BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/06 pp81–97
Volume 5, Number 2, June 2022
DOI: 10.26599/BDMA.2021.9020028

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

Isaac Kofi Nti*, Juanita Ahia Quarcoo, Justice Aning, and Godfred Kusi Fosu

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[4] 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 machinery and 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[6, 7]. 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

• Isaac Kofi Nti is with the Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani BS2103, Ghana. E-mail: isaac.nti@uenr.edu.gh.
• Juanita Ahia Quarcoo is with the Department of Electrical & Electronic Engineering, Sunyani Technical University, Sunyani BS2103, Ghana. E-mail: juan83p@yahoo.com.
• Justice Aning and Godfred Kusi Fosu are with the Department of Computer Science, Sunyani Technical University, Sunyani BS2103, Ghana. E-mail: aning421@gmail.com; godfred17@gmail.com.
* To whom correspondence should be addressed.

Manuscript received: 2021-10-14; revised: 2021-12-11; accepted: 2021-12-13

this field and organisations' IT departments[3]. Hence, integrating these tools and platforms seamlessly into an organisation'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 choice out 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 data come in a different format (structured, semi-structured, or unstructured) regarding the data source or industry. Subsequently, Ref. [9] 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.[10] 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 scientific community. (ii) The outcome of this study will serve as a guide to the enhancement of techniques and tools that blend big data and cognitive computing.

The main contribution of the current study are as follows:

(1) An all-inclusive and detailed valuation of previous state-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 metrics);

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

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

(4) 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 et al.[8], 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 given by Doug Laney cited in Refs. [8, 11, 12] is branded by volume, velocity, and variety, acknowledged as 3Vs. However, most studies[1,7,10] expand the concept of Doug Laney to 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[11]. The first "V" (volume) denotes the data size, which swells exponentially with time[4]. It is argued that the healthcare industry generates enormous amounts of data in electronic medical records compared with most industries[11]. The second "V" (velocity) refers to the swiftness at which data are generated and acquired from various industries. The third "V" (variety) denotes the multiplicity and heterogeneity of data. The fourth "V", value, to some researchers, is the most vital and

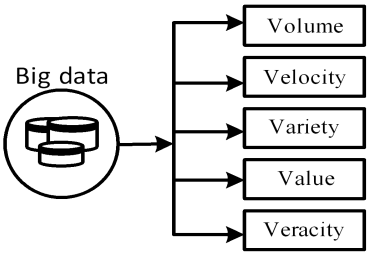

Fig. 1 General concept of BD.

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

According to Russom[13], 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[5]. 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[5, 14] (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, companies, geographic areas, as well as among individuals, to help businesses and individuals make data-driven decisions to accomplish desired business goals[5, 15–18]. 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 processing statistical analysis, data visualisation, AI, ML, text analytics, MongoDB, Hadoop, and MapReduce. 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.

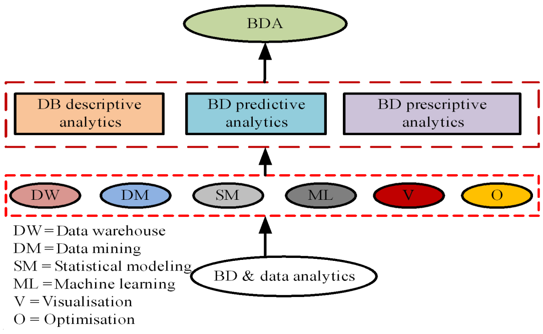

Fig. 2 A taxonomy of BDA.

Figure 3 shows the basic applied theoretical architecture of BDA[11, 19]. ML algorithms have become dominant in analysing, visualising, 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

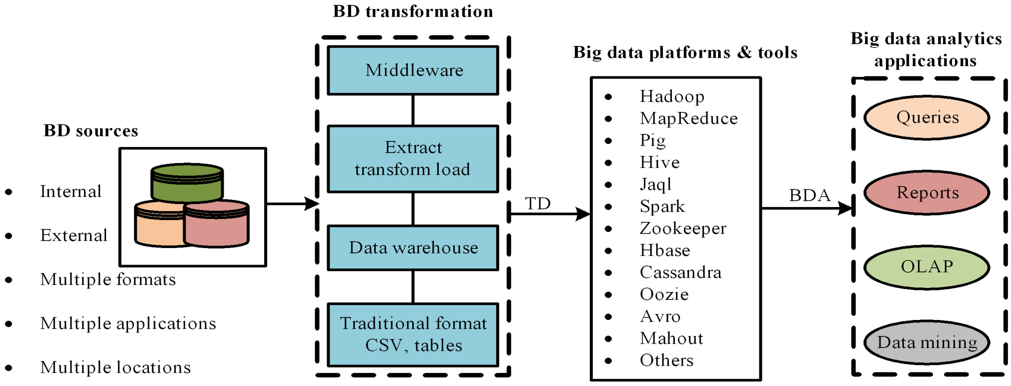

Fig. 3 A theoretical architecture of BDA. Here TD represents transformed data.

Table 1 Machine learning application in different economic sectors 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 disease detection)
[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).

2.2 Related works

The vast number of BDA papers makes it difficult for practitioners and researchers to keep up with developments in the field. Therefore, some past studies attempted to summarise different ML applications in BDA and its advancement to help novices select the correct ML algorithm for BDA.

However, quite a number of these studies were narrowed to BDA on specific industries, such as healthcare[17,39–45], air quality[46], Internet of Things (IoT)[47–49], agriculture[50], and information security[51]. Similarly, Refs. [1, 7, 13, 19, 52–61] focused on the overview, challenges, and approach in BDA. Likewise, Ale[62] presented the risk analysis of BD. In contrast, Ref. [63] covered BDA in diverse areas yet focused on model efficiency and computational cost. Furthermore, the work of Chong and Shi[64] provided an overview of big data analytics' content, scope, and findings, and discussed its future evolution. Finally, Fathi et al.[65] reviewed BDA in weather forecasting.

According to our knowledge, little has been done on ML advancement and big data analytics in diverse industries, like healthcare, agriculture, energy, engineering, and more. None of the earlier studies has adequately addressed this challenge; however, big data make waves across every economic sector. Hence, it will be unfair to narrow the big data analytics review in a single or dual industry. Furthermore, in light of the aforementioned efforts, we could identify the following flaws in prior works:

(1) The paper selection process in some papers, such as Refs. [1,63,66], was not clear.

(2) Most of the studies, like Refs. [1,7,13,19,52–61], did not consider the origin of the papers.

(3) None of the existing demonstrated clear statistical information on the BDA platforms and modelling tools.

(4) None of the studies discussed above considered the objectives of the papers they reviewed.

As a result of the stated reasons, we are happy to propose a Mini Literature Review (MLR) report that addresses all of the flaws mentioned above.

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 format are the same. Likewise, there is no substantial difference between the steps involved in MLR and the literature review[67,68]. 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 data and data analytics", "analytics of big data", "business analytics", "big data business analytics", "analytics big

† https://scholar.google.com.

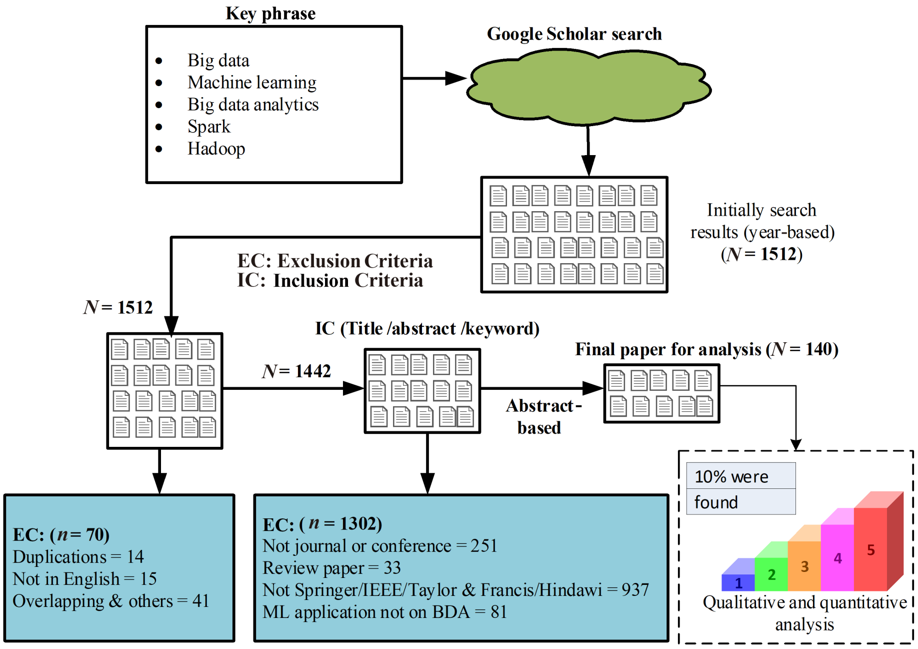
Key phrase
Big data
Machine learning
Big data analytics
Spark
Hadoop
Google Scholar search
Initially search results (year-based) (N = 1512)
EC: Exclusion Criteria
IC: Inclusion Criteria
N = 1512
IC (Title /abstract /keyword)
N = 1442
Final paper for analysis (N = 140)
Abstract-based
EC: (n = 70)
Duplications = 14
Not in English = 15
Overlapping & others = 41
EC: (n = 1302)
Not journal or conference = 251
Review paper = 33
Not Springer/IEEE/Taylor & Francis/Hindawi = 937
ML application not on BDA = 81
10% were found
1
2
3
4
5
Qualitative and quantitative analysis

Fig. 4 Systematic literature review process.

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 written in the English language, (ii) the article must relate to 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 excluded from the qualitative and quantitative analysis of this 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, and papers that were not connected with the proposed study were 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.

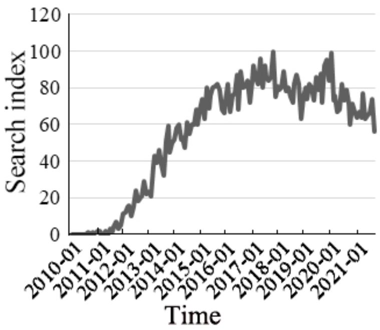

Fig. 5 Trend on Google search queries on big data analytics.

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 association between 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 that were analysed quantitatively. The review shows that current research on BDA can be categorised under five different themes, namely, (i) core BD area to handle the scale, (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.[5] (see Fig. 2), this study grouped the type of big data analytics into three, namely, (i) BD descriptive analytics (denoted as "A"), (ii) BD predictive analytics (denoted as "B"), and (iii) BD prescriptive analytics (denoted as "C"). Out of the 66 papers, 56 indicated the application domain.

It was observed that 58% of the reviewed papers[15, 69–99] were based on BD predictive analytics, 18% BD prescriptive analytics[103–105], 11% BD descriptive analytics[15, 69–99], whiles 9% (A+B)[100–105] and 5% (B+C)[106–108]. Few studies in BDA used prescriptive analytics (see Table A1 in Appendix); this can be attributed to fact that big data prescriptive analytics is in its early stage. However, Bousdekis et al.[109] argued that the development of information fusion algorithms for merging human-driven and sensor-driven learning is the way for big data prescriptive analytics. Cleland et al.[110] argued that BD's rapid evolution is changing several economic sectors, including the health and medical sectors. Accordingly, the healthcare industry is one area where massive historical data are generated daily for various reasons, like conformity and regulatory requirements, keeping records, and patient care[111].

Hence, it is not a shock to see such a massive study in the healthcare industry (30%), followed by anomaly detection (11%), cybersecurity, data privacy & IoT (5%) and automobile and transportation (5%); (see Table A1 in Appendix). Concerning the data size, some studies indicated their data size in terms of the storage space ranging from 708 MB[72] to 600 GB[103], while in terms of the number of observations, it ranges from 1789[93] to 3 billion records[112]. Based on the data size (e.g., 708 MB[72]), one can say that the study by Jallad et al.[72] 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 employed for the empirical analysis also count. Detailed study results based on the proposed taxonomy are presented in the 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 in Ref. [11]. 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 can be effectively derived from this availability of data to make an informed decision. Figure 6b shows the

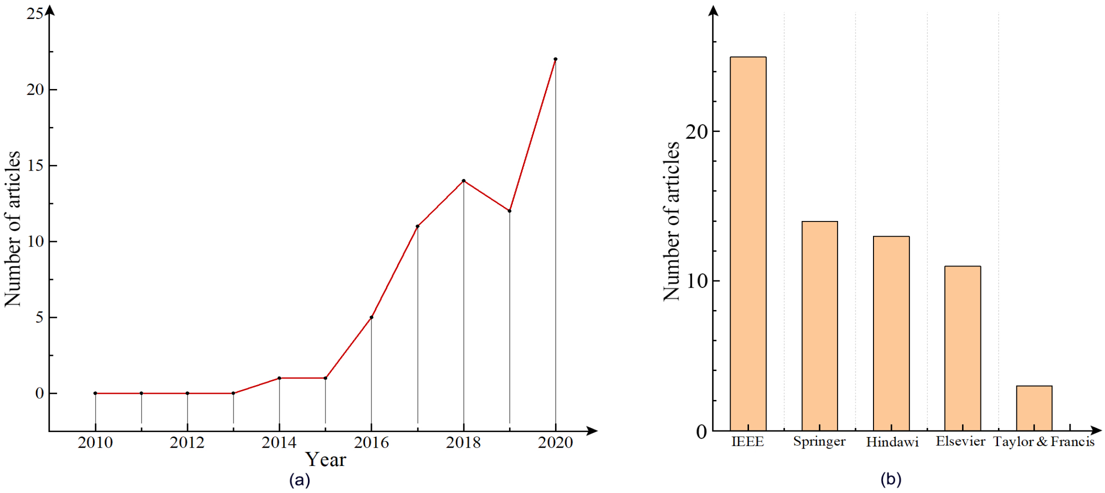

Fig. 6 Time trend publication and publisher wise distribution.

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 that high impact publication houses have seen the need to make available big data analytics with ML applications to the scientific community.

4.2 Big data platforms tool in BDA

Table 2 shows the typically used DBA platforms and tools. 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[8, 113], 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 MapReduce. Also, it is believed that Spark offers high processing speed than Hadoop[103].

Furthermore, Hadoop does not offer data-pipelining, and it is not easy to use. For example, Gu and Li[113] pointed out that it is not suitable for reiterative operations due to the associated cost of reloading disk data at each reiteration. Likewise, Refs. [114, 115] reported that Spark is highly efficient in handling massive amounts of data than Hadoop. Hence, these factors might contribute to its (Spark) familiarity among research work in big data.

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 in caches for some time[103, 114, 115]. 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 Apache frameworks' open-source has seen massive market expansion, as more firms and researchers have found the saccharine spot to adopt these platforms.

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

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-day business, maximising profit, while minimising risk. The SVM 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. According to Celli et al.[93], DTs are smooth to understand. Moreover, they validate the model with statistical tests (like entropy or information gain), contributing to its popularity in BDA.

Another thrilling revelation is that most studies[70, 77, 79, 84, 94, 97, 106, 108, 116–121] used Deep Neural Networks (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 internal structure 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 said that BDA and DL seem to be dependent on one another and show a reciprocally beneficial association. Large amounts of data allow DL techniques to realise better

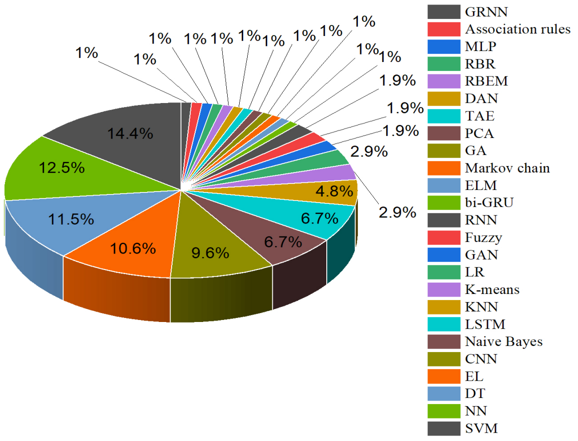

Fig. 7 ML techniques mainly used for BDA.

generalisation; thus, yielding meaningfully and more valuable results in the field of BDA. Hence, in support of Ref. [117], this paper suggests that ML techniques, such as DNN, are worth exploring in future works to gain the acceptance of the platforms. Again, GAN proposed by Ian Goodfellow in 2014 is seen as a robust deep learning algorithm[121], yet it has received little attention in BDA. Therefore, future studies can virtually explore this technique to examine its ability in BDA.

Additionally, it was observed that some studies[15, 69, 73, 87, 89, 93, 95, 98, 101, 122] adopted ensemble learning methods, like the random forest, boosting, and bagging, to enhance the power of EL techniques in BDA. ML algorithm's hybridisation is an excellent technique to compensate for the weakness in the individual algorithm[9]. However, it was revealed that few studies out of 66 papers reviewed adopted it[70, 83, 85, 88, 101, 102, 107, 108, 118, 122]. From Fig. 7, it is 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 studies were undertaken in China (36.4%), followed by the USA (18.2%). Though Srivastava[123] reports that the USA leads China interns of DB adaptation in 2019, this study shows 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 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.

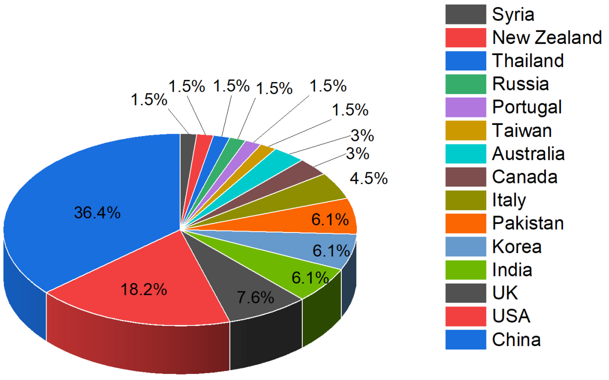

Fig. 8 Publication origin distribution.

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[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 metric because most papers were BD predictive analytics. Therefore, it is argued that combining accuracy and error metrics offers a better ML model evaluation[10].

We observed that several studies[70, 83, 85, 87, 88, 92, 95, 97, 104, 108, 117, 119, 122, 124, 125] combined two or more metrics to evaluate their models and framework.

4.5 Major keywords (use keywords of papers)

Figure 10 shows a plot of the keywords used in the BDA research. All keywords used in the 66 papers were copied in a text file; using a python API (word clouds), we plotted an image of commonly used keywords for

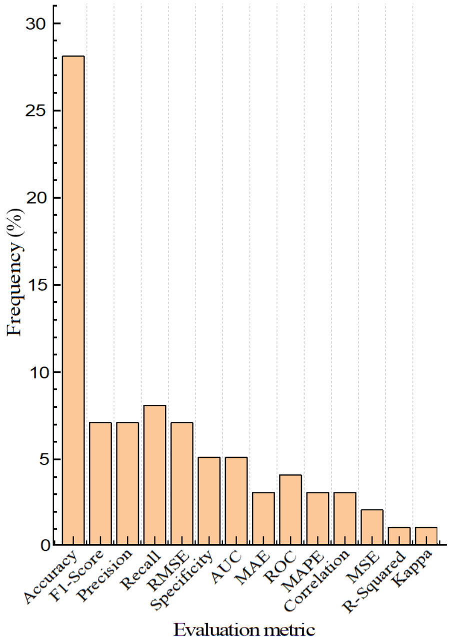

Fig. 9 Evaluation metric distribution.

§ https://www.worldometers.info.

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; it is obvious to envision the most used tools for this review, 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 studies to comprehend the complete context of big data analytics with machine learning and its applications in several industries. The following section puts forward a summary of challenges and openings.

5.1 Key challenges in BDA

Big data analytics challenges have been echoed in past studies[1, 7, 12, 13, 19, 52–61] of similar objectives as this paper; however, the following are few identified to add up or affirm literature.

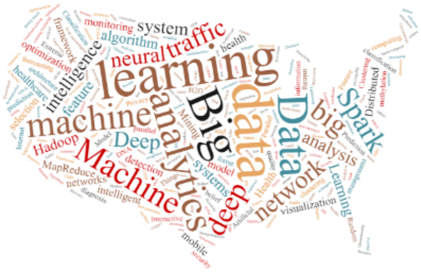

Fig. 10 Keywords commonly used in BDA research.

(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.[42]

(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 handling big data analytics, make it initially expensive for small-scale and medium-scale businesses and organisations in small-income and medium-income countries.

(3) Currently, big data analytics has become a hot topic for discussion among scientists and industry professionals worldwide[114, 115] because of the emergence of new sophisticated technologies, tools, and platforms. However, several studies[1, 7, 12, 13, 19, 52, 52–61] have pointed out the challenge associated with handling this massive and unprecedented amount of data made up of 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 effectively perform well with high accuracy. Nevertheless, this paper can affirm that this challenge still exist in BDA today.

5.2 Prospects for future studies

Several openings that need to be tackled by future research 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 campaign put up a model of classifying individual voters that led their campaign strategy to an efficacious result[126]. 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 cultural perspectives.

(2) 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[127]. 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.

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

(4) 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[120]. Likewise, Khan et al.[128] pointed out that Hadoop has more than 190 configuration constraints, which can substantially affect the performance and time constraints 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.

(5) Even though BDA has seen rapid growth in recent years (see Fig. 6), the study by Shahbaz et al.[129] 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 use BDA. 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.

(6) 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.

Appendix

Table A1 is in the following.

Table A1 A summary of reviewed papers.

No.	Reference	Application area	Objective	Data size
1.	[15]	Disaster (rain damage)	Proposed a forecasting model for heavy rain damage using ML and BD	528 521 data points
2.	[69]	Energy	Predicted offshore wind farm based on BD and ML algorithms	396 000 observations
3.	[70]	Healthcare	Predicted disease in healthcare using ML and BD	20 320 848 records
4.	[71]	E-commerce	Predicted consumer product demands using BD	35 203 observations
5.	[72]	NS	Optimised anomaly detection using BD	708 MB
6.	[73]	Cybersecurity	Detected and classified malicious command and reply packets in an SCADA network	64 100 instances
7.	[74]	Automobile, power system, and road weather dataset	Proposed an improvement in distance variable of time-series DB stream evaluation	34 435 268 instances
8.	[75]	NS	Proposed a Random Bit's Regression (RBR) as a predictor for DB	NA
9.	[76]	NS	Sentiment classification using Rule-Based Emission Model (RBEM)	12 809 observations
10.	[77]	Geochemical anomalies	Detected geochemical anomalies	NA
11.	[78]	Data privacy (wireless network)	Proposed and implemented an ML tactic for smart-edges based on differential privacy	222 048 observations
12.	[79]	Healthcare	Proposed a predictive framework for emergency biomedical BD using DL H_2O	35 233 samples
13.	[80]	Communication systems & network	Proposed a BD and ML-enabled wireless channel model	NA
14.	[81]	Transportation (traffic speed)	Predicted traffic speeds	70 000 samples
15.	[82]	Transportation	Identified unlicensed taxi using BDA	340 679 449 records

(To be continued)

(Continued)

No.	Reference	Application area	Objective	Data size
16.	[83]	Disaster (flood)	Predicted flood using ML and DB	NA
17.	[84]	Communication systems & network	Predicted cellular traffic	300 million records
18.	[85]	Healthcare	Scaled up the ML algorithms that are breast cancer prediction	32 154 records
19.	[86]	Economics, network security, NLP, and more	Proposed a distribution preserving kernel SVM	2 405 562 observations
20.	[87]	Healthcare, divorce, and musk	Cooperative co-evolution for future selection in big data	17 463 observations
21.	[88]	Healthcare	Predicted unpleasant occurrences after surgery	Over 27 million records
22.	[89]	NS	Distributed classifier training based on label-aware distributed EL	10 543 802 records
23.	[90]	Electricity generation	Forecasted electricity generation ML and BD	NA
24.	[91]	Transportation	Predicted train delay using BDA	NA
25.	[92]	Healthcare	Predicted at-risk profiles upon admission at the hospital	1 271 733 records
26.	[93]	Healthcare	Identified cancer drivers based on large DNA methylation datasets classification	1789 samples
27.	[94]	Face detection	Proposed a face detection model based on a fast deep-convolutional network	600 000 samples
28.	[95]	Finance	Proposed a BDA for complex credit risk assessment	2284 records
29.	[96]	Entertainment industry (movie)	Proposed a technique for movie review cataloguing and summarisation	53 000 records
30.	[97]	E-commerce	Proposed a four-dimension (4D), word-based representation model, using question-answering interaction-level and hyper interaction-level	NA
31.	[98]	Social media (Web forums)	Proposed answer discovery in discussion forums using ML and BDA	NA
32.	[99]	Construction and architectural	Enhanced the performance of embedded smart-city architecture based on BDA	3.1 GB
33.	[100]	Sports	Forecasted ice hockey outcomes using BD and ensemble methods in ML	43 million observations
34.	[101]	Healthcare	Assessed the performance of 5 ML methods using outpatients' visits data	2 011 813 observations
35.	[102]	Healthcare and aviation	Analysed diabetes prediction and flight delays	414 116 observations
36.	[103]	NS	All-inclusive experimental performance evaluation between Spark and MapReduce frameworks	600 GB
37.	[104]	Healthcare	Quantified psychiatric diagnosis using ML algorithms	NA
38.	[105]	NS	Proposed a clustering algorithm using fuzzy c-Means for handling BD	92.3 GB
39.	[106]	Manufacturing	Proposed the design and implementation of a DB-driven computer numerical control machining process	NA
40.	[107]	Anomalies detection	Proposed an ML and BD based mobile IoT security monitoring	7 009 270 records
41.	[108]	Intrusion detection	Proposed hierarchical DL system for intrusion detection	7.741 million samples
42.	[109]	IoT	Prescriptive analytics time-dependent parameters-based sensor-driven learning	NA
43.	[110]	Healthcare	Understanding of antidepressant prescribing	NA
44.	[112]	Healthcare	BDA using distributed data over HBase	3 billion records
45.	[116]	Image pre-processing	Presented a deep computing paradigm for big data feature learning	101 300 observations
46.	[117]	Transportation	Proposed an online ML-based traffic monitoring	190 million records
47.	[118]	Healthcare	Proposed hybrid multi-modal DL based on collaborative concat layer	610 488 records
48.	[119]	Fault recognition	Proposed a portable robotic roller bearing fault diagnosis	2 800 000 sample

(To be continued)

(Continued)

No.	Reference	Application area	Objective	Data size
49.	[120]	NS	Proposed automatic tuning configurations of BD frameworks using GAN	NA
50.	[121]	NS	GPGPU NN for DL big data analysis	NA
51.	[122]	NS	Feature selection in BD with random feature grouping	NA
52.	[124]	Healthcare	Perceived of cardiac Arrhythmias	109 000 records
53.	[125]	Entertainment industry (movie)	Proposed a movie recommendation system based on an improved hybrid sentiment analyser for BDA	NA
54.	[127]	Communication systems & network	Proposed a big data analytic framework for mobile cellular networks	NA
55.	[128]	NS	Proposed an improved parallel detrended fluctuation analysis algorithm for scalable analytics on big data	NA
56.	[130]	Aviation	Predicted flight delay based on aviation BD and ML	NA
57.	[131]	Healthcare	Proposed a BD and IoT based patient behaviour monitoring system	NA
58.	[132]	NS	Modeled firm's performance and quality dynamics using BD	NA
59.	[133]	NS	Assessed associative classification techniques for BD	200 GB
60.	[134]	Healthcare	Explored the associations amid medical college standings and performance with DB	602 770 instances
61.	[135]	Sports	Analysed the influence of BD on sports	NA
62.	[136]	NS	Proposed a distributed and weighted extreme learning for imbalanced BD	NA
63.	[137]	Healthcare	Proposed a knowledge-based technique for BDA in immunology	NA
64.	[138]	Finance	Proposed a dynamic feedback warning using BDA	NA
65.	[139]	NS	Developed a multiple BDA platform with rapid response	NA
66.	[140]	Cybersecurity	A BDA framework for detecting targeted cyber-attacks	NA

Note: NA = Not Applicable; NS = Not Specified (authors did not make known the industry where their data came from).

References

[1] J. F. Qiu, Q. H. Wu, G. R. Ding, Y. H. Xu, and S. Feng, A survey of machine learning for big data processing, *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 67, 2016.

[2] B. Aragona and R. De Rosa, Big data in policy making, *Math. Popul. Stud.*, vol. 26, no. 2, pp. 107–113, 2019.

[3] G. Kaur, P. Tomar, and P. Singh, Design of cloud-based green IoT architecture for smart cities, in *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence*, N. Dey, A. E. Hassanien, C. Bhatt, A. S. Ashour, and S. C. Satapathy, eds. Cham, Germany: Springer, 2018, pp. 315–333.

[4] A. Holst, Amount of information globally 2010–2024, https://www.statista.com/statistics/871513/worldwide-data-created/, 2020.

[5] 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.

[6] S. Debortoli, O. Müller, and J. vom Brocke, Comparing business intelligence and big data skills, *Bus. Inf. Syst. Eng.*, vol. 6, no. 5, pp. 289–300, 2014.

[7] B. K. Sarkar, Big data for secure healthcare system: A conceptual design, *Complex Intell. Syst.*, vol. 3, no. 2, pp. 133–151, 2017.

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

[9] I. K. Nti, A. F. Adekoya, and B. A. Weyori, A systematic review of fundamental and technical analysis of stock market predictions, *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 3007–3057, 2020.

[10] I. K. Nti, A. F. Adekoya, and B. A. Weyori, A comprehensive evaluation of ensemble learning for stock-market prediction, *J. Big Data*, vol. 7, no. 1, p. 20, 2020.

[11] R. Raja, I. Mukherjee, and B. K. Sarkar, A systematic review of healthcare big data, *Sci. Program.*, vol. 2020, p. 5471849, 2020.

[12] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, Big data analytics: A survey, *J. Big Data*, vol. 2, no. 1, p. 21, 2015.

[13] P. Russom, Introduction to big data analytics, https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf, 2011.

[14] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks, *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[15] 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.

[16] T. T. Le, W. X. Fu, and J. H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, *Bioinformatics*, vol. 36, no. 1, pp.

250–256, 2020.

[17] K. Y. Ngiam and I. W. Khor, Big data and machine learning algorithms for health-care delivery, *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, 2019.

[18] F. Wang, M. G. Li, Y. D. Mei, and W. R. Li, Time series data mining: A case study with big data analytics approach, *IEEE Access*, vol. 8, pp. 14322–14328, 2020.

[19] W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: Promise and potential, *Health Inf. Sci. Syst.*, vol. 2, p. 3, 2014.

[20] I. K. Nti, A. F. Adekoya, and B. A. Weyori, Random forest based feature selection of macroeconomic variables for stock market prediction, *Am. J. Appl. Sci.*, vol. 16, no. 7, pp. 200–212, 2019.

[21] A. F. Adekoya and K. I. Nti, The COVID-19 outbreak and effects on major stock market indices across the globe: A machine learning approach, *Indian J. Sci. Technol.*, vol. 13, no. 35, pp. 3695–3706, 2020.

[22] I. K. Nti, A. F. Adekoya, and B. A. Weyori, Predicting stock market price movement using sentiment analysis: Evidence from Ghana, *Appl. Comput. Syst.*, vol. 25, no. 1, pp. 33–42, 2020.

[23] I. K. Nti, A. F. Adekoya, and B. A. Weyori, Efficient stock-market prediction using ensemble support vector machine, *Open Comput. Sci.*, vol. 10, no. 1, pp. 153–163, 2020.

[24] I. K. Nti, M. Teimeh, A. F. Adekoya, and O. Nyarko-boateng, Forecasting electricity consumption of residential users based on lifestyle data using artificial neural networks, *ICTACT J. Soft Comput.*, vol. 10, no. 3, pp. 2107–2116, 2020.

[25] I. K. Nti, A. Y. Appiah, and O. Nyarko-Boateng, Assessment and prediction of earthing resistance in domestic installation, *Eng. Rep.*, vol. 2, no. 1, p. e12090, 2020.

[26] I. K. Nti, A. A. Samuel, and A. Michael, Predicting monthly electricity demand using soft-computing technique, *Int. Res. J. Eng. Technol.*, vol. 6, no. 6, pp. 1967–1973, 2019.

[27] I. K. Nti, A. F. Adakoya, and O. Nyarko-Boateng, A multifactor authentication framework for the national health insurance scheme in ghana using machine learning, *Am. J. Eng. Appl. Sci.*, vol. 13, no. 4, pp. 639–648, 2020.

[28] S. Akyeramfo-Sam, A. A. Philip, D. Yeboah, N. C. Nartey, and I. K. Nti, A web-based skin disease diagnosis using convolutional neural networks, *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 11, pp. 54–60, 2019.

[29] D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, Partial derivative nonlinear global pandemic machine learning prediction of COVID-19, *Chaos, Solitons Fractals*, vol. 139, p. 110056, 2020.

[30] I. K. Nti, A. F. Adekoya, M. Opoku, and P. Nimbe, Synchronising social media into teaching and learning settings at tertiary education, *Int. J. Soc. Media Interact. Learn. Environ.*, vol. 6, no. 3, pp. 230–243, 2020.

[31] I. K. Nti and J. A. Quarcoo, Self-motivation and academic performance in computer programming language using a hybridised machine learning technique, *Int. J. Artif. Intell. Expert Syst.*, vol. 8, no. 2, pp. 12–30, 2019.

[32] R. Ghorbani and R. Ghousi, Comparing different resampling methods in predicting students' performance using machine learning techniques, *IEEE Access*, vol. 8, pp. 67899–67911, 2020.

[33] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, Predicting at-risk university students in a virtual learning environment via a machine learning algorithm, *Comput. Human Behav.*, vol. 107, p. 105584, 2020.

[34] I. K. Nti, G. Eric, and Y. S. Jonas, Detection of plant leaf disease employing image processing and gaussian smoothing approach, *Int. J. Comput. Appl.*, vol. 162, no. 2, pp. 20–25, 2017.

[35] A. Sharifi, Yield prediction with machine learning algorithms and satellite images, *J. Sci. Food Agric.*, vol. 101, no. 3, pp. 891–896, 2021.

[36] A. Hamrani, A. Akbarzadeh, and C. A. Madramootoo, Machine learning for predicting greenhouse gas emissions from agricultural soils, *Sci. Total Environ.*, vol. 741, p. 140338, 2020.

[37] A. Boukerche and J. H. Wang, Machine learning-based traffic prediction models for intelligent transportation systems, *Comput. Netw.*, vol. 181, p. 107530, 2020.

[38] P. P. Hanzelik, S. Gergely, C. Gáspár, and L. Györy, Machine learning methods to predict solubilities of rock samples, *J. Chemom.*, vol. 34, no. 2, p. e3198, 2020.

[39] A. L. Beam and I. S. Kohane, Big data and machine learning in health care, *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.

[40] J. E. Bibault, P. Giraud, and A. Burgun, Big Data and machine learning in radiation oncology: State of the art and future prospects, *Cancer Lett.*, vol. 382, no. 1, pp. 110–117, 2016.

[41] S. Siuly and Y. C. Zhang, Medical big data: Neurological diseases diagnosis through medical data analysis, *Data Sci. Eng.*, vol. 1, no. 2, pp. 54–64, 2016.

[42] A. K. U. Haq, A. Khattak, N. Jamil, M. A. Naeem, and F. Mirza, Data analytics in mental healthcare, *Sci. Program.*, vol. 2020, p. 2024160, 2020.

[43] G. R. Chen and M. Islam, Big data analytics in healthcare, in *Proc.* 2^{nd} *Int. Conf. Safety Produce Informatization*, Chongqing, China, 2019, pp. 227–230.

[44] Z. F. Khan and S. R. Alotaibi, Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective, *J. Healthc. Eng.*, vol. 2020, p. 8894694, 2020.

[45] Z. He, C. Tao, J. Bian, M. Dumontier, and W. R. Hogan, Semantics-powered healthcare engineering and data analytics, *J. Healthc. Eng.*, vol. 2017, p. 7983473, 2017.

[46] G. K. Kang, J. Z. Gao, S. Chiao, S. Q. Lu, and G. Xie, Air quality prediction: Big data and machine learning approaches, *Int. J. Environ. Sci. Dev.*, vol. 9, no. 1, pp. 8–16, 2018.

[47] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, Deep learning for IoT big data and streaming analytics: A survey, *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 2923–2960, 2018.

[48] Z. M. Bi and D. Cochran, Big data analytics with applications, *J. Manag. Anal.*, vol. 1, no. 4, pp. 249–265, 2014.

[49] S. Choudhury, Q. Ye, M. X. Dong, and Q. C. Zhang, IoT big data analytics, *Wirel. Commun. Mob. Comput.*, vol. 2019, p. 9245392, 2019.

[50] C. Ma, H. H. Zhang, and X. F. Wang, Machine learning for Big Data analytics in plants, *Trends Plant Sci.*, vol. 19, no. 12, pp. 798–808, 2014.

[51] K. Szczypiorski, L. Q. Wang, X. Y. Luo, and D. P. Ye, Big data analytics for information security, *Secur. Commun. Netw.*, vol. 2018, p. 7657891, 2018.

[52] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, Machine learning with big data: Challenges and approaches, *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[53] L. N. Zhou, S. M. Pan, J. W. Wang, and A. V. Vasilakos, Machine learning on big data: Opportunities and challenges, *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[54] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, Big data, analytics and the path from insights to value, *MIT Sloan Manag. Rev.*, vol. 52, no. 2, pp. 21–31, 2011.

[55] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, Interactions with big data analytics, *Interactions*, vol. 19, no. 3, pp. 50–59, 2012.

[56] H. Chen, R. H. L. Chiang, and V. C. Storey, Business intelligence and analytics: From big data to big impact, *MIS Q.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[57] J. Ram, C. Y. Zhang, and A. Koronios, The implications of big data analytics on business intelligence: A qualitative study in China, *Procedia Comput. Sci.*, vol. 87, pp. 221–226, 2016.

[58] T. Condie, P. Mineiro, N. Polyzotis, and M. Weimer, Machine learning on Big Data, in *Proc. 29th Int. Conf. Data Engineering*, Brisbane, Australia, 2013, pp. 1242–1244.

[59] B. Wixom, T. Ariyachandra, D. Douglas, M. Goul, B. Gupta, L. Iyer, U. Kulkarni, J. G. Mooney, G. Phillips-Wren, and O. Turetken, The current state of business intelligence in academia: The arrival of big data, *Commun. Assoc. Inf. Syst.*, vol. 34, p. 1, 2014.

[60] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, Trends in big data analytics, *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2561–2573, 2014.

[61] Z. Obermeyer and E. J. Emanuel, Predicting the future — big data, machine learning, and clinical medicine, *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, 2016.

[62] B. Ale, Risk analysis and big data, *Saf. Reliab.*, vol. 36, no. 3, pp. 153–165, 2016.

[63] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, Efficient machine learning for big data: A review, *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.

[64] D. Z. Chong and H. Shi, Big data analytics: A literature review, *J. Manag. Anal.*, vol. 2, no. 3, pp. 175–201, 2015.

[65] M. Fathi, M. H. Kashani, S. M. Jameii, and E. Mahdipour, Big data analytics in weather forecasting: A systematic review, *Arch. Comput. Methods Eng.*, doi: 10.1007/s11831-021-09616-4.

[66] H. Hassani and E. S. Silva, Forecasting with big data: A review, *Ann. Data Sci.*, vol. 2, no. 1, pp. 5–19, 2015.

[67] L. Collins, Mini literature review: A new type of literature review article, https://www.emeraldgrouppublishing.com/archived/products/journals/call_for_papers.htm%3Fid%3D5730, 2021.

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

[69] X. X. Yin and X. W. Zhao, Big data driven multi-objective predictions for offshore wind farm based on machine learning algorithms, *Energy*, vol. 186, p. 115704, 2019.

[70] M. Chen, Y. X. Hao, K. Hwang, L. Wang, and L. Wang, Disease prediction by machine learning over big data from healthcare communities, *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[71] A. Y. L. Chong, E. Ch'ng, M. J. Liu, and B. Y. Li, Predicting consumer product demands via Big Data: The roles of online promotional marketing and online reviews, *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5142–5156, 2017.

[72] K. A. Jallad, M. Aljnidi, and M. S. Desouki, Anomaly detection optimization using big data and deep learning to reduce false-positive, *J. Big Data*, vol. 7, no. 1, p. 68, 2020.

[73] K. M. Paramkusem and R. S. Aygun, Classifying categories of SCADA attacks in a big data framework, *Ann. Data Sci.*, vol. 5, no. 3, pp. 359–386, 2018.

[74] A. Wibisono, P. Mursanto, J. Adibah, W. D. W. T. Bayu, M. I. Rizki, L. M. Hasani, and V. F. Ahli, Distance variable improvement of time-series big data stream evaluation, *J. Big Data*, vol. 7, no. 1, p. 85, 2020.

[75] Y. Wang, Y. Li, M. M. Xiong, Y. Y. Shugart, and L. Jin, Random bits regression: A strong general predictor for big data, *Big Data Anal.*, vol. 1, p. 12, 2016.

[76] E. Tromp, M. Pechenizkiy, and M. M. Gaber, Expressive modeling for trusted big data analytics: Techniques and applications in sentiment analysis, *Big Data Anal.*, vol. 2, no. 1, p. 5, 2017.

[77] R. G. Zuo and Y. H. Xiong, Big data analytics of identifying geochemical anomalies supported by machine learning methods, *Nat. Resour. Res.*, vol. 27, no. 1, pp. 5–13, 2018.

[78] M. Du, K. Wang, Z. Q. Xia, and Y. Zhang, Differential privacy preserving of training model in wireless big data with edge computing, *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 283–295, 2020.

[79] A. S. Elsayad, A. I. El Desouky, M. M. Salem, and M. Badawy, A deep learning H_2O framework for emergency prediction in biomedical big data, *IEEE Access*, vol. 8, pp. 97231–97242, 2020.

[80] J. Huang, C. X. Wang, L. Bai, J. Sun, Y. Yang, J. Li, O. Tirkkonen, and M. T. Zhou, A big data enabled channel model for 5G wireless communication systems, *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 211–222, 2020.

[81] D. Jo, B. Yu, H. Jeon, and K. Sohn, Image-to-image

learning to predict traffic speeds by considering area-wide spatio-temporal dependencies, *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1188–1197, 2019.

[82] W. Yuan, P. Deng, T. Taleb, J. F. Wan, and C. F. Bi, An unlicensed taxi identification model based on big data analysis, *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1703–1713, 2016.

[83] S. Puttinaovarat and P. Horkaew, Flood forecasting system based on integrated big and crowdsource data by using machine learning techniques, *IEEE Access*, vol. 8, pp. 5885–5905, 2020.

[84] C. T. Zhang, H. X. Zhang, J. P. Qiao, D. F. Yuan, and M. G. Zhang, Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data, *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, 2019.

[85] S. Alghunaim and H. H. Al-Baity, On the scalability of machine-learning algorithms for breast cancer prediction in big data context, *IEEE Access*, vol. 7, pp. 91535–91546, 2019.

[86] D. Singh, D. Roy, and C. K. Mohan, DiP-SVM: Distribution preserving kernel support vector machine for big data, *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 79–90, 2017.

[87] A. N. M. B. Rashid, M. Ahmed, L. F. Sikos, and P. Haskell-Dowland, A novel penalty-based wrapper objective function for feature selection in big data using cooperative co-evolution, *IEEE Access*, vol. 8, pp. 150113–150129, 2020.

[88] S. B. Roy, M. Maria, T. N. Wang, A. Ehlers, and D. Flum, Predicting adverse events after surgery, *Big Data Res.*, vol. 13, pp. 29–37, 2018.

[89] S. Khalifa, P. Martin, and R. Young, Label-aware distributed ensemble learning: A simplified distributed classifier training model for big data, *Big Data Res.*, vol. 15, pp. 1–11, 2019.

[90] M. N. Rahman, A. Esmailpour, and J. H. Zhao, Machine learning with big data an efficient electricity generation forecasting system, *Big Data Res.*, vol. 5, pp. 9–15, 2016.

[91] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and D. Anguita, Train delay prediction systems: A big data analytics perspective, *Big Data Res.*, vol. 11, pp. 54–64, 2018.

[92] P. Genevès, T. Calmant, N. Layaïda, M. Lepelley, S. Artemova, and J. L. Bosson, Scalable machine learning for predicting at-risk profiles upon hospital admission, *Big Data Res.*, vol. 12, pp. 23–34, 2018.

[93] F. Celli, F. Cumbo, and E. Weitschek, Classification of large DNA methylation datasets for identifying cancer drivers, *Big Data Res.*, vol. 13, pp. 21–28, 2018.

[94] D. Triantafyllidou, P. Nousi, and A. Tefas, Fast deep convolutional face detection in the wild exploiting hard sample mining, *Big Data Res.*, vol. 11, pp. 65–76, 2018.

[95] A. W. Niu, B. Q. Cai, and S. S. Cai, Big data analytics for complex credit risk assessment of network lending based on SMOTE algorithm, *Complexity*, vol. 2020, p. 8563030, 2020.

[96] A. Khan, M. A. Gul, M. I. Uddin, S. A. A. Shah, S. Ahmad, M. D. Al Firdausi, and M. Zaindin, Summarizing online movie reviews: A machine learning approach to big data analytics, *Sci. Program.*, vol. 2020, p. 5812715, 2020.

[97] H. Q. Wu, M. M. Liu, S. B. Zhang, Z. K. Wang, and S. L. Cheng, Big data management and analytics in scientific programming: A deep learning-based method for aspect category classification of question-answering-style reviews, *Sci. Program.*, vol. 2020, p. 4690974, 2020.

[98] A. Khan, I. Ibrahim, M. I. Uddin, M. Zubair, S. Ahmad, M. D. Al Firdausi, and M. Zaindin, Machine learning approach for answer detection in discussion forums: An application of big data analytics, *Sci. Program.*, vol. 2020, p. 4621196, 2020.

[99] B. N. Silva, M. Khan, and K. Han, Big data analytics embedded smart city architecture for performance enhancement through real-time data processing and decision-making, *Wirel. Commun. Mob. Comput.*, vol. 2017, p. 9429676, 2017.

[100] 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.

[101] T. Daghistani, H. AlGhamdi, R. Alshammari, and R. H. AlHazme, Predictors of outpatients' no-show: Big data analytics using apache spark, *J. Big Data*, vol. 7, p. 108, 2020.

[102] T. Nibareke and J. Laassiri, Using Big Data-machine learning models for diabetes prediction and flight delays analytics, *J. Big Data*, vol. 7, p. 78, 2020.

[103] N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A. Rashid, A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench, *J. Big Data*, vol. 7, no. 1, p. 110, 2020.

[104] F. Saeed, Towards quantifying psychiatric diagnosis using machine learning algorithms and big fMRI data, *Big Data Anal.*, vol. 3, no. 1, p. 7, 2018.

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

[106] K. P. Zhu, G. C. Li, and Y. Zhang, Big data oriented smart tool condition monitoring system, *IEEE Trans. Ind. Inform.*, vol. 16, no. 6, pp. 4007–4016, 2020.

[107] I. Kotenko, I. Saenko, and A. Branitskiy, Framework for mobile internet of things security monitoring based on big data processing and machine learning, *IEEE Access*, vol. 6, pp. 72714–72723, 2018.

[108] W. Zhong, N. Yu, and C. Y. Ai, Applying big data based deep learning system to intrusion detection, *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 181–195, 2020.

[109] A. Bousdekis, N. Papageorgiou, B. Magoutas, D. Apostolou, and G. Mentzas, Sensor-driven learning of time-dependent parameters for prescriptive analytics, *IEEE Access*, vol. 8, pp. 92383–92392, 2020.

[110] B. Cleland, J. Wallace, R. Bond, M. Black, M. Mulvenna, D. Rankin, and A. Tanney, Insights into antidepressant prescribing using open health data, *Big Data Res.*, vol. 12, pp. 41–48, 2018.

[111] M. Giacalone, C. Cusatelli, and V. Santarcangelo, Big data

compliance for innovative clinical models, *Big Data Res.*, vol. 12, pp. 35–40, 2018.

[112] D. Chrimes and H. Zamani, Using distributed data over HBase in big data analytics platform for clinical services, *Comput. Math. Methods Med.*, vol. 2017, p. 6120820, 2017.

[113] 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. Embedded and Ubiquitous Computing*, Zhangjiajie, China, 2013, pp. 721–727.

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

[115] 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.

[116] Q. C. Zhang, L. T. Yang, and Z. K. Chen, Deep computation model for unsupervised feature learning on big data, *IEEE Trans. Serv. Comput.*, vol. 9, no. 1, pp. 161–171, 2016.

[117] 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.

[118] J. C. Kim and K. Chung, Hybrid multi-modal deep learning using collaborative concat layer in health bigdata, *IEEE Access*, vol. 8, pp. 192469–192480, 2020.

[119] G. M. Xian, Parallel machine learning algorithm using fine-grained-mode spark on a mesos big data cloud computing software framework for mobile robotic intelligent fault recognition, *IEEE Access*, vol. 8, pp. 131885–131900, 2020.

[120] M. Y. Li, Z. Q. Liu, X. H. Shi, and H. Jin, ATCS: Auto-tuning configurations of big data frameworks based on generative adversarial nets, *IEEE Access*, vol. 8, pp. 50485–50496, 2020.

[121] A. Fonseca and B. Cabral, Prototyping a GPGPU neural network for deep-learning big data analysis, *Big Data Res.*, vol. 8, pp. 50–56, 2017.

[122] A. N. M. B. Rashid, M. Ahmed, L. F. Sikos, and P. Haskell-Dowland, Cooperative co-evolution for feature selection in Big Data with random feature grouping, *J. Big Data*, vol. 7, no. 1, p. 107, 2020.

[123] S. Srivastava, Top 10 countries & regions leading the big data adoption in 2019, https://www.analyticsinsight.net/top-10-countries-regions-leading-the-big-data-adoption-in-2019/, 2020.

[124] H. M. Rai and K. Chatterjee, A novel adaptive feature extraction for detection of cardiac arrhythmias using hybrid technique MRDWT & MPNN classifier from ECG big data, *Big Data Res.*, vol. 12, pp. 13–22, 2018.

[125] Y. B. Wang, M. M. Wang, and W. Xu, A sentiment-enhanced hybrid recommender system for movie recommendation: A big data analytics framework, *Wirel. Commun. Mob. Comput.*, vol. 2018, p. 8263704, 2018.

[126] R. J. Dalton, The potential of big data for the cross-national study of political behavior, *Int. J. Sociol.*, vol. 46, no. 1, pp. 8–20, 2016.

[127] Y. He, F. R. Yu, N. Zhao, H. X. Yin, H. P. Yao, and R. C. Qiu, Big data analytics in mobile cellular networks, *IEEE Access*, vol. 4, pp. 1985–1996, 2016.

[128] M. Khan, Z. W. Huang, M. Z. Li, G. A. Taylor, P. M. Ashton, and M. Khan, Optimizing hadoop performance for big data analytics in smart grid, *Math. Probl. Eng.*, vol. 2017, p. 2198262, 2017.

[129] M. Shahbaz, C. Y. Gao, L. L. Zhai, F. Shahzad, and M. R. Arshad, Moderating effects of gender and resistance to change on the adoption of big data analytics in healthcare, *Complexity*, vol. 2020, p. 2173765, 2020.

[130] G. Gui, F. Liu, J. L. Sun, J. Yang, Z. Q. Zhou, and D. X. Zhao, Flight delay prediction based on aviation big data and machine learning, *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 140–150, 2020.

[131] K. T. Chui, R. W. Liu, M. D. Lytras, and M. B. Zhao, Big data and IoT solution for patient behaviour monitoring, *Behav. Inf. Technol.*, vol. 38, no. 9, pp. 940–949, 2019.

[132] S. J. F. Ren, S. F. Wamba, S. Akter, R. Dubey, and S. J. Childe, Modelling quality dynamics, business value and firm performance in a big data analytics environment, *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5011–5026, 2017.

[133] F. Padillo, J. M. Luna, and S. Ventura, Evaluating associative classification algorithms for Big Data, *Big Data Anal.*, vol. 4, no. 1, p. 2, 2019.

[134] A. R. Rao and D. Clarke, Exploring relationships between medical college rankings and performance with big data, *Big Data Anal.*, vol. 4, no. 1, p. 3, 2019.

[135] D. Patel, D. Shah, and M. Shah, The intertwine of brain and body: A quantitative analysis on how big data influences the system of sports, *Ann. Data Sci.*, vol. 7, pp. 1–16, 2020.

[136] Z. Q. Wang, J. C. Xin, H. X. Yang, S. Tian, G. Yu, C. R. Xu, and Y. D. Yao, Distributed and weighted extreme learning machine for imbalanced big data learning, *Tsinghua Science and Technology*, vol. 22, no. 2, pp. 160–173, 2017.

[137] G. L. Zhang, J. Sun, L. Chitkushev, and V. Brusic, Big data analytics in immunology: A knowledge-based approach, *Biomed Res. Int.*, vol. 2014, p. 437987, 2014.

[138] W. R. Li, M. G. Li, Y. D. Mei, T. Li, and F. Wang, A big data analytics approach for dynamic feedback warning for complex systems, *Complexity*, vol. 2020, p. 7652496, 2020.

[139] B. R. Chang, Y. D. Lee, and P. H. Liao, Development of multiple big data analytics platforms with rapid response, *Sci. Program.*, vol. 2017, p. 6972461, 2017.

[140] A. K. Ju, Y. B. Guo, Z. W. Ye, T. Li, and J. Ma, HeteMSD: A big data analytics framework for targeted cyber-attacks detection using heterogeneous multisource data, *Secur. Commun. Netw.*, vol. 2019, p. 5483918, 2019.

Isaac Kofi Nti received the Higher National Diploma (HND) in electrical & electronic engineering from Sunyani Technical University, Ghana in 2007, the BS degree in computer science from Catholic University College, Ghana in 2011, the MS degree in information technology from Kwame Nkrumah University of Science and Technology, Ghana in 2016, and the PhD degree in computer science from University of Energy and Natural Resources, Ghana in 2021. He is a lecturer at the Department of Computer Science and Informatics, University of Energy and Natural Resources (UENR), Sunyani, Ghana. His research interests include artificial intelligence, energy system modelling, intelligent information systems, social and sustainable computing, business analytics, and data privacy and security.

Juanita Ahia Quarcoo received the MPhil degree in computer engineering from Kwame Nkrumah University of Science and Technology, Ghana in 2011, and PgD (Postgraduate Diploma) in Wireless and Mobile Computing (WiMC) from Advanced Information Technology Institute–Kofi Annan Centre of Excellence (AITI–KACE), Ghana in 2013. She is a lecturer at the Department of Electrical & Electronic, Sunyani Technical University. Her research interests are in the area of computer networks, educational and instructional technology, wireless and mobile computing, artificial intelligence, cloud computing, and mobile app development.

Justice Aning received the MS degree in information technology from Kwame Nkrumah University of Science and Technology (KNUST), Ghana in 2017. He is currently a lecturer at the Department of Computer Science, Sunyani Technical University, Sunyani, Ghana. He has authored more than five papers in journals. His current research interests include Web design, machine learning, and intelligent systems for modelling and optimisation.

Godfred Kusi Fosu received the BS degree in computer science from All Nation University College, Ghana in 2009, and the MS degree in information technology from Sikkim Manipal University, India in 2014. He is a computer scientist with a broad interest in and strong passion for computer science and other research. He is currently an assistant lecturer at the Department of Computer Science, Sunyani Technical University. His research areas include artificial intelligence application, data science, computer and data security, and wireless and mobile computing.