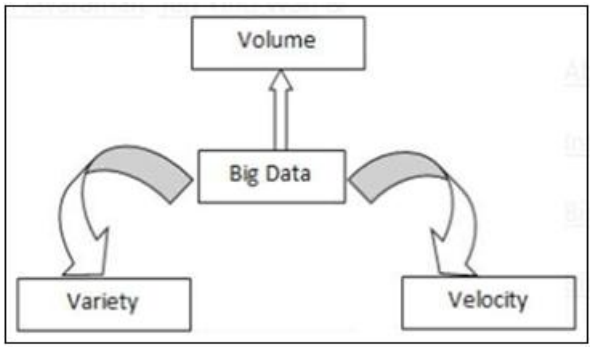

Figure 2: Schematic diagram of three Vs of the Big data

Later, more number of V's have been added to the big data such as veracity [(Uddin and Gupta 2014),(Demchenko, De Laat et al. 2014),(Jin, Wah et al. 2015)], value [(Uddin and Gupta 2014),(Demchenko, De Laat et al. 2014),(Jin, Wah et al. 2015)], validity [(Uddin and Gupta 2014)], variability[(Demchenko, De Laat et al. 2014)][(Philip Chen and Zhang 2014)], volatility [(Uddin and Gupta 2014)] and virtual [(Uddin and Gupta 2014)]. These different proposed V's have created lots of controversy and mystification among the V's of big data.

The term "big data" describes the collection of tools, methods, and technologies that can efficiently handle data of any size. The discipline of data processing has developed as a result of growing memory capacities and sophisticated storage technologies, increased processing capacity of modern computers, and the availability of vast amounts of data. Large volumes of data may now be managed, controlled, processed, and analyzed like never before because to modern hardware and software technology.

A flood of data is being produced as an outcome of the expansion of the Internet, social media technology, gadgets, and applications. Patterns, trends, and relationships pertaining to human behavior and interaction can be found by gathering, analyzing, and correlating data sets with incredibly vast dimensions. Utilizing big data helps to enhance operational efficiency, save costs and risks, create more focused marketing efforts, and gain a deeper understanding of consumer behavior. International Data Corporation (IDC), a global provider of market intelligence and information technology advisory services, estimates that the global big data and analytics market will surge in times to come [(Press 2014)]. As a result of this wealth of information, businesses face the challenge of determining how to best utilize it.

The representation and quality of data is very important for machine learning algorithms. Knowledge extraction becomes difficult during the training phase if the data is irrelevant, redundant or noisy. Therefore pre -processing of data is very important but on the contrary, it is very time consuming. Pre-processing of data include data cleaning, normalization, transformation, feature extraction and selection, etc. Machine learning algorithms can be applied on clean training data set which is the output of data pre-processing. Data pre-processing enhances the performance of machine learning algorithms but preprocessing is the very much time consuming approach [(Dridi 2021)].

Since data is always being generated and we are inundated with it, today's big data might not be big tomorrow. Instead, it needs to be harnessed and analysed to reveal fresh insights. As a result, old-fashioned data processing tools that cannot handle huge data will eventually be rendered useless.

Big data's enormous sample size and high dimensionality provide unique computational and statistical problems, such as scalability and storage constraints, noise accumulation, false correlation, unexpected heterogeneity, and measurement mistakes. These could lead to inaccurate statistical conclusions, which could subsequently produce inaccurate scientific results. [(Fan, Han et al. 2014)].

The big data era has entered a new phase. Better analysis of the increasing volumes of data could lead to faster advancements in a number of scientific sectors and boost the profitability and success of many enterprises. However, a number of technological challenges need to be addressed before this promise can be fully realized. Addressing these technological issues in the context of a single domain would not be cost-effective because they are common across a variety of application domains. Furthermore, these issues will not be automatically addressed by the next generation of industrial products; rather, they will require revolutionary solutions. [(Khalil, Kim et al. 2020)].

3. Machine Learning: A Brief Introduction.

The basics of machine learning, its history, the various applications of ML techniques, and advanced machine

learning techniques recently presented are all covered in this part. The field of big data Processing is in dire need of ML-based problem solving.

3.1 Machine Learning Techniques: Classifications & Use

In the field of computing, the concept of machine learning is not new, but because of the constantly shifting demands of the modern world, it has taken on an entirely new 'Avatar'. Everyone is now discussing ML-based solution ideas for a certain issue set. In machine learning (ML), computer algorithms are used to automatically learn from data and information. More digital information is being produced as a result of the internet's growth, which indicates there is more data for machines to evaluate and "learn" from [(Bhatnagar 2018)]. As a result, machine learning is once again becoming popular. Thanks to machine learning algorithms because of which computers can now interact with people, drive themselves, write and post reports on sporting events, and even identify potential terrorists. The fastest-growing area in computer science is machine learning (ML) [(Pugliese, Regondi et al. 2021)].

Some of the widely used machine learning techniques/methods include classification [(Dridi 2021)], regression [(Salihoun 2020)], topic modelling [(Bharadiya 2023),(Dietrich, Heller et al. 2015)], time series analysis [(Dietrich, Heller et al. 2015)], cluster analysis [(Dietrich, Heller et al. 2015), (Naeem, Jamal et al. 2022)], association rules [(Salihoun 2020)], (Dietrich, Heller et al. 2015)], collaborative filtering [(AL WAILI 2023) ,(Ryzko 2020),(Wu, Wu et al. 2023)], and dimensionality reduction [(Vurgaft 2023), (Rodionova, Kucheryavskiy et al. 2021)]. Based on the current patterns and correlations among the data in the given dataset, they are used to do analytics and forecast future trends.

In Fig 3, compare three ML subdomains from various angles and describe the ML techniques for data processing. Supervised learning, unsupervised learning, and reinforcement learning are the three primary subfields of machine learning (ML). By using labeled data to train models, supervised learning enables them to forecast results in response to novel inputs. Conversely, unsupervised learning focuses on clustering or dimensionality reduction and works with finding patterns or groupings in data without specified labels. By using rewards or penalties to educate agents to make decisions, reinforcement learning focuses on helping them gradually learn the best course of action. By making sure the data is ready for training, data normalization, feature extraction, and data augmentation are three methods for data processing in machine learning that are crucial for increasing model accuracy and generalization (Qiu, Wu et al. 2016).

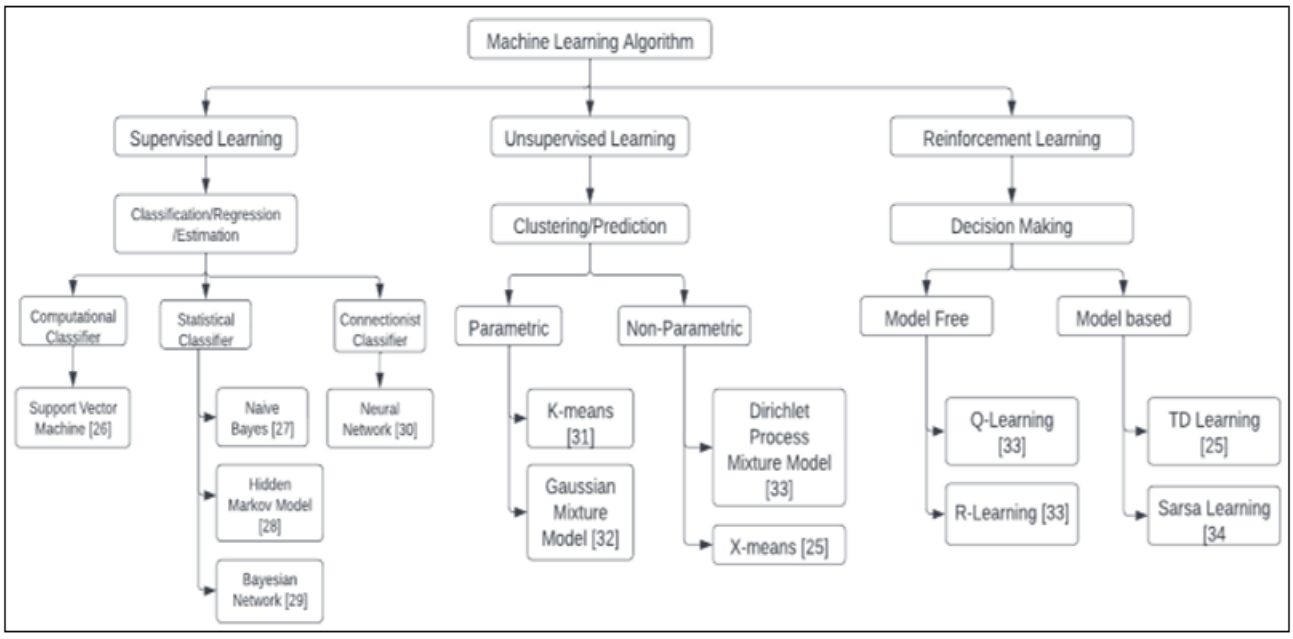

Figure 3: Comparison of Machine Learning Technology

To apply machine learning (ML) to big data, big data processing frameworks such as Apache Spark offer machine learning features and components. The various components of Apache Spark, a broad data processing structure, are used by researchers worldwide for a variety of reasons [(Sutton 1988)]. MLlib is the name of Spark's machine learning (ML) library. Its goal is to make actual machine learning simple and sustainable. Apache Flink is an open-source stream processing framework designed for networked, efficient, always-available, and precise information streaming systems.

4. Big Data Processing Using Machine Learning: Issues, Challenges and Opportunities

Managing this "messy" The quick rise in the "3 Vs" of big data i.e volume, velocity, and variety of data, present special difficulties that traditional data processing techniques cannot handle [(Laney 2001), (Uddin and Gupta 2014),(Bansal, Chana et al. 2020)]. Machine learning (ML) has become a crucial tool for managing these complexity as a result of the development of sophisticated technologies. Big data frequently includes massive datasets that are impossible to

process with conventional tools, necessitating the development of new techniques to effectively extract valuable information [(Fan, Han et al. 2014), (Salihoun 2020), (Shafiq, Tian et al. 2020)]. However, integrating ML with big data processing is not without its hurdles. Data quality, scalability, algorithmic efficiency, and the requirement to manage disparate data sources are the main concerns [(Bhatnagar 2018), (Ryzko 2020), (Najafabadi, Villanustre et al. 2015), (Rathore, Shah et al. 2021, Kumar, Sharma et al. 2022)]. Because noisy or redundant data can have a major impact on model performance, researchers have shown that data purification is essential before implementing ML algorithms. data to maintain dataset balance is still a crucial component of big data analytics [(Dridi 2021), (Han, Huang et al. 2017), (Devi and Sabrigiriraj 2018)] .

4.1 Data Quality and Preprocessing Challenges

The preprocessing phase is one of the main obstacles. Steps like resolving missing values, removing outliers, and normalizing datasets are all part of data preparation and cleaning. These problems might result in inaccurate analysis, erroneous insights, and subpar model predictions if they are not properly resolved. These problems can be successfully mitigated by hybrid approaches that combine machine learning-based models with conventional data pretreatment techniques [(Dridi 2021), (Dietrich, Heller et al. 2015)]. Researchers can greatly enhance the quality of the processed datasets and increase the overall performance of downstream machine le arning applications by creating adaptive models that can learn from the discrepancies in the data [(Alhenawi, Al-Sayyed et al. 2022), (Liu, Kitouni et al. 2022),(Raj, Shobana et al. 2020)]. Efficiency is further increased and less manual intervention is required when preprocessing chores can be automated utilizing cutting-edge methods like deep learning and unsupervised learning [(Sarker 2021)].

In order to address the particular difficulties posed by big data, recent developments have created hybrid approaches that combine increasingly complex ML-based solutions with conventional data pretreatment techniques [(Nti, Quarcoo et al. 2022), (Al-Marghilani 2022),(Rahmani, Yousefpoor et al. 2021)]. During data preprocessing, noise and outliers, for example, pose major challenges. To overcome them, a variety of techniques have been developed, ranging from basic statistical filtering to sophisticated machine learning models that can dynamically identify and reject irrelevant data [(Keswani, Vijay et al. 2020), (Chen, Zobel et al. 2023), (Cao, Yan et al. 2023)]. However, despite these developments, issues related to data scalability, heterogeneity, and real-time processing continue to pose challenges for practitioners [(Bhatnagar 2018), (Almutiri, Alhabeeb et al. 2022), (Luengo, García-Gil et al. 2020), Zhang & Yang, 2023]. For example, managing data from diverse sources, such as text, images, and videos, requires ML models that can adapt to varying formats without losing efficiency or accuracy. This adaptability is still underdeveloped in many existing systems, which struggle to scale effectively as data size increases [(Bharadiya 2023), (Dietrich, Heller et al. 2015)].

4.2 Scalability and Distributed Frameworks

Using distributed computing frameworks like Hadoop and Apache Spark, which enable the effective management of massive datasets, is one of the most exciting opportunities in big data processing [8, 16,(Ketu, Mishra et al. 2020)]. By parallelizing the calculation process, these frameworks can expedite machine learning model deployment and training. These frameworks are made to manage the "volume" component of big data by dividing up the work across several computers so that they can work on different sections of the data at the same time. But even though this technique can drastically cut down on processing time, the intricacy of these systems necessitates careful management to avoid problems like scalability limitations, data bottlenecks, and node failures [(Salihoun 2020), (Ryzko 2020), (Vurgaft 2023)]. It has also been investigated to integrate distributed frameworks with cloud services, allowing for dynamic resource allocation that can improve system resilience even more [(Ali 2023)].

The use of cloud-based systems, such Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure, to implement machine learning models for Big Data processing has gained more attention in recent years. Large data volumes can be handled more easily with these platforms' scalable architecture, which eliminates the need for infrastructure administration [(Naeem, Jamal et al. 2022), (Wu, Wu et al. 2023), (Noman 2024)]. Research has indicated that the integration of cloud-based solutions with distributed frameworks can result in notable cost reductions and enhanced processing speeds. Furthermore, these platforms include a range of tools for model training, data preprocessing, and deployment, which facilitates the implementation of end-to-end solutions by businesses with minimal costs [(Gupta and Sharma 2023)].

4.3 Feature Selection and Dimensionality Reduction

The feature selection procedure presents yet another important obstacle. Because big data frequently consists of both structured and unstructured data, it can be challenging to determine which properties are pertinent for study [(Pugliese, Regondi et al. 2021), (Schneckenreither 2020)]. Efficient feature selection increases processing efficiency in addition to model correctness. According to recent research, feature extraction and unsupervised learning approaches can be used to increase data representation and ML algorithm performance [(Tang, Lin et al. 2022), (Pérez Pupo, Piñero Pérez et al. 2020), (Devulapalli, Potti et al. 2023)]. The speed and accuracy of machine learning algorithms are directly impacted by the reduction of data dimensionality, which is largely achieved by feature selection. But choosing the most significant features is a difficult undertaking, particularly when working with high-dimensional data sets [(Chong, Khaw et al. 2023)].

In feature selection challenges, new methods such as AutoML and neural architecture search (NAS) have begun to gain popularity. Without requiring extensive manual involvement, autoML platforms provide optimal feature selection by automatically exploring various configurations of ML models and feature sets. In big data applications, where the feature space can be daunting, this can be especially helpful. In order

to improve generalization and performance across a variety of datasets, researchers are now investigating the possibilities of unsupervised feature learning techniques, in which models are trained to identify patterns and choose features without explicit supervision [(Zhou, Jin et al. 2021);(Nguyen, Nguyen et al. 2023)].

4.4 Real-Time Processing and Edge Computing

In many businesses, real-time data processing has become essential, especially when quick decisions are needed. For example, real-time data processing can result in notable enhancements in service quality and operational efficiency in the domains of finance, healthcare, and autonomous systems [(Alqahtani, Changalasetty et al. 2023), (Can, Yavuz et al. 2020)]. Recent advancements in edge computing have made it possible to process data closer to the source, reducing latency and bandwidth usage [(DHAMELIYA, PATEL et al. 2024)]. Machine learning and edge computing work together to install lightweight models on IoT sensors, smartphones, and embedded systems, delivering real-time analytics and insights right at the data source [(Alhenawi, Al-Sayyed et al. 2022)].

A hybrid approach is made possible by the integration of edge computing and cloud infrastructure, where more complex analyses are carried out in the cloud while crucial data processing takes place on the edge. It has been demonstrated that this hybrid method improves the overall effectiveness of ML models, particularly in situations involving streaming or highly velocityd data. Businesses are using edge-based machine learning to create real-time monitoring systems for predictive maintenance, industrial automation, and even real-time customer interaction in retail settings [28, 53]. However, as edge devices frequently have constrained memory and processing capability, creating models that function well on them necessitates careful optimization [(Devi and Sabrigiriraj 2018)].

4.5 Applications Across Various Sectors

Notwithstanding these difficulties, there are a ton of opportunities in a variety of industries when machine learning is included into big data processing. Businesses in a variety of sectors, including healthcare and finance, have used machine learning (ML) techniques to extract valuable insights from massive datasets, enhancing operational efficiency and decision-making [(Demchenko, De Laat et al. 2014), (Alghunaim and Al-Baity 2019)]. Healthcare systems, for example, use machine learning algorithms to evaluate patient data, forecast disease outbreaks, and enhance treatment regimens [(Wu, Wu et al. 2023), (Tahmassebi, Gandomi et al. 2017)]. By examining trends in genetic data, patient histories, and medical imaging, machine learning models can help with early diagnosis, which could result in better results and earlier therapies [(Nguyen, Nguyen et al. 2023)].

These models are also used in the finance sector for consumer behavior analysis, risk assessment, and fraud detection. Businesses may act on information in real time and gain a competitive edge by processing and analyzing enormous amounts of data quickly. Big data is used by financial organizations to identify transaction irregularities, forecast market trends, and improve client experiences by making tailored suggestions [(Gupta and Sharma 2023)]. Furthermore, machine learning models are used in the manufacturing industry for quality control, supply chain optimization, and predictive maintenance, which lowers downtime and saves money [(Zeng and Ge 2020), (Lu 2020), (Chen, Zobel et al. 2023)].

Big data and machine learning are being utilized in environmental research to evaluate the effects of human activity on ecosystems, forecast natural disasters, and study climate data [(Schneckenreither 2020)]. In order to monitor deforestation, ocean temperatures, and air quality and help guide decisions, researchers can now examine vast amounts of sensor data and satellite pictures. Additionally, investigating novel hybrid techniques that integrate the advantages of many machine learning methodologies may provide more thorough answers to current problems. Furthermore, coordinating the use of machine learning (ML) in big data with international projects like the Sustainable Development Goals (SDGs) of the UN could promote more socially conscious and sustainable inventions, improving the globe [(Lemarchand, McKeever et al. 2022)].

4.6 Research Opportunities

In summary, while machine learning and big data are both effective techniques in and of itself, their combination is essential to the advancement of contemporary data analytics. Unlocking big data's full potential will require addressing the issues of processing efficiency, scalability, and data quality [(Dridi 2021), (Nti, Quarcoo et al. 2022),(Devulapalli, Potti et al. 2023)]. Managing the size and complexity of contemporary datasets requires the development of increasingly complex preprocessing methods in conjunction with distributed computing frameworks.

The techniques and resources for processing large amounts of data will also change as technology does, creating new avenues for development and innovation. Future systems that make use of artificial intelligence (AI), cloud-based solutions, and other cutting-edge technologies like blockchain and quantum computing are probably going to change the way that big data processing is done. Quantum algorithms may be able to process complicated datasets more quickly than conventional systems in the developing field of quantum machine learning (QML), for example, offering a fresh approach to managing the enormous volume of Big Data [Zhou et al., 2022;]. On the other side, blockchain can improve data security and integrity, guaranteeing data privacy and reliability, particularly in distributed big data contexts [(Wei, Wang et al. 2020)].

Further promising methods for improving the comprehension of unstructured data, including text, photos, and videos, are hybrid approaches that combine machine learning models with computer vision and natural language processing (NLP). New applications in fields like healthcare, where combining data from clinical trials, diagnostic imaging, and medical records could improve patient outcomes, may result from this [Nguyen & Lee, 2023; Chen et al., 2022]. Similar to this, the industrial and logistics industries may greatly increase productivity by combining real-time data analytics with

automation and predictive maintenance. This will minimize downtime and optimize the supply chain.

Additionally, investigating novel hybrid techniques that integrate the advantages of many machine learning methodologies may provide more thorough answers to current problems. Advances in federated learning (FL), which allows models to learn across dispersed devices without combining raw data, can help with privacy issues, which are crucial in industries like healthcare and finance [Smith et al., 2023]. Recent international data privacy laws like the GDPR, which encourage firms to embrace safer data practices without sacrificing the quality of insights, are in line with this shift towards privacy-preserving machine learning [(Gupta and Sharma 2023)].

Furthermore, integrating machine learning (ML) into big data with international projects like the Sustainable Development Goals (SDGs) of the UN may encourage more socially conscious and sustainable solutions. SDG 11 (Sustainable Cities and Communities) can be addressed, for instance, by leveraging Big Data analytics to improve urban planning and make cities smarter and more sustainable [(Lemarchand, McKeever et al. 2022)]. Similar to this, predictive analytics in agriculture can help SDGs 2 (Zero Hunger) and 12 (Responsible Consumption and Production) by increasing agricultural yields, lowering waste, and promoting sustainable farming methods [Smith et al., 2023].

In conclusion, machine learning and big data have a shared future as new technologies constantly transform their uses. To advance this subject, it will be crucial to address the shortcomings of existing systems, including data heterogeneity, real-time processing, and scalability, while promoting an atmosphere of cooperation between government agencies, business, and academia. Leveraging the potential of big data and machine learning will be essential in propelling advancement in a variety of fields as breakthroughs continue to emerge, enhancing worldwide economic and societal results [29, 53, Zhou et al., 2021].

5. Conclusion

In order to handle complicated problem-solving requirements, machine learning integrates a variety of technologies, including statistical analysis, mathematics, data mining, deep learning, and artificial intelligence. Due to the inefficiency of traditional algorithms in handling such large volumes of data, the rise of big data has greatly increased the role of machine learning in data-intensive fields. Researchers are drawn to machine learning's advanced techniques in an effort to extract insightful information from big data, which is now essential for contemporary economic progress.

The sheer volume and complexity of big data has made data analytics an important field of study because it makes it easier to find trends and make data-driven decisions across industries. For instance, the rise of big data in finance has made financial scam more probable. The complexity of big data is often too complex for traditional methods, but machine learning and data mining offer potential alternatives by accurately identifying fraudulent activities. Big data-driven machine learning allows for prediction-based insights that are beneficial in a variety of sectors.

Nevertheless, there are still difficulties in creating machine learning algorithms, particularly when dealing with complicated datasets for categorization tasks. This study emphasizes that a significant obstacle is data imbalance, which causes misclassification to produce erroneous conclusions even when accuracy scores are high. To combat this, comprehensive pre-processing, also known as data preparation, is necessary to clean data, simplify it, and enhance machine learning results.

6. Future Scope

Future studies can concentrate on creating a brand-new single hybridization method that streamlines imbalanced big data and cleans it at the same time. A cohesive strategy like this could address several issues by:

1) Effectively eliminating outliers, cutting down on redundancy, and cleaning up messy big data.
2) Using better strategies that get around the drawbacks of conventional approaches to handle the complexity of unbalanced big data.
3) Providing a cleaner, more balanced dataset to improve the performance of different machine learning algorithms.

This all-encompassing method has the potential to revolutionize data preparation, increasing its efficacy and efficiency for practical uses.

Acknowledgements
For providing the resources necessary to conduct the current research, the authors would like to express their gratitude to Shobhit Institute of Engineering & Technology (Deemed to-be-University), Meerut-250110, Uttar Pradesh.

References

[1] Al-Marghilani, A. (2022). "Artificial intelligence-enabled cyberbullying-free online social networks in smart cities." International Journal of Computational Intelligence Systems **15**(1): 9.
[2] AL WAILI, A. R. (2023). "High Performance Scalable Big Data and Machine Learning using Apache Mahout." Journal of Misan Researches **13**(26-1).
[3] Alghunaim, S. and H. H. Al-Baity (2019). "On the scalability of machine-learning algorithms for breast cancer prediction in big data context." IEEE Access **7**: 91535-91546.
[4] Alhenawi, E. a., et al. (2022). "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." Computers in biology and medicine **140**: 105051.
[5] Ali, S. A. (2023). "Designing Secure and Robust E-Commerce Plaform for Public Cloud." The Asian Bulletin of Big Data Management **3**(1): 164-189.
[6] Almutiri, R., et al. (2022). "A survey of machine learning for big data processing." J. Big Data **4**(2): 97-111.
[7] Alqahtani, A. S., et al. (2023). "Effective spectrum sensing using cognitive radios in 5G and wireless body

area networks." Computers and Electrical Engineering **105**: 108493.
[8] Bansal, M., et al. (2020). "A survey on iot big data: current status, 13 v's challenges, and future directions." ACM Computing Surveys (CSUR) **53**(6): 1-59.
[9] Bharadiya, J. P. (2023). "A comparative study of business intelligence and artificial intelligence with big data analytics." American Journal of Artificial Intelligence **7**(1): 24.
[10] Bhatnagar, R. (2018). Machine learning and big data processing: a technological perspective and review. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018), Springer.
[11] Can, B., et al. (2020). "A closer look into the characteristics of fraudulent card transactions." IEEE Access **8**: 166095-166109.
[12] Cao, L., et al. (2023). "Autood: Automatic outlier detection." Proceedings of the ACM on Management of Data **1**(1): 1-27.
[13] Chen, Q., et al. (2023). "Benchmarks for measurement of duplicate detection methods in nucleotide databases." Database **2023**: baw164.
[14] Chong, A. Y. W., et al. (2023). "Customer churn prediction of telecom company using machine learning algorithms." Journal of Soft Computing and Data Mining **4**(2): 1-22.
[15] Demchenko, Y., et al. (2014). Defining architecture components of the Big Data Ecosystem. 2014 International conference on collaboration technologies and systems (CTS), IEEE.
[16] Devi, S. G. and M. Sabrigiriraj (2018). Feature selection, online feature selection techniques for big data classification:-a review. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), IEEE.
[17] Devulapalli, S., et al. (2023). "Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques." Materials Today: Proceedings **81**: 983-988.
[18] DHAMELIYA, N., et al. (2024). "Edge computing in network-based systems: Enhancing latency-sensitive applications." Journal of Computing and Digital Technologies **2**(1): 1-21.
[19] Dietrich, D., et al. (2015). Data science & big data analytics: discovering, analyzing, visualizing and presenting data, Wiley.
[20] Dridi, S. (2021). "Supervised learning-a systematic literature review." preprint, Dec.
[21] Fan, J., et al. (2014). "Challenges of big data analysis." National science review **1**(2): 293-314.
[22] Gupta, U. and R. Sharma (2023). A Study of Cloud-Based Solution for Data Analytics. Data Analytics for Internet of Things Infrastructure, Springer**:** 145-161.
[23] Han, S., et al. (2017). "Automatically Redundant Features Removal for Unsupervised Feature Selection via Sparse Feature Graph." arXiv preprint arXiv:1705.04804.
[24] Jin, X., et al. (2015). "Significance and challenges of big data research." Big data research **2**(2): 59-64.
[25] Keswani, B., et al. (2020). Adapting machine learning techniques for credit card fraud detection. International Conference on Innovative Computing and Communications: Proceedings of ICICC 2019, Volume 1, Springer.
[26] Ketu, S., et al. (2020). "Performance analysis of distributed computing frameworks for big data analytics: hadoop vs spark." Computación y Sistemas **24**(2): 669-686.
[27] Khalil, M. I., et al. (2020). "Challenges and opportunities of big data." Journal of Platform Technology **8**(2): 3-9.
[28] Kumar, S., et al. (2022). "Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research." Annals of Operations Research: 1-44.
[29] Laney, D. (2001). "3D data management: Controlling data volume, velocity and variety." META group research note **6**(70): 1.
[30] Lemarchand, P., et al. (2022). "A computational approach to evaluating curricular alignment to the united nations sustainable development goals." Frontiers in Sustainability **3**: 909676.
[31] Liu, Z., et al. (2022). "Towards understanding grokking: An effective theory of representation learning." Advances in Neural Information Processing Systems **35**: 34651-34663.
[32] Lu, W. (2020). "Improved K-means clustering algorithm for big data mining under Hadoop parallel framework." Journal of Grid Computing **18**(2): 239-250.
[33] Luengo, J., et al. (2020). "Big data preprocessing." Cham: Springer.
[34] Naeem, M., et al. (2022). Trends and future perspective challenges in big data. Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania, Springer.
[35] Najafabadi, M. M., et al. (2015). "Deep learning applications and challenges in big data analytics." Journal of big data **2**: 1-21.
[36] Nguyen, T., et al. (2023). "Vision-and-Language Pretraining: Methods, Applications, and Future Challenges."
[37] Noman, M. (2024). "Comparative Analysis of Big Data Processing in AWS and GCP Cloud Environments."
[38] Nti, I. K., et al. (2022). "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects." Big Data Mining and Analytics **5**(2): 81-97.
[39] Pérez Pupo, I., et al. (2020). Discovering fails in software projects planning based on linguistic summaries. International Joint Conference on Rough Sets, Springer.
[40] Philip Chen, C. and C.-Y. Zhang (2014). "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data."
[41] Press, G. (2014). "6 predictions for the $125 billion big data analytics market in 2015." Forbes, Dec.
[42] Pugliese, R., et al. (2021). "Machine learning-based approach: Global trends, research directions, and regulatory standpoints." Data Science and Management **4**: 19-29.

[43] Qiu, J., et al. (2016). "A survey of machine learning for big data processing." EURASIP Journal on Advances in Signal Processing **2016**: 1-16.
[44] Rahmani, A. M., et al. (2021). "Machine learning (ML) in medicine: Review, applications, and challenges." Mathematics **9**(22): 2970.
[45] Raj, R. J. S., et al. (2020). "Optimal feature selection-based medical image classification using deep learning model in internet of medical things." IEEE Access **8**: 58006-58017.
[46] Rathore, M. M., et al. (2021). "The role of ai, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities." IEEE Access **9**: 32030-32052.
[47] Rodionova, O., et al. (2021). "Efficient tools for principal component analysis of complex data—A tutorial." Chemometrics and Intelligent Laboratory Systems **213**: 104304.
[48] Ryzko, D. (2020). Modern big data architectures: a multi-agent systems perspective, John Wiley & Sons.
[49] Salihoun, M. (2020). "State of art of data mining and learning analytics tools in higher education." International Journal of Emerging Technologies in Learning (iJET) **15**(21): 58-76.
[50] Sarker, I. H. (2021). "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." SN computer science **2**(6): 420.
[51] Schneckenreither, M. (2020). "Average reward adjusted discounted reinforcement learning: Near-blackwell-optimal policies for real-world applications." arXiv preprint arXiv:2004.00857.
[52] Shafiq, M., et al. (2020). "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey." Sustainable Cities and Society **60**: 102177.
[53] Sutton, R. S. (1988). "Learning to predict by the methods of temporal differences." Machine learning **3**: 9-44.
[54] Tahmassebi, A., et al. (2017). An evolutionary approach for fmri big data classification. 2017 IEEE Congress on Evolutionary Computation (CEC), IEEE.
[55] Tang, J., et al. (2022). "Using domain adaptation for incremental SVM classification of drift data." Mathematics **10**(19): 3579.
[56] Uddin, M. F. and N. Gupta (2014). Seven V's of Big Data understanding Big Data to extract value. Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education, IEEE.
[57] Vurgaft, A. (2023). Multi-dimensional Data Visualization for Analyzing Materials. Australasian Conference on Data Science and Machine Learning, Springer.
[58] Wei, P., et al. (2020). "Blockchain data-based cloud data integrity protection mechanism." Future Generation Computer Systems **102**: 902-911.
[59] Wu, C., et al. (2023). "Personalized news recommendation: Methods and challenges." ACM Transactions on Information Systems **41**(1): 1-50.
[60] Zeng, L. and Z. Ge (2020). "Improved Population-Based Incremental Learning of Bayesian Networks with partly known structure and parallel computing." Engineering Applications of Artificial Intelligence **95**: 103920.
[61] Zhou, D., et al. (2021). Autospace: Neural architecture search with less human interference. Proceedings of the IEEE/ CVF International Conference on Computer Vision.