A R DIGITECH
International Journal Of Engineering, Education And Technology (ARDIJEET)
www.ardigitech.in ISSN 2320-883X, VOLUME 12 ISSUE 03 pub: 26/07/2024

A Mini-Review of Machine Learning in Big Data Analytics: Applications, Challenges, and Prospects

Prof. Sheetal Prakash Ingole*1

**1(Head of Department, Computer Department, Trinity Polytechnic Pune, Maharashtra, India)*
*Email id-meenakshi.yadav100@gmail.com*1*

Abstract: The availability of digital technology in the hands of every citizenry worldwide makes an available unprecedented massive amount of data. The capability to process these gigantic amounts of data in real-time with Big Data Analytics (BDA) tools and Machine Learning (ML) algorithms carries many paybacks. However, the high number of free BDA tools, platforms, and data mining tools makes it challenging to select the appropriate one for the right task. This paper presents a comprehensive mini-literature review of ML in BDA, using a keyword search; a total of 1512 published articles was identified. The articles were screened to 140 based on the study proposed novel taxonomy. The study outcome shows that deep neural networks (15%), support vector machines (15%), artificial neural networks (14%), decision trees (12%), and ensemble learning techniques (11%) are widely applied in BDA. The related applications fields, challenges, and most importantly the openings for future research, are detailed.

Key words: Big Data Analytics (BDA); Machine Learning (ML); Big Data (BD); Hadoop; MapReduce

1 Introduction

Huge volumes of data are being generated every day in a variety of fields, from social networks to engineering and commerce to biomolecular research and phycology[1,2]. Digital data generated from various digital platforms and devices are growing at astounding rates worldwide. In 2011, digital information grew nine times in volume compared with 2006, and it is estimated to reach 44 zettabytes by 2020[1,3]. As of 16th December 2020, the volume of daily generated data globally was 59 zettabytes. It is anticipated to reach 149 zettabytes in 2024 as we go into an even more data-driven future.

The escalating volume in data is the principal attribute of "big data", a jargon that has become a household name in the research communities, organisations, and the Internet.
Recently, Big Data (BD) and its emerging machineryand techniques, like Big Data Analytics (BDA), have transformed the way that organisations and businesses operate, delivering new significant prospects for enterprises, professionals, and academia[5]. Besides businesses and research institutions, governmental and non-government organisations now regularly generate massive unique scope and complexity data. Therefore, picking up meaningful information and valuable advantages from these available big data has become vital to organisations worldwide. However, the literature shows that it is challenging to efficiently and skillfully derive helpful insights from BD quickly and easily[8]. So, BDA has become indistinguishably essential to realise BD's total value to improve business performance and increase market share to most organisations.
Even though most Artificial Intelligence (AI) and Machine Learning (ML) algorithms and their enabling platforms for performing BDA are free, they require a new skill set that is uncommon to most practitioners in
this field and organisations' IT departments. Hence, integrating these tools and platforms seamlessly into anorganisation's internal and external data on a common platform is a challenge.
Also, the availability of several ML algorithms possesses a challenge in making a good choiceout of them, i.e., "searching for a needle in a haystack". Therefore, performing a comprehensive comparative analysis of BDA in different industries with ML algorithms is necessary. Additionally, big datacome in a

different format (structured, semi-structured, or unstructured) regarding the data source or industry. Subsequently, has proven that ML algorithms perform differently concerning input data format. Thus,an ML algorithm might fit better (high accuracy)on a structured dataset than on a semi-structured or unstructured dataset. It was also evident in Nti et al[1] that an ML algorithm might perform differently under different ML tasks. For example, the same algorithm capable of regression or classification task might perform better in classification than regression.

Hence, this paper investigates various literature on big data analytics with ML algorithms. We sought to review journal and conference published research papers relating to BDA using ML methods, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Deep Learning (DL), K-Nearest Neighbour (KNN), and many more, by (i) discussing some critical issues related to big data analytics, and highlighting current research efforts and the potential challenges in BDA and future openings and trends, and

(ii) identifying various ML techniques for BDA from different viewpoints.

The worth of the current study are as follows:

(i) This study will help researchers, IT departments, and professionals appreciate the right BDA tools and algorithms when carrying out big data analytics. (ii) It will also help new researchers in BDA make an informed decision and a helpful contribution to the scientificcommunity. (ii) The outcome of this study will serve asa guide to the enhancement of techniques and tools thatblend big data and cognitive computing.

The main contribution of the current study are as follows:

An all-inclusive and detailed valuation of previousstate-of-the-art studies on BDA with ML techniques; based on a novel taxonomy (i.e., the type BDA, data size, study origin, ML task and method, and evaluation

(1) A concise representation of the valuable features of compared techniques in BDA with ML;

(2) A concise representation of the valuable features of compared techniques in BDA with ML;

(3) We finally provide the potential challenges, research trends, and opportunities for future studies in BDA.

We organised the remaining sections as follows: Section 2 presents the review of the literature and related works. Section 3 studies methodology, Section 4 outlines the study results and discussions. Finally, we conclude the study in Section 5.

2 Research Literature

This section presents the concept of big data, big data analytics, and a review of related works.

2.1 Concept of big data and big data analytics

According to Sujitparapitaya , BD is the gathering of data in huge volume enabled by the recent advances made in technologies tools and platforms that support high-velocity data capture, storage, and analysis. The concept of BD is branded by volume, velocity, and variety, acknowledged as 3Vs. However, most studies expand the concept of Doug Laneyto five key characteristics (5 Vs), namely, volume, velocity, variety, value, and veracity (see Fig. 1), i.e., the definition for BD keeps varying following the advancement in technology, storage capacity for data, the transmission rate of data, and other system abilities. The first "V" (volume) denotes the data size,which swells exponentially with time. It is argued that the healthcare industry generates enormous amounts of data in electronic medical records compared with most industries' The second "V" (velocity) refers to the swiftness at which data are generated and acquired from various industries. The third "V" (variety) denotesthe multiplicity and heterogeneity of data. The fourth "V", value, to some researchers, is the most vital and

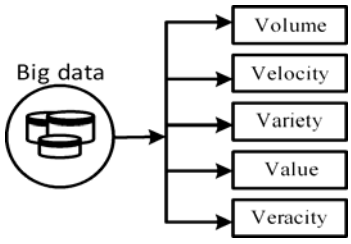


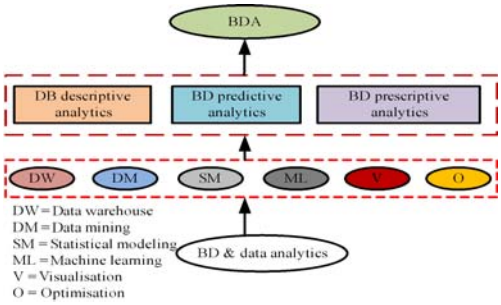
Fig. 1 General concept of BD.

irreplaceable characteristic of all the 5 Vs of BD, as they believe it has the power to transmute industry datainto a piece of valuable information. The fifth "V" (veracity) refers to the credibility of the data, whichin this context is very similar to quality assurance of data. It gives a degree of genuineness about a particularsector knowledge.

According to Russom, BDA is applying advanced analytical methods and techniques on big datasets. Similarly, BDA can be defined as the process of collecting, systematising, and scrutinising DB to envisage and display patterns, discover knowledge and intelligence along with other information in the BD Thus, BDA practically involves two things, big data and analytics, and how these two have teamed up to create one of the overwhelming current trends in Business Intelligence (BI). BDA consists of BD descriptive, BD predictive, and BD prescriptive analytics (see Fig. 2). Thus, BDA uses data analysis techniques to uncover patterns in call logs, mobile banking transactions, and online user-generated content. The ground rules of BDA consist of engineering, mathematics, human interface, statistics, information technology, and computer science.

Currently, BDA is the new big data technology that has become widely embraced across sectors,

Fig. 2 A taxonomy of BDA



companies, geographic areas, as well as among individuals, to help businesses and individuals make data-driven decisions to accomplish desired business goals Of late, BDA can be enabled by several analytic platforms and tools, including those based on Structured Query Language (SQL) queries, fact clustering, data mining, natural language processingstatistical analysis, data visualization, AI, ML, text analytics, MongoDB, Hadoop, and Map Reduce. The platforms and tools available to handle the 5Vs of big data have improved dramatically in recent years. In general, these technologies and tools are not absurdly expensive, and much of the available software is open source.

Figure 3 shows the basic applied theoretical architecture of BDA. ML algorithms have becomedominant in analyzing, visualizing, and modelling big data. ML makes machines learn from a dataset in its basic definition, apply their knowledge and insight on unseen data, and make predictions. The literature reports the success of ML algorithms in different application areas (see Table 1).

From the literature, ML can be categorised into four classes, namely: (i) supervised learning, (ii) unsupervised learning, (iii) semi-supervised learning, and (v) reinforcement learning. Countless ML algorithms are available freely for performing ML tasks, like classification, regression, clustering, dimensionality reduction, and ranking. To name a few, SVM, ANN, DT, Naive-Base (NB), Tensor Auto-Encoder (TAE), Ensemble Learning (EL), KNN, Hidden Markov Model (HMM), Singular-Value Decomposition (SVD), Radial Basis Function Neural Network (RBF-NN), Principal Component Analysis (PCA), Generative Adversarial Networks (GANs), Natural Language Processing (NLP), Recurrent Neural Network (RNN), Bidirectional Gated Recurrent Unit.

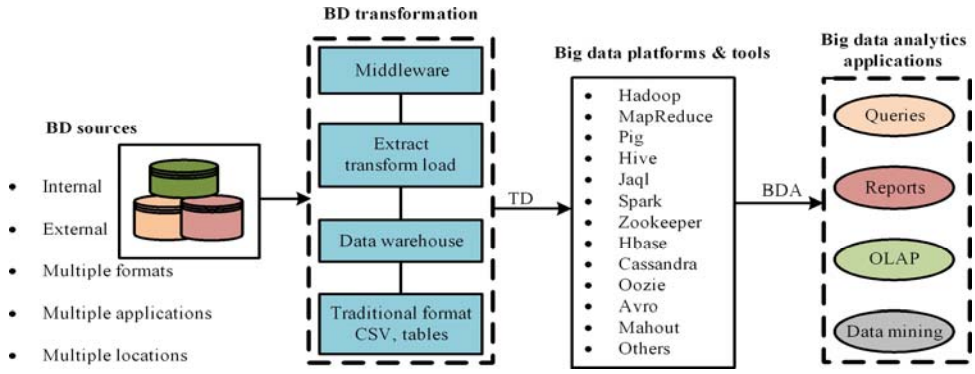


Fig. 3 A theoretical architecture of BDA. Here TD represents transformed data.

Table 1 Machine learning application in different economicsectors by academicians and industry professionals.

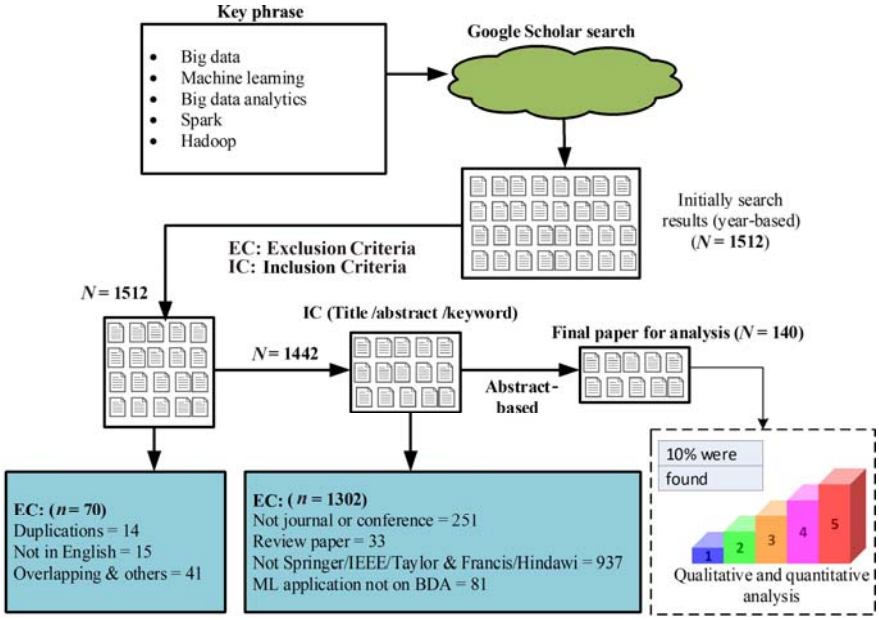
Reference	Application area
[10, 20–23]	Finance and stock market
[24–26]	Energy system forecasting and faults detection
[27–29]	Healthcare
[30–33]	Teaching and learning
[34–36]	Agriculture (crop yields, emissions, and diseasedetection)
[37]	Transportation
[38]	Petrology

(Bi-GRU), Generalized Discriminant Analysis (GDA), Deep Q Network (DQN), General Regression Neural Network (GRNN), Feed-Forward Neural Network(FNN), Long Short-Term Memory (LSTM), Deep Autoencoder Network (DAN), MultiLayer Perceptron (MLP), Extreme Learning Machines (ELM), DL, and Deep Belief Network (DBN).

3 Methodology

According to Lorna[67] and Elfar[68], MLR seeks to quickly and easily showcase, a specific issue or set of related topics, and emphasise where there are gaps in the literature and possibilities for further research. An MLR is typically shorter than a full-fledged literature review. That is because, unlike a literature review, which focuses on synthesising findings from several studies to develop conclusions on a broad area of research, an MLR focuses on a single subject or topic. It is vital to remember that the literature review and the MLR formatare the same. Likewise, there is no substantial differencebetween the steps involved in MLR and the literature review. However, the only difference between the two is that one is broader while the other is narrower. We adopted MLR because its concise format makes it simple to grasp those themes in the literature, allowing more practitioners to profit from them. Figure 4 shows the review process in this paper; five guidelines are followed, i.e., (i) search strategy, (ii) selection criteria, (iii) study selection process, (iv) quality assurance, and (v) qualitative and quantitative analysis. Finally, we explain in detail what is accomplished in each step. Google Scholar† was adopted as the central search engine platform for collecting relevant articles due

to its open access and its date restriction flexibility. However, only relevant journal and conference articles were downloaded. Five principal phrases defined by the authors were used in the search, "big data", "machine learning", "big data analytics", "Apache Spark", and "Hadoop". However, we obtain several related queries to our five keywords using Google trends. The following are a few of them that were adopted in these study as auxiliary words "big dataand data analytics", "analytics of big data", "business analytics", "big data business analytics", "analytics big



data", "data analytics", "analytics", "Hadoop", "big data Hadoop", "deep machine learning", "deep learning", "Hadoop spark", "spark Apache tutorial", "Scala", and "Hadoop hive". Figure 5 shows the trend on big data analytics from Google trends.

The articles were considered based on an agreed inclusion-exclusion criterion by all authors. The inclusion criteria are as follows: (i) the article is writtenin the English language, (ii) the article must relateto big data and DBA, (iii) article published between 2010–2021, and (iv) article must be published in a journal or conference. While the exclusion criteria were

(i) articles not published within 2010–2021 and (ii)papers published not in a journal or conference. Also, review-based articles on BDA, papers not published in Elsevier, Springer, Taylor & Francis, IEEE, and Hindawi, and ML applications where dataset and tools used do not fall into the concept of big data were excludedfrom the qualitative and quantitative analysis of this

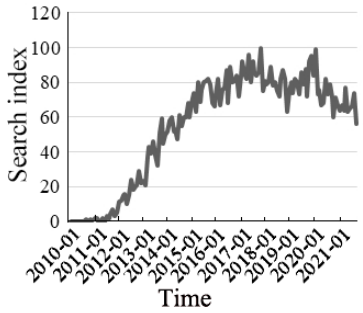


Fig. 5 Trend on Google search queries on big data analytics.

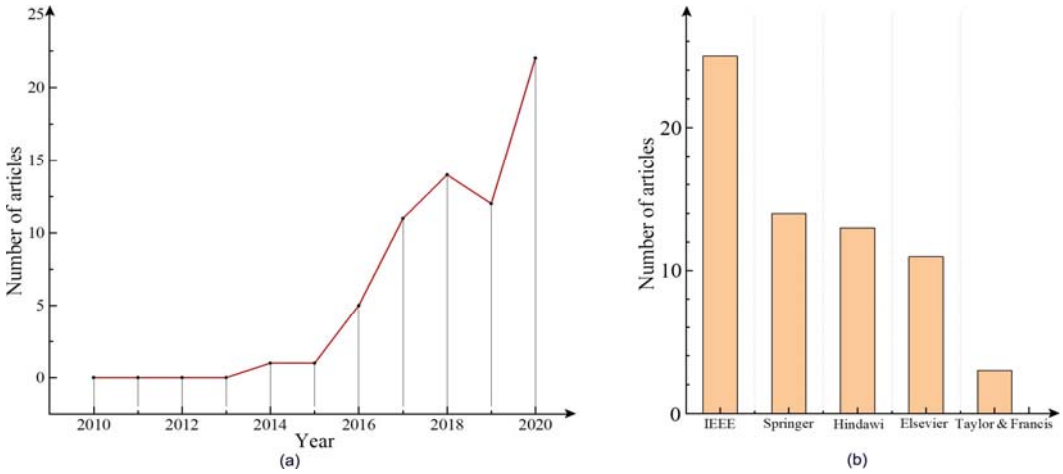
study. Initially, all the articles relevant to the current study (big data and BDA) were carefully chosen in the primary screening phase. Based on our inclusion- exclusion criteria defined above, the downloaded papers(1512) were screened in Stage one (see Fig. 4), and inappropriate papers, i.e., papers not published in English and duplication papers, overlapping papers, were excluded. We further screened the remaining based on its title, abstract, publishers and publication type, andpapers that were not connected with the proposed studywere discarded. Finally, these articles were filtered on the basis of abstracts using the Boolean AND operator on all of the defined search terms in the final step of screening.
As a result of the comprehensive screening, one hundred and forty (140) articles pertinent to the research domain were selected from the 1512 initially downloaded articles. Of the 140 articles, 74 were analysed qualitatively and 66 quantitatively. Quality evaluation plays a substantial role in a systematic literature review procedure. Therefore, all authors of this study did the Quality Assessment (QA) of papers after analysing and evaluating abstracts of selected papers. Some of the QA criteria were as follows: (i) Is the researcher objective of the article clear? (ii) Is the methodology effectively applied? (iii) Are the results undoubtedly explained? and (iv) Is there an associationbetween the introduction, results, and conclusion? Figure 4 shows the details of articles excluded and included.

4 Result and Discussion

This section presents the outcome of the 66 articles thatwere analysed quantitatively. The review shows that current research on BDA can be categorised under five different themes, namely, (i) core BD area to handle thescale, (ii) managing noise and vagueness in the data, (iii)privacy and security aspects, (iv) data engineering, and
(v) rendezvous of BD and data science. Table A1 in Appendix summarises the papers reviewed in this study;it presents the application area, the papers' objective, and the data size used.
Based on the ontology proposed in Sun et al. (see Fig. 2), this study grouped the type of big data analytics into three, namely, (i) BD descriptive analytics (denotedas "A"), (ii) BD predictive analytics (denoted as "B"), and (iii) BD prescriptive analytics (denoted as "C"). Outof the 66 papers, 56 indicated the application domain.
It was observed that 58% of the reviewedpapers were based on BD predictive analytics,18% BD prescriptive analytics, 11% BD
Hence, it is not a shock to see such a massive studyin the healthcare industry (30%), followed by anomaly detection (11%), cybersecurity, data privacy & IoT (5%)and automobile and transportation (5%); (see Table A1in Appendix). Concerning the data size, some studies indicated their data size in terms of the storage space ranging from 708 MB to 600 GB, while in terms of the number of observations, it ranges from 1789to 3 billion records[112]. Based on the data size (e.g., 708 MB), one can say that the study by Jallad et al.is not related to big data. However, we believe that the data size used in research only does not classify it as a big data study. However, the tools and platforms employedfor the empirical analysis also count. Detailed study results based on the proposed taxonomy are presented inthe following sections of this paper.

4.1 Time trend publication

Figure 6 shows the time trend in publication and publisher wise distribution. Although the review limited the literature search between 2010 – 2020, it was observed that research work in ML application in BDA started receiving a rise in attention from academicians and professionals in the last five years and since has increased progressively, supporting the report . Nonetheless, BD has been around for decays; however, it has only taken on from a word-buzzed in recent years. Furthermore, it can be inferred that the tremendous rise in BD in current years[4] has attracted researchers' attention to examine the benefit that canbe effectively derived from this availability of datato make an informed decision. Figure 6b shows the



publisher wise distribution; of the 66 articles, 38% were published in IEEE, 21% Springer, 20% Hindawi, whiles 17% and 5% were published in Elsevier and Taylor & Francis, respectively. The results suggest thathigh impact publication houses have seen the need to make available big data analytics with ML applicationsto the scientific community.

4.2 Big data platforms tool in BDA

Table 2 shows the typically used DBA platforms andtools. Out of the 66 articles, 43 (65%) indicated the BDA tools used for their experimental analysis. Even though Hadoop is believed to be the most potent and popular tool in BDA, our results show otherwise.We observed that Spark is the top most used (34.88%)tool for DBA among researchers in this field, followed by Hadoop (30.23%). This finding can be attributed to Spark being faster and easier to utilize for big data analytics than Hadoop Map Reduce. Also, it is believed that Spark offers high processing speed than Hadoop[103]

On the other hand, a comparative study shows that Spark consumes more memory in operation than Hadoop since it loads all processes in memory and keeps them incaches for some time. Therefore, this paper recommends choosing between these two platforms to be grounded on different features. Like, cost, ease of use, memory limitations, fault tolerance, performance level, type of data processing, and security show their appropriateness for a project at hand and organisation needs present and future. In summary, the Spark and

Table 2 BDA platforms and tools.

BD platforms and tools	Number of papers	Percentage (%)
Flink	1	2.33
Apache Mahout	1	2.33
HiBench	1	2.33
H2O	1	2.33
MATLAB	5	11.63
MapReduce	6	13.95
Apache Hadoop	13	30.23
Apache Spark	15	34.88

Apache frameworks' open-source has seen massive market expansion, as more firms and researchers have found the saccharine spot to adopt these platforms.

Figure 7 shows the ML techniques mainly used for BDA. The outcome suggests that artificial intelligence and ML will continue to hit the roof as more firms and industries look forward to

transforming their day-to-daybusiness, maximising profit, while minimising risk. TheSVM is found to be the most widely (14 out of 66)used machine learning algorithms in big data analytics because of its ability to work (i) virtually without prior knowledge of the dataset and (ii) on high dimensional and the risk of over-fitting. The Decision Tree (DT) algorithm, though simple, is the third most used ML algorithm in BDA.

Deep NeuralNetworks (DNN), such as LSTM and Convolutional Neural Network (CNN), as shown in Fig. 6. This outcome can be attributed to LSTM's storing memory and solving the gradient vanishing problem; CNN can automatically notice and extract the appropriate internalstructure from a time series dataset to create in-depth input features, using convolution and pooling operations. Also, CNN and LSTM algorithms are resilient to noise tolerance and accuracy for time-series classification.

Suppose we consider the situation where these techniques were hybrid with other techniques, approximately 25 out of 66 adopted DNN. It can be saidthat BDA and DL seem to be dependent on one anotherand show a reciprocally beneficial association. Large amounts of data allow DL techniques to realise better

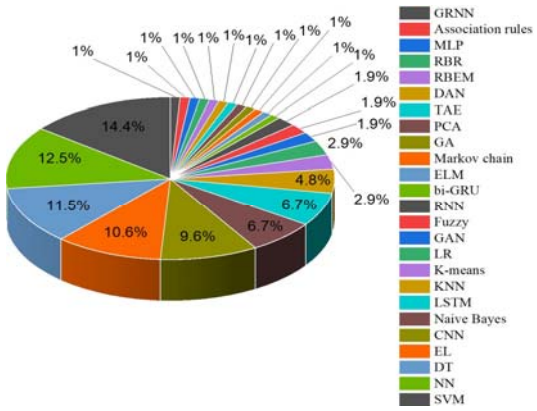


Fig. 7 ML techniques mainly used for BDA.

generalisation; thus, yielding meaningfully and more valuable results in the field of BDA. Hence, in support of Ref. [16], this paper suggests that ML techniques, such as DNN, are worth exploring in future works to gain the acceptance of the platforms it has received little attention in BDA. Therefore, future studies can virtuallyexplore this technique to examine its ability in BDA.

methods, like the random forest, boosting, and bagging, to enhance the power of EL techniquesin BDA. ML algorithm's hybridisation is an excellenttechnique to compensate for the weakness in the evident that most ML algorithms apply to BDA. However, two areas require further enhancement: (i) the huge computational cost associated with most ML algorithms and (ii) the communication cost for diverse computer nodes in parallel computing.

4.3 Country-wise distribution of publications

Figure 8 shows the distribution of papers across countries. It was observed that most of the studieswere undertaken in China (36.4%), followed by the USA (18.2%). Though Srivastava reports that the USA leads China interns of DB adaptation in 2019, this studyshows that more studies in BDA analytics evolve from China than USA. This outcome is no surprise since a report by www.statista.com shows that China has the highest mobile phone users, followed by India and USA.The outcome is no surprise since China's population is 18.47% of the total world population[§]. Therefore, it can

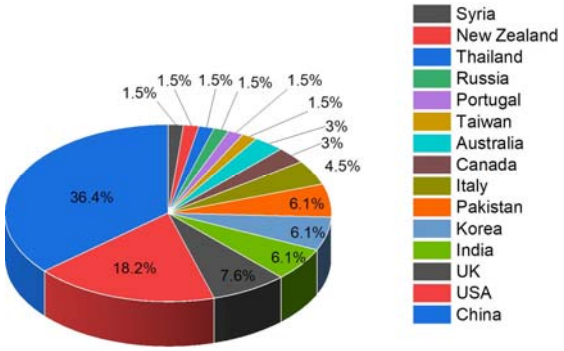


Fig. 8 Publication origin distribution.

be inferred that China generates more electronic data from mobile phone users than any other country; hence,more research is needed in big data. Interestingly, in the 66 papers, Africa is recorded none; therefore, we encourage analytics in the big data generated from the continent to enhance business decisions.

4.4 Evaluation metrics used in BDA

Several evaluation metrics can be used to measure a machine learning model's performance, depending on the ML task. Some of the most common are Mean- Square-Error (MSE), Root-Mean-Square-Error (RMSE), Mean-Absolute-Error (MAE), accuracy, precision, recall, Area-Under-the-Curve (AUC), and F-score. For more details and definitions of the various evaluation metric, readers are referred to Ref. [9]. Figure 9 shows the distribution of error metrics used in BDA. It was observed that accuracy (37%) was the most used metricbecause most papers were BD predictive analytics. Therefore, it is argued that combining accuracy and error metrics offers a better ML model evaluation

combined two or more metrics to evaluate their modelsand framework.

be inferred that China generates more electronic data from mobile phone users than any other country; hence,more research is needed in big data. Interestingly, in the 66 papers, Africa is recorded none; therefore, we encourage analytics in the big data generated from the continent to enhance business decisions.

4.5 Evaluation metrics used in BDA

Several evaluation metrics can be used to measure a machine learning model's performance, depending on the ML task[9]. Some of the most common are Mean- Square-Error (MSE), Root-Mean-Square-Error (RMSE), Mean-Absolute-Error (MAE), accuracy, precision, recall, Area-Under-the-Curve (AUC), and F-score. For more details and definitions of the various evaluation metric, readers are referred to Ref. [9]. Figure 9 shows the distribution of error metrics used in BDA. It was observed that accuracy (37%) was the most used metricbecause most papers were BD predictive analytics. Therefore, it is argued that combining accuracy and error metrics offers a better ML model evaluation.
combined two or more metrics to evaluate their modelsand framework.

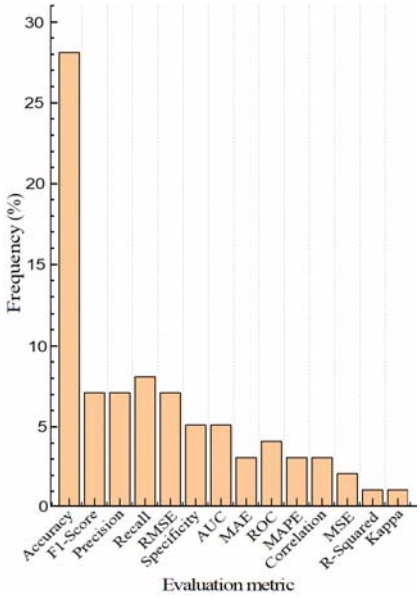


Fig. 9 Evaluation metric distribution

research papers in BDA. The aim was to measure the correlation between keywords used in big data analytics.A total of 360 words were obtained after pre-processing;it is essential to note that words with a frequency less than or equal to 2 were exempted from the plot. As seen in Fig. 10, big data, machine learning, analytics, network, and Spark are frequently used by research papers in BDA. A careful look at Fig. 9 shows that Spark is more significant than Hadoop; this affirms the results discussed in Table 2, that Spark is commonly used among researchers in BDA than Hadoop. From Fig. 10, it is evident that there is a high relation among the keywords used by different researchers in BDA.

5 Conclusion

The current studies reviewed past studies on big data analytics with machine learning applications in different economic areas. A total of 1512 papers published in journals and conferences were downloaded using keywords search in Google scholar. The downloaded papers were screened through several stages, and the final selected papers (140) were reviewed based on a proposed taxonomy. Based on the analysis we have presented in previous subsections and the summarised information in Tables 2 and A1 and Figs. 6 – 10; itis obvious to envision the most used tools for thisreview, the trend of research work over the years, besides mentioning several openings for future works. We hope that this study will aid researchers and industry professionals with a valuable base for further studiesto comprehend the complete context of big data analytics with machine learning and its applications in several industries. The following section puts forward asummary of challenges and openings.

5.1 Key challenges in BDA

Big data analytics challenges have been echoed in past studies of similar objectives as this paper; however, the following are few identified to add

Fig. 10 Keywords commonly used in BDA research

up or affirm literature.

(1) Ethical issues: It was observed that there are several ethical issues, such as copyright (reusability without permission), privacy, and involvement of the rival organisation associated with big data. This outcome confirms the report of Haq et al.

(2) Skills are another crucial challenge; big data analytics involves teamwork with expert and skilled personnel from different areas, such as computer science, biologists, mathematics, physicians, economists, etc., for a particular task. Also, infrastructures, such as software and hardware (preferably supercomputers) for handlingbig data analytics, make it initially expensive for small-scale and medium-scale businesses and organisations insmall-income and medium-income countries.

(3) Currently, big data analytics has becomea hot topic for discussion among scientists and industry professionals worldwide[14,15] because of the emergence of new sophisticated technologies, tools, andplatforms. However, several studies have pointed out the challenge associated with handling this massive and unprecedented amount of data made upof different data types and even streaming data as well[12]. Thus, big data are high-dimensional, diverse, gigantic, complex, incomplete, amorphous, noisy, and erroneous,making data pre-processing difficult in BDA. However, it is essential to make machine learning models effectivelyperform well with high accuracy. Nevertheless, this paper can affirm that this challenge still exist in BDAtoday.

5.2 Prospects for future studies

Several openings that need to be tackled by futureresearch efforts were identified; we discuss a few.

(1) Big data analytics on social media has been applied in developing countries' political arena to enhance campaign strategies. For example, the Obama campaignput up a model of classifying individual voters that led their campaign strategy to an efficacious result. However, it is not assured that these models can be pass-on to elections and other political activities and behaviour in underdeveloped and developing countries globally. Since individual sentiments and expression on social media are ethically influenced by several culturalperspectives.

Mobile cellular networks are one sector that has gotten little attention in big data analytics but creates and transmits large amounts of data on a regular basis.

However, DBA can enhance the performance of MCN and maximise the profits of operators. Hence, an opportunity for further studies in this area. Similarly, exploring different cross-domain datasets for cellular traffic forecasting and knowledge transfer amid different cities is an exciting direction in the future.

(1) Some studies have looked at traffic management based on big data analytics. However, a high percentageof these works concentrate on a single data source.However, with the increases in several data sources in the transportation industry, we expect future research tocombine data from diverse data sources, like weather information, security cameras, and other transportation-related data sources.

It is observed that BDA platforms' developers usually provide their expected users with highly configurable constraints (parameters) to enhance their frameworks' forcefulness and performance. Nevertheless, these constraints' complexity and high- dimensionality make manual tuning in BDA on these frameworks time-consuming, challenging, and unproductive. Likewise, It was pointed out that Hadoop has more than 190 configuration constraints, which can substantially affect the performance and

timeconstraints of the Hadoop framework. Thus, leading to several challenges in putting up high-performance models for DB frameworks. Hence, future studies can explore methods to make automation's configuration process more flexible.

(1) Even though BDA has seen rapid growth in recentyears (see Fig. 6), the study by Shahbaz et al.revealed no gender balance in the adaptation of BDA in most industries. That is, males, as compared with females, are dominant towards the positive intent to useBDA. On the other hand, the same report said females create more resistance to change than males while adopting BDA in healthcare and allied organisations. Hence since BDA has come to stay with us, more advocacy and orientation studies should be carried out in the future to propel gender balance in BDA.

(2) To facilitate easy handling of big data, identified issues, incomplete and diverse data sources, noisy and erroneous data that affect data analytics' performance need to be addressed going forward. Therefore, big data analytics designers need to highly and efficiently automate the data pre-processing (e.g., data clean, sampling, and compression) stage where possible with less human effort.

References

1) Z. H. Sun, L. Z. Sun, and K. Strang, Big data analytics services for enhancing business intelligence, *J . Comput*.
Inf. Syst., vol. 58, no. 2, pp. 162–169, 2018.

2) J. Zakir, T. Seymour, and K. Berg, Big data analytics,*Issues Inf. Syst*., vol. 16, no. 2, pp. 81–90, 2015.

3) C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S.Kim, Development of heavy rain damage prediction model using machine learning based on big data, *Adv. Meteorol*.,vol. 2018, p. 5024930, 2018.

4) L. Collins, Mini literature review: A new type of literature review article, https://www.emeraldgroup publishing.com/archived/products/journals/call for papers. htm%3Fid%3D5730, 2021.

5) J. C. Elfar, Introduction to mini-review, *Geriatr. Orthop.Surg. Rehabil*., vol. 5, no. 2, p. 36, 2014.

6) K. A. Jallad, M. Aljnidi, and M. S. Desouki, Anomaly

7) detection optimization using big data and deep learning to reduce false-positive, *J . Big Data*, vol. 7, no. 1, p. 68, 2020.

8) W. Gu, K. Foster, J. Shang, and L. R. Wei, A game-predicting expert system using big data and machine learning, *Expert Syst. Appl*., vol. 130, pp. 293–305, 2019.

9) T. Daghistani, H. AlGhamdi, R. Alshammari, and R. H. AlHazme, Predictors of outpatients' no-show: Big dataanalytics using apache spark, *J . Big Data*, vol. 7, p. 108, 2020

10) T. Nibareke and J. Laassiri, Using Big Data-machine
11) learning models for diabetes prediction and flight delaysanalytics, *J . Big Data*, vol. 7, p. 78, 2020.
12) N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A.
13) Rashid, A comprehensive performance analysis of ApacheHadoop and Apache Spark for large scale data sets using

14) F. Saeed, Towards quantifying psychiatric diagnosis using machine learning algorithms and big

fMRI data, *Big DataAnal*., vol. 3, no. 1, p. 7, 2018.

15) N. Bharill, A. Tiwari, and A. Malviya, Fuzzy based scalable clustering algorithms for handling big data usingapache spark, *IEEE Trans. Big Data*, vol. 2, no. 4, pp. 339–352, 2016.

16) L. Gu and H. Li, Memory or time: Performance evaluation for iterative operation on hadoop and spark,in *Proc. 10th Int. Conf. High Performance Computing and Communications & 2013 IEEE Int. Conf. Embeddedand Ubiquitous Computing*, Zhangjiajie, China, 2013, pp.721–727.

17) [114]Y. Samadi, M. Zbakh, and C. Tadonki, Comparative study between Hadoop and Spark based on Hibenchbenchmarks, in *Proc. 2nd Int. Conf. Cloud Computing Technologies and Applications*, Marrakech, Morocco, 2016, pp. 267–275.

18) Y. Samadi, M. Zbakh, and C. Tadonki, Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks, *Concurr. Comput.: Pract. Exp.* vol. 30, no. 12, p. e4367, 2018.

19) D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera, Online incremental machine learning platform for big data-driven smart traffic management, *IEEE Trans. Intell. Transp. Syst*., vol. 20, no. 12, pp. 4679–4690, 2019.