Data Analytics and Techniques: A Review

Safa S. Abdul-Jabbar[1] and Alaa K. Farhan[2]

[1]Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq
[2]Department of Computer Science, University of Technology, Baghdad, Iraq

***Abstract*—Big data of different types, such as texts and images, are rapidly generated from the internet and other applications. Dealing with this data using traditional methods is not practical since it is available in various sizes, types, and processing speed requirements. Therefore, data analytics has become an important tool because only meaningful information is analyzed and extracted, which makes it essential for big data applications to analyze and extract useful information. This paper presents several innovative methods that use data analytics techniques to improve the analysis process and data management. Furthermore, this paper discusses how the revolution of data analytics based on artificial intelligence algorithms might provide improvements for many applications. In addition, critical challenges and research issues were provided based on published paper limitations to help researchers distinguish between various analytics techniques to develop highly consistent, logical, and information-rich analyses based on valuable features. Furthermore, the findings of this paper may be used to identify the best methods in each sector used in these publications, assist future researchers in their studies for more systematic and comprehensive analysis and identify areas for developing a unique or hybrid technique for data analysis.**

***Index Terms*—Big data analysis, Data analytics, Data analysis, Data management, Machine learning**

I. Introduction

Every company collects a considerable amount of data from various sources. So two prerequisites are needed to secure this data and use techniques to extract useful information from this data (Khoshbakht, Shiranzaei and Quadri, 2021; Farhan and Ali, 2017). The use of big data has rapidly progressed from a theory to a reality with the rapid progression of data resources and the creation of companies specializing in big data (Zheng and Guo, 2020; Do Nascimento, et al., 2021; Mariani and Baggio, 2022). For example, clients struggle to find relevant and acceptable material that satisfies their needs because the amount of data on the internet is constantly rising. When a customer submits a query for information or data to an Internet search engine, the result is typically many pages. Hence, he faces the repetitious task of locating the appropriate data from this flood of results. The term describing this problem is called "Data Overloading" (Kan and Klavans, 2002). Hence, the primary objective of this decade of electronic revolution is to construct and ensure a better manner of managing, collaborating, and developing via the use of computer and information technology-based knowledge and information-oriented services (Rajon, Shamim and Arif, 2011; Russell and Norvig, 2020). The process of analyzing and discovering hidden patterns, undiscovered correlations, and other valuable business information from a vast volume of data is known as big data analytics (Patel, Singh and Kazi, 2017; Faizan, et al., 2020). Therefore, data analytics is a crucial subject for many systems, such as those that work with strings or information retrieval operations (Abdul-jabbar and George, 2017). Furthermore, data analytics can be used to check the privacy issues in social media, such as tags and image uploading, as we can see on Flickr and Facebook (Smith, et al., 2013; Abkenar, et al., 2020). Besides social media applications, data analytics can provide many services for applications in different fields, such as audio and video (Verma and Agrawal, 2016).

This paper has three overarching goals:

1. It will provide a brief history of data analytics techniques and methods for documents and describe how data analytics tools utilize the knowledge from all input documents.
2. Presents how the previous studies are based on multi-algorithms and multi objective to optimize the traditional methods and explain the researcher with a comprehensive overview that helps him choose the suitable algorithms and integrate them into a model according to the task at hand.
3. Finally, this paper also illustrates the limitations of each proposed method to present new directions in future works.

The paper is structured as follows, Section II introduces the proposed data analytics techniques and methods, and Section III interprets and describes the significance of our findings in the published paper. Finally, Sections IV and V present a compelling discussion and conclusions that inform researchers on what they can learn from published research papers mentioned in this research.

II. Data Analytics and its Methods

Data analysis primarily entails big data analytical methodologies, systematic architecture, data mining, and analysis tools. The most crucial phase in big data is data

ARO-The Scientific Journal of Koya University
Vol. X, No. 2 (2022), Article ID: ARO.10975. 11 pages
DOI: 10.14500/aro.10975
Received 01 May 2022; Accepted: 28 August 2022
Review paper; Published: 08 October 2022
Corresponding author's e-mail: safa.s@csw.uobaghdad.edu.iq

 http://doi.org/10.14500/aro.10975

research, which involves examining significant values, making recommendations, and making judgments and decision support tools that have gained popularity, such as executive information systems and online analytical processing. Therefore, data analysis and interpretation complexity encourage researchers and companies to use algorithms that process real-time data, analyze it, and produce highly accurate analytics results. In addition, data analysis can be used to investigate potential values where this information can be used for business development and performance enhancement, such as predictive analytics that can make future predictions. Data analytics is a wide, dynamic and complex field because data comes in different types and grows significantly. Furthermore, the purpose of the analysis varies depending on the type of application required (Schwarz, Schwarz and Black, 2014; Harfouchi, et al., 2017; Rajaraman, 2016). Hence, data analytics aims to answer three categories of questions in general. As shown in Fig. 1, these elucidate what happened in the past, what is happening now, and what is anticipated (Ghavami, 2020).

As a result, processing and obtaining the necessary information from an extensive database cost a lot of time and processing power (Abdul Majeed, Kadhim and Subhi Ali, 2017). Moreover, interdisciplinary investigation makes it difficult for businesses to identify the specialist skills needed to conduct a large-scale reality check. Therefore, viable research provides critical features for completing this activity and overcoming the inaccessibility of analytical capabilities (Kashyap, 2019).

In other words, data analytics can be defined as a data science used to break data into individual components for personal inspection and integrate these components to create knowledge. Informally, Oracle and Cloudera have proposed a seven-step "value-chain" approach for extracting value using data analytics; these steps are as follows (Ghavami, 2020):

1. Objectives identification.
2. Business levers identification.
3. Data collecting.
4. Data cleaning.
5. Data modeling.
6. Data science team creating (i.e., building solid teams).
7. Optimize and repeat.

On the other hand, Dr. Carol Anne Hargreaves proposed another seven steps for the business analytics process in her data science process model, which also can be listed as follows (Ghavami, 2020):

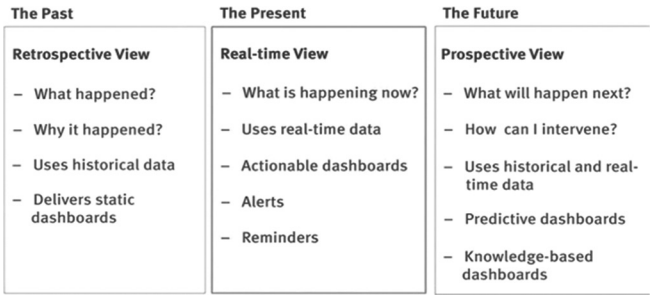

Fig. 1. Big data analytics' temporal questions (Peter Ghavami, 2020).

1. Business needs identification.
2. Explore the data.
3. Analyze the data.
4. Predict what is likely to happen.
5. Optimize (find the best solution).
6. Make a decision and measure the outcome.
7. Update the system with the results of the decision.

All kinds of data analytics processes, including the traditional Knowledge Discovery in Databases (KDD) process, and others such as (Mishra and Sharma, 2014), who proposed six steps for data analytics and (Chen, Mao and Liu, 2014) suggested three primary steps only and many others. These proposed systems depend on big data analytics tools that provide valuable knowledge for enhancing business. Typically, these methods can be used in different analytics models that can be divided into the following types:

A. Advanced analytics and predictive modeling

Machine learning, data science, and predictive modeling have grown widespread in every area where data analysis plays a key role (Butcher and Smith, 2020). Data mining is a sophisticated technique for evaluating large amounts of data. There are two forms of data analytics: Supervised/ unsupervised data analytics. Based on the findings of previous research studies, predictive modeling works in different scopes with solid chances of achieving efficient results when used with unsupervised analytics than with supervised analytics (Fan, et al., 2018). A prediction model is built by learning a dataset with a known outcome (classified results) and then determining the effects of unclassified cases (Shouval, et al., 2014). De Fortuny, Martens and Provost, showed in 2013 that when predictive models are created depending on varied and accurate data, they can provide a performance improvement even on a large amount of data. In this study, the researchers trained models and made predictions on sparse datasets using the Naive Bayes classifier. Data from several different predictive modeling applications are used to test the proposed method. The proposed method can conclude that the system with big data might be more efficient for predictive analytics operations. Consequently, organizations with more data and better understanding may gain significant competitiveness (De Fortuny, Martens and Provost, 2013). The data analytics technologies can be used in health-care systems as in 2014 when Pourhomayoun, et al., 2014 proposed a new system for remote health monitoring (Pourhomayoun, et al., 2014).

On the other hand, machine learning is one of the most critical data analytics approaches with significant facilitators of knowledge-intensive automation that can be used in many applications (Mishra and Sharma, 2014; Cearley, et al., 2018). Therefore, ML Algorithms are used in different applications such as medical, roads and many other applications with the risk of facing many problems in robustness, monitoring, alignment and systemic safety that should be handled (Rajpurkar, et al., 2017; Hendrycks, et al., 2021).

As an example of using neural networks for advanced data analytics systems, in 2017, Jain presented the implementation details for detecting telecommunication fraud using Data Stream Analytics and Neural Network classification-based Data Mining. The proposed method depends on Microsoft Azure's Event Hub and Stream Analytics components for fraud detection using a self-coded algorithm and a Data Mining Neural Network Pattern Recognition tool. The findings indicate that the proposed methodologies are accurate and efficient and may be extended to various cloud analytics systems and provide a foundation for big data analytics and mining (Jain, 2017). Furthermore, Talasila, et al. (2020) presented a novel neural network-based method for medical data analytics and disease prediction in 2020. They employed rough set theory to choose the most significant characteristics and then fed them into a Recurrent Neural Network for disease forecasting. As a result, the new technique had a 98.57% accuracy, more than the current accuracy presented by the existing methods for the heart disease dataset (Talasila, et al., 2020). Furthermore, dealing with big data can be aided by deep learning, which has the potential to extract complicated abstractions (Vu, et al., 2021). A new analytics model for distant physiological data was proposed based on powerful clustering techniques and multi-model classification. The proposed model is decomposed into several steps. The first is remote health monitoring and body sensor networks, which collect the data and send it to the analytical system. Then the data preprocessing and feature extraction step should be done to the received data. Followed by data sample clustering and group-specific feature selection, the multiple model classification must be done as a final step in this model. The proposed model was evaluated using a subset of data acquired from 600 heart failure patients through a remote health monitoring system. The proposed model dramatically improved prediction accuracy and performance (Pourhomayoun, et al., 2014). Whereas in 2019, Corizzo, Ceci and Malerba, 2019 were inspired by the goals of scientific studies sponsored by the European Commission and several national governments (Corizzo, Ceci and Malerba, 2019). They employed recommended methodologies based on distributed architectures, big data analytics, and predictive modeling research domains. The results of the proposed system give accurate predictions (temporal and geographical) that are scalable in big data. While in 2021(Hamarashid, Saeed and Rashid, 2021), a new paper was published to present a novel model for predicting the next word depending on the N-gram method with a sufficient increase in the number of N-grams used to reduce the time for predicting the next word in Kurdish dataset. The proposed model achieved results with accuracy up to 96.3%. Also, in 2021 another research was presented to produce a prediction model for healthcare centers based on machine learning algorithms and analysis methods (Moharram, Altamimi and Alshammari, 2021). In this paper, they analyze the input data to reduce the number of training data. Then, three machine learning algorithms were applied (Logistic Regression, JRip, and Hoeffding tree) to compare the results and select the best one for the proposed system. The proposed model produces similar effects in predicting appointment no-shows in pediatric outpatient clinics with roughly 90% classification accuracy. Furthermore, in the same year, Rocha, et al. (2021) used Principal Component Analysis techniques and unsupervised algorithms to perform better clustering. As a result, K-mean clustering algorithm shows the best results for clustering operation (Rocha, et al., 2021).

B. Model accuracy and optimization

There are several optimization strategies available by multiobjective optimization approaches (Zarchi and Attaran, 2019), (Wang, et al., 2011), (Jaouadi, et al., 2020). In 2020, Castellanos, et al. showed how to specify, deploy and track performance metrics in big data analytics applications based on domain-specific modelling and DevOps using a design process methodology based on the Attribute-Driven and Architecture analysis method technique (Castellanos, et al., 2020). Furthermore, many researchers employ the approximation model instead of the accurate numerical simulation model to improve the effectiveness of the current multiobjective optimization approaches in dealing with complicated engineering issues (Choi, Cho and Kim, 2018). Therefore, employing optimization techniques is the best method for identifying suitable model parameters (Kumar, et al., 2018). A data-driven predictive modeling strategy for forecasting surface roughness in additive manufacturing is developed to optimize the integrity of fabricated components. Various sensors of various sorts are used to collect data on temperature and vibration. An ensemble learning approach is used to train the surface roughness prediction model. A subset of these characteristics is chosen to enhance computational complexity and accuracy rate. As a result, the proposed model can provide accurate predicting results. At the same time, the frequency amplitude of the build plate vibrations, the extruder vibrations, and the temperatures influence the outcome (Li, et al., 2019). Whereas in the education sector, Tran, et al. (2019) have published a paper that described the benefits of Federated Learning and suggested a new system by establishing Federated Learning over a wireless network. This paper fills the trade-offs between computation and communication latencies caused by learning accuracy level, Federated Learning time, and energy consumption of mobile user equipment. They found the globally optimal solution by finding the confined methods to all sub-problems. This solution provides exciting insights into design issues through the ideal Federated Learning over wireless network learning (time, accuracy, and user equipment's energy cost) obtained through numerical and theoretical analysis (Tran, et al., 2019). In 2019 Zou, et al., proposed a new vehicle evaluation prediction model (Zou, et al., 2019). This model is used to optimize the traditional logistic regression algorithm by studying the logistic mathematical model, designing the error function, using the gradient descent method to discover the regression coefficient, and optimizing the sigmoid function. Consequently, the training time and classification effect were enhanced, and the accuracy is maintained.

In 2020, Liu, et al. developed a new adaptive model for efficient multiobjective optimization. This model depends

on micro multi objective genetics to improve performance. The optimization results further demonstrate the proposed model's usefulness in real-life applications. However, this model needs more samples and local-densifying iterations to provide reliable optimization results (Liu, et al., 2020). Also, in 2020 Ben Seghier, et al. (2020) proposed a hybrid Artificial Intelligence model that aims to create a hybrid framework for predicting and analyzing stress intensity factors. This framework was built by building an adaptive neuro-fuzzy inference system, tuned using two meta-heuristic algorithms: genetic algorithm and particle swarm optimization. The proposed model outperformed the other AI models for accurate prediction, with R^2 = 0.9913, RMSE = 23.6, and MAE = 18.07. However, increasing the datasets generated based on actual test results or bigger finite element method computations that include a variety of ranges and materials might enhance prediction performance (Seghier, et al., 2020).

C. Natural language processing

Researchers in the discipline can leverage techniques developed to appropriately and accurately analyze language. For example, natural language techniques have computational assessments of various language features about specific goals, and deep learning techniques such as CNN are widely used in this area (McNamara, et al., 2017; Shamsaldin, et al., 2019). Hence, the Natural Processing techniques allow researchers to collect and analyze data to extract the information (Rajput, 2019). However, one of the significant obstacles in text categorization is the optimization problem. This problem can consider an analytical issue for document summarization, prompting a group of academics to create a nature-inspired optimization technique based on a multi-criteria optimization model linked to Artificial Bee Colonies (ABC). The suggested technique in 2018 yielded significant gains, with average increases of 31.09% (8.43%) and 18.63% (6.09%) in ROUGE-2 (ROUGE-L) compared to the best single-objective and multiobjective findings in the published studies (Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018). In the same year (Rashid, Mustafa and Saeed, 2018), Rashid, Mustafa and Saeed (2018) applied a stemmer to Kurdish text documents (KDC-4007 dataset). They used three algorithms: Support Vector Machine, Naïve Bays, and Decision Tree, to classify Kurdish text. After the preprocessing phase, they found that the support vector machine achieved the best accuracy among all the applied algorithms. In 2019, researcher Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018 continued developing the research proposed in the previous year by creating an indicator based on a multi objective Artificial Bee Colony. The developed system was tested on several datasets (the same datasets used in their previous research) and evaluated the results using a variety of measures. Consequently, the results for ROUGE-2 and ROUGE-L have improved to between 7.37% and 40.76% and 2.59% and 11.24%, respectively (Sanchez-Gomez, et al., 2019).

On the other hand, Yadav and Chatterjee (2016) describe an efficient and robust summarizing approach based on the meaning of essential words in the content for text summarization. Sentiment analysis is constantly utilized for large-scale text data analysis and subjectivity analysis. This study demonstrates that sentiment analysis may be used well for text summarization and provides an efficient way to summarize the content, particularly for 50% (Yadav and Chatterjee, 2016). Furthermore, researchers can employ a lexicon-based technique to examine students' responses. A new algorithm has been suggested to establish teachers' opinion results by extracting semantic meaning from students' comments, including intensifier words, and determining the amount of positive or negative thoughts. This method displays the instructors' opinion results, categorized according to the strength of the positive or negative sentences. However, utilizing a lexicon approach to sentiment analysis is not optimal because some crucial details might be lost (Aung and Myo, 2017).

A summarization system can be designed depending on the dataset's similarity or dissimilarity measures. The research performed by Saini, et al., 2019 presented effective feature summarization for text as a binary analysis issue. They use a multi objective binary differential evolution-based optimization technique. Differential evolution's solutions encode a potential subcategory of sentences to be included in the summary, and then assessed using objective functions such as the sentence's location in the document. The results show that good improvements were obtained depending on the dataset used and the objective function (Saini, et al., 2019). In 2021, another paper was proposed to perform data analysis using state-of-the-art techniques. Using syntax analysis, they developed a method capable of extracting the recent Toolkit for ATM Incidence investigative process taxonomy factors from free-text safety reports. Finally, they modify a Data-Driven Method capable of automatically determining the cause of the aircraft accident. The results demonstrate that when merely elevated predictions are considered, the model provided pilots' contribution is around 97% accurate and 94% for ATCo (Buselli, et al., 2021). In the same year, Vargas-Calderón, et al. (2021) presented a model used in healthcare applications to evaluate the quality of service in hospitals depending on client reviews. After the text extraction and cleaning step, the model was designed depending on multi-ML algorithms (Vargas-Calderón, et al., 2021). Furthermore, in 2021 Hryshchenko and Yaremenko implemented the bloom filter, naïve Bayesian classifier, and neural networks to categorize a batch of text data and determine both disadvantages and advantages of each method (Hryshchenko and Yaremenko, 2021). At the same time, Yaremenko, Rogoza and Spitkovskyi (2021) developed a neural network architecture that can process a large amount of data in real-time systems and handle the determined limitation of the applied mathematical models of the standard Neural Networks and Naive Bays (Yaremenko, Rogoza and Spitkovskyi, 2021).

D. Quantitative analysis (prediction and prognostics)

Quantitative analysis is concerned with quantifying and analyzing variables to arrive at conclusions. It entails using

statistical tools to analyze numerical data for answering questions such as who and when. Apuke published his work on predictor measurement heterogeneity by altering the degree of measurement error across derivation and validation scenarios. Hence, he generated hybrid predictor measurements using measurement error models (Apuke, 2017; Pajouheshnia, et al., 2019; Luijken, et al., 2019; and Luijken, et al., 2020). In 2021, Admiraal, et al. used 12 quantitative features gathered from various patient situations to train several types of machine learning algorithms. The research results show that machine learning employing quantitative features derived from collected data has a better precision than visual data analysis in predicting poor prognosis following cardiac arrest, making it a potential alternative to visual analysis (Admiraal, et al., 2021). In 2022, Luijken, Song and Groenwold proposed a paper to analyze the expected predictor measurement diversity impact. In period outcome data, simulation research was conducted to examine the influence of predictor measurement variation across validation and implementation settings. The application of quantitative prediction error analysis was demonstrated with an illustration of forecasting the 6-year probability of acquiring second type diabetes with variability in the predictor body mass-index measurement. As a result of this paper, all situations of predictor measurement variability resulted in the poor measurement of prediction models, and overall accuracy was lowered. Furthermore, it increased random predictor measurement variability (Luijken, Song and Groenwold, 2022).

Moreover, artificial systems generate high-level data representations from large-scale data, particularly unlabeled data, which is plentiful in Big Data (Chen and Lin, 2014), (Najafabadi, et al., 2015). In 2020, Zhong, Yu and Ai proposed the big data-based hierarchical deep learning system in the context of employing deep learning for data analytics. This system uses behavioral and content features to interpret network traffic patterns and information encoded in the payload. When several machines are deployed, the findings of this suggested system demonstrate that it may boost the detection rate of intrusive attacks and reduce the time spent significantly (Zhong, Yu and Ai, 2020).

E. Ensemble of models (data analytics prediction framework)

In many real-world applications, the availability of classified data is restricted, making it challenging to detect and eliminate duplicate and unnecessary variables from the feature-set, particularly in high-dimensional applications. This circumstance naturally happens in many real-life situations when a large amount of data can be acquired inexpensively and quickly. Yet, the manual classification of samples is time-intensive and cannot be assumed. Many approaches were suggested to improve accuracy in machine learning; one of these approaches is to aggregate the output of several learners. Ensemble Learning is a term used to describe this approach. Bagging, boosting, stacking, and error-correcting output are the four methods for merging several models (Wang, et al., 2014). The learning under supervision finds clusters with high probability densities in individual classes. It is employed when there is a reference value and training set with the variables to cluster (Dean, 2014). Whereas in unsupervised learning, feature selection seeks to locate meaningful subsets of features that yield best groupings clustering by clustering "similar" items together using any similarity metric (Nag and Mitra, 2002), (Dy and Brodley, 2004), (Hong, et al., 2008), (Elghazel and Aussem, 2010). In 2011, Rajon, Shamim and Arif proposed a complete framework that designed and implemented a generic product-independent e-market model for emerging economies. This paper's fundamental contribution is creating and executing a generic e-marketplace model for emerging economies where agriculture is widely practiced and a thriving manufacturing sector. A comprehensive examination of the utility and efficacy of establishing e-Commerce and e-Commerce services has also been presented (Rajon, Shamim and Arif, 2011). Rajon, Shamim and Arif proposed a method based on the random sample partition model that retains the statistical features of the data set in each data block in 2018. They presented the Alpha framework, which consists of three primary layers for data administration, batch management, and data analysis, to enhance the efficiency of big data analysis with Random Sample Partition blocks. The results show that the proposed method can provide approximate results for data analysis tasks such as data summarization and the Alpha framework for Big Data Analysis tasks (Salloum, et al., 2018). In the same year, Yu, et al. (2018) design a model to demonstrate how boosting and bagging approaches can be compared to produce better explanatory models to prove that the ensamble approaches are more suitable for some problems than other approaches (Yu, et al., 2018).

On the other hand, Kumar, Singh and Buyya proposed a new ensemble learning-based workload prediction model in 2020, which makes use of excessive learning machines and weights their estimates with a voting engine. The optimized weights are chosen using a metaheuristic algorithm motivated by the black hole theory. The results demonstrate the approach's superiority over conventional methods, with a reduction in RMSE of up to 99.20% (Kumar, Singh and Buyya, 2020). Whereas in 2021 a new framework based on features modeling and ensemble learning to predict query performance was proposed by Zaghloul, Salem and Ali-Eldin (2021) using Machine learning algorithm attempting to predict a performance metric based on the amount of time elapsed and ensemble learning (Zaghloul, Salem and Ali-Eldin, 2021).

III. Summary and Comparisons

Data analytics methods and techniques have more and more applications in life, and performance enhancement solutions are widely applied. It is essential to improve the efficiency of any application when dealing with big data by enhancing the result accuracy and processing time; this will be done by analyzing the input data and extracting

only the relevant information that the application needs. Depending on this principle, many types of research papers in different scopes were published to propose new techniques that can be used to enhance the analysis results. This paper provides a survey of these research papers as summarized in Table I. The research papers that proposed new methods in the literature review section were presented in this table; to illustrate the research scope: The field in which this research was developed, the main research issues: Used to show the problems that the research tries to solve, the research techniques used to describe the method and tools that used to implement this research, and the main research findings used to describe the most important results obtained or concluded from the published research.

It can be seen from the summary of the research papers presented in Table I that most researchers suggested methods that used machine learning algorithms. Accordingly, this paper discusses the proposed methods' details and their impact on the results, as shown in Table II. This table focuses on the strengths or offers the capabilities of ML algorithms in

Table I
The Categorization of the Discussed Previous Works in Terms of Researches Scope, Main Research Issues, Main Research Techniques, and Main Research Findings

References	Research scope	Main research issues	Main research techniques	Main research findings
De Fortuny, Martens, and Provost, 2013	Big data	Critical prediction jobs	Multivariate Bernoulli Naive Bayes	Organizations with more data selection provide a better understanding of improving the system performance
Pourhomayoun, et al., 2014	Healthcare systems	Remote health monitoring	Multi-model approach	Improved the prediction accuracy and performance
Corizzo, Ceci and Malerba, 2019	Energy sector	Accurate big data analytics and predictive modeling	Multi-model approach	Accurate results of predictions with big data
Moharram, Altamimi, and Alshammari, 2021	Healthcare systems	Reduce the number of training data	ML algorithms	The best algorithm used was the Hoeffding tree, with a classification accuracy of 90%
Rocha, et al., 2021	Human development indicators	Classify the departments of peru according to their human development index using clustering techniques	Multi-model approach	
Li, et al., 2019	Extrusion-based additive manufacturing	Optimize the integrity of fabricated components	ML algorithms	Providing accurately predicted results
Tran, et al., 2019	Education sector	Analyze the trade-offs between (computation *vs*. communication latencies) (Learning time *vs*. energy consumption)	Confined methods to all sub-problems	Provide effective cost, time, and accuracy
Zou, et al., 2019	Vehicle evaluation	Many iterations and training vast amounts of data take a long time	ML algorithms	The training time is reduced, the classification effect is enhanced, and the accuracy is maintained
Liu, et al., 2020	Multiobjective optimization	High computational cost	Multi-model approach	The proposed model's useful in real-life applications
El and Ben, 2020	Stress intensity factor	Prediction of stress intensity factor	ML algorithms	Provide accurate predictions
Sanchez-Gomez, Vega-Rodríguez and Pérez, 2018	Text summarization	Find the essential information from a document collection	Optimization algorithms	Provide essential improvements compared to the traditional approaches
Sanchez-Gomez, et al., 2019	Text summarization	Find the essential information from a document collection	Optimization algorithms	Provide improved performance and accurate results compared to the previous version of this proposed model
Yadav and Chatterjee, 2016	Text summarization	Find an efficient way to summarize texts	Sentiment analysis	Provide an efficient and robust summarizing approach based on feelings of significant words
Aung and Myo, 2017	Education system	Analysis of students' comment	Sentiment analysis	Provide good results but not optimal because some crucial details might be lost
Saini, et al., 2019	Text summarization	Design an automatic text-summarization	ML algorithms	Provide good improvements to the traditional approaches
Buselli, et al., 2021	Air traffic management	Automation and digitization to maintain safety aviation	Multi-model approach	The model provides pilots' contribution is around 97% accurate and 94% for ATCo
Admiraal, et al., 2021	Healthcare systems	electroencephalography reactivity quantitative analysis of neurological prognostication following cardiac arrest	ML algorithms	Provide better precision than visual data analysis in predicting poor prognosis following cardiac arrest

(Contd...)

Table I
(Continued)

References	Research scope	Main research issues	Main research techniques	Main research findings
Luijken, Song and Groenwold, 2022	Healthcare systems	Analyze the effect predictor measurement heterogeneity	Quantitative prediction error analysis	Proves that the increasing random predictor measurement heterogeneity will decrease the model discrimination
Rajon, Shamim and Arif, 2011	Electronic commerce	Design a general prototype for the e-marketplace	General framework designing tools	They are creating and developing an e-market prototype that is product-agnostic
Salloum, et al., 2018	Big data	Making the extensive data analysis is feasible when the volume of data exceeds the available computation power	ML algorithms	Enhance the efficiency of extensive data analysis
Kumar, Singh, and Buyya, 2020	Cloud systems	These systems must allocate and deallocate resources with low operational cost and maintain the quality of services	ML algorithms	Provide accurate predictions by reducing the error prediction
Zaghloul, Salem, and Ali-Eldin, 2021	Query optimizer	Attempt to predict a performance metric	ML algorithms	Provide effective query performance metric
Jain, 2017	Fraud detection	Provide an efficient method for fraud detection using a self-coded	ML algorithms	Provide accurate and efficient cloud analytics
Talasila, et al., 2020	Healthcare systems	Provide an accurate method for diseases prediction	ML algorithms	Provide prediction accuracy up to 98.57%
Zhong, Yu, and Ai, 2020	Intrusion detection	Intrusion detection depends on data analytics	ML algorithms	Boost the detection rate of intrusive attacks and reduce the time spent significantly

ML: Machine learning

the proposed methods. It has been formulated to include: the ML algorithms used to design the proposed method and the techniques used to describe how these algorithms are used to build the proposed method. Furthermore, this table illustrates the effect of the proposed method on the system performance and the efficiency of the results. The most vital limitations are also presented in this table to describe the challenges and problems in the proposed methods. As we have seen in (Admiraal, et al., 2021) and (Zhong, Yu, and Ai, 2020), there are limitations in the number of models used to test the proposed method, so it may have some implementation issues when used with a large amount of data. Furthermore, (Kumar, Singh, and Buyya, 2020), there is another issue: people must determine the number of networks and nodes in hidden networks. This causes the system to need human intervention at the data entry stage and is not entirely automated. While in (Jain, 2017), the proposed model was designed without any dynamic implementation, which causes the possibility of facing several problems when applied to real-time data.

IV. Discussion

Data analytics aims to provide meaningful and relevant information. However, many users are uncertain about the sort of analysis to perform on data collection and which kinds of visual data presentation are appropriate. This paper presented a comprehensive review of data analytics techniques to help researchers construct an accurate and effective analytics tool to be used efficiently in the intentional system. This leads to utilizing these

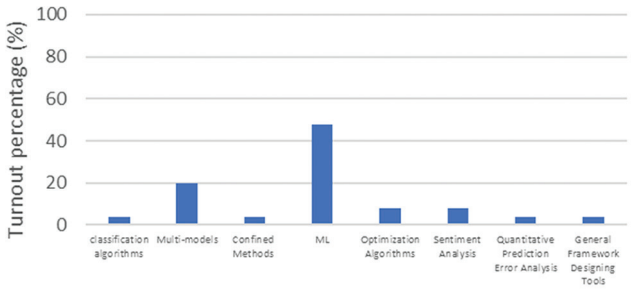

Fig. 2. The percentage of turnout for each method.

analytics tools in the best way to provide better privacy, less power, more efficient results, and economic services rather than relying only on standard analytics tools to solve the research problem in different scopes. According to Table I, ML algorithms have shown the highest usability in data analysis systems than other algorithms because they provide good accuracy with higher performance capacity in multiple areas of industrial, commercial, agricultural, health, education, and mining activity. As a result, these algorithms contribute to the development, increased employment, the contribution of mining techniques and increased business investment. Furthermore, Fig. 2 illustrates the turnout percentage, which shows that ML algorithms are superior to other methods. Therefore, the ML improvements were presented in Table II to show the effectiveness of each proposed method.

Finally, it is worth mentioning that many challenges can be faced when adopting ML algorithms in data analytics systems. For example, processing time, computation power… etc. However, these challenges can be addressed using

Table II
Machine Learning Algorithms in Each Reference and its Contributions

References	The used ML algorithms	The design techniques	Improvement effects			Research limitation
			Time	Accuracy	Processing	
Moharram, Altamimi and Alshammari, 2021	logistic regression, JRip, and Hoeffding tree	The optimum machine-learning method was discovered by comparing three ML algorithms by computing recall and prescience for each	*	*	*	-
Li, et al., 2019	RF, AdaBoost, CART, SVR, RR, RVFL	Data was collected depending on multiple sensors, extracted and selected features from these data, then applied selective algorithms to these features		*		-
Zou, et al., 2019	Logistic regression	Using the logistic regression algorithm based on the gradient descent method and optimizing the sigmoid function	*	*	*	-
Seghier, et al., 2020	Genetic and particle swarm algorithms	Hybrid AI model (GA and PSO) for accurate prediction trained on data values obtained using the finite element method calculations	*	*	*	-
Saini, et al., 2019	Genetic operation inspired form genetic algorithm	Using the binary differential evolution and self-organizing map (based on genetic operations) to select a subset of sentences, then evaluate these sentences using the objective functions		*	*	-
Admiraal, et al., 2021	LR, SVM, RF, NN, and GTB algorithms	The RF classifier was the best machine-learning classifier after ML models were trained with twelve quantitative variables derived from data of 134 adult cardiac arrest patients		*		A limited number of samples for training operation
Salloum, et al., 2018	Hadoop MapReduce and data mining and machine learning algorithms	The data is assigned to the blocks sampling algorithm, followed by blocks analysis and ensemble estimates techniques. Finally, an ensemble model evaluation was applied to check whether it fits the criteria	*	*	*	-
Kumar, Singh, and Buyya, 2020	Extreme learning machines	They were using extreme learning machines and a voting engine to optimize the weights of the prediction outcomes. These weights are defined using a black hole-inspired metaheuristic algorithm		*	*	Heuristics are used to manually define the number of networks and the number of hidden nodes in each network
Zaghloul, Salem, and Ali-Eldin, 2021	XGBoost algorithm	They used real-world datasets to predict query performance using feature modelling and ensemble learning based on the duration of time elapsed	*	*	*	-
Jain, 2017	Neural network	The event hub and Stream analytics of microsoft azure with the neural network and self-coded algorithms were employed to detect fraud telecommunication		*	*	It cannot be applied to real-time dynamic data
Talasila, et al., 2020	Recurrent Neural Network	The most significant characteristics were selected using rough set and then fed into a recurrent neural network to predicate the system results	*	*		-
Zhong, Yu, and Ai, 2020	Deep learning	The hierarchical deep learning system used the behavioural and content features to interpret network traffic patterns and information encoded in the payload	*	*	*	Effective when applied to a small number of samples

ML: Machine learning

parallel and distributed frameworks and choosing appropriate algorithms to implement for each system.

V. Conclusion

In this paper, several articles are reviewed in different analytics models. In addition, this study addressed the specified standards algorithms used for designing each system. Furthermore, it highlighted the advantages and disadvantages of the specified big data analytics methodologies, evaluating them in terms of scalability, efficiency, precision, and privacy. Furthermore, the suitable employment of data analytics in different scopes and applications has been adjusted to construct how it can be used to provide a high-quality performance. It should be noted that this paper recognizes that the core function of machine learning is to offer analytical answers that can be developed based on the behavior of previous data models. As a result, this paper intends to provide simple research examining several proposed analytics technologies from various perspectives and fills in the gaps clearly in unknown information. From this examination, we can conclude that when the ML algorithms and data analytics self-tuning system feature selections have been used, they will improve performance compared to other approaches and techniques. Several paper investigations in many sectors have demonstrated the possibility of using machine learning algorithms in data analytics systems to improve performance speed and accuracy.

References

Abdul Majeed, G., Kadhim, A. and Subhi Ali, R. (2017). Retrieving encrypted query from encrypted database depending on symmetric encrypted cipher system method. *Diyala Journal For Pure Science*, 13(1), pp.183-207.

Abdul-jabbar, S.S. and George, L.E. (2017). Fast text analysis using symbol enumeration and Hashing methodology. *Fast Strings Search Process*, 58(1), pp.345-354.

Abkenar, S.B. Kashani, M.H., Mahdipour, E. and Jameii, S.M. (2020). Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 57, 101517.

Admiraal, M.M., Ramos, L.A., Delgado Olabarriaga, S., Marquering, H.A., Horn, J. and van Rootselaar, A.F. (2021). Quantitative analysis of EEG reactivity for neurological prognostication after cardiac arrest. *Clinical Neurophysiology*, 132(9), pp.2240-2247.

Apuke, O.D. (2017). Quantitative research methods : A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, 6(11), pp.40-47.

Aung, K.Z. Myo, N.N. (2017). Sentiment Analysis of Students'Comment Using Lexicon Based Approach. *IEEE/ACIS 16th International Conference on Computer and Information Science*, pp.149-154.

Ben Seghier, M., Carvalho, H., Keshtegar, B. and Correia, J.A.F. (2020). Novel hybridized adaptive neuro-fuzzy inference system models based particle swarm optimization and genetic algorithms for accurate prediction of stress intensity factor. *FFEMS*, 43(11), pp.2653-2667.

Buselli, I., Oneto, L., Dambra, C., Gallego, C.V., Martínez, M.G., Smoker, A. and Martino, P.R. (2021). Natural Language Processing and Data-Driven Methods for Aviation Safety and Resilience : From Extant Knowledge to Potential Precursors. Open Research Europe.

Butcher, B. and Smith, B.J. (2020). Feature engineering and selection: A practical approach for predictive models. *The American Statistician*, 74(3), pp.308-309.

Castellanos, C., Pérez, B., Varela, C.A. and Correal, D. (2020). A Model-Driven Architectural Design Method for Big Data Analytics Applications. *Proceedings 2020 IEEE International Conference on Software Architecture Companion*, *ICSA-C 2020*, pp.89-94.

Cearley, D.W., Natis, Y., Walker, M. and Burke, B. (2018). Top 10 Strategic Technology Trends for 2018. Gartner, Stamford. Available fromt: https://www.gartner.com/ngw/globalassets/en/information-technology/documents/top-10-strategic-technology-trends-for-2018.pdf [Last accessed on 2017 Oct 3].

Chen, M., Mao, S. and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), pp.171-209.

Chen, X.W. and Lin, X. (2014). Big data deep learning: Challenges and perspectives. *IEEE Access*, 2, pp.514-525.

Choi, B.C., Cho, S. and Kim, C.W. (2018). Kriging Model Based Optimization of MacPherson Strut Suspension for Minimizing Side Load using Flexible Multi-Body Dynamics. *International Journal of Precision Engineering and Manufacturing*, 19(6), pp. 873-879.

Corizzo, R., Ceci, M. and Malerba, D. (2019). Big Data Analytics and Predictive Modeling Approaches for the Energy Sector. *2019 IEEE International Congress on Big Data* (*BigDataCongress*), pp.55-63.

De Fortuny, E.J., Martens, D. and Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data*, 1(4), pp.215-226.

Dean, J. (2014). Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley and Sons, Hoboken. Available from: https://doc.lagout.org/Others/Data Mining/Big Data, Data Mining, and Machine Learning_Value Creation for Business Leaders and Practitioners %5BDean 2014-05-27%5D.pdf. [Last accessed on 2022 Apr 01].

Do Nascimento, I.J.B., Marcolino, MS., Abdulazeem, H.M, Weerasekara, I., Azzopardi-Muscat, N., Gonçalves, M.A. and Novillo-Ortiz D. (2021). Impact of big data analytics on people's health: Overview of systematic reviews and recommendations for future studies. *Journal of Medical Internet Research*, 23(4), p.e27275.

Dy, J.G. and Brodley, C.E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5, pp. 848-889.

Elghazel, H. and Aussem, A. (2010). Feature selection for unsupervised learning using random cluster ensembles. *Proceedings IEEE International Conference on Data Mining ICDM*, pp.168-175. Doi: 10.1109/ICDM.2010.137.

Faizan, M. Zuhairi, M.F., Ismail, S. and Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. *International Journal of Advanced Computer Science and Applications*, 11(12), pp.146-153.

Fan, C., Xiao, F., Li, Z. and Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159, pp.296-308.

Farhan, K.A. and Ali, M.A. (2017). Database Protection System Depend on Modified Hash Function. *2nd International Conference of Cihan University-Erbil on Communication Engineering and Computer Science*, p.2520-4777.

Ghavami P. (2020). *Big Data Analytics Methods*. 2nd ed. De Gruyter, Berlin.

Hamarashid, H.K., Saeed, S.A. and Rashid, T.A. (2021). Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji. *Neural Computing and Applications*, 33(9), pp. 4547-4566.

Harfouchi, F., Habbi, H., Ozturk, C. and Karaboga, D. (2017). Modified multiple search cooperative foraging strategy for improved artificial bee colony optimization with robustness analysis. *Soft Computing A Fusion of Foundations Methodologies and Applications*, 22(19), pp.6371-6394.

Hendrycks, D. Carlini, N., Schulman, J., Steinhardt, J. (2021) Unsolved Problems in ML Safety. ArXiv, Cornell Tech, pp.1-28. Available from: https://arxiv.org/abs/2109.13916

Hong, Z., Smart, G., Dawood, M., Kaita, K., Wen, S.W., Gomes, J. and Wu, J. (2008). Hepatitis C Infection and Survivals of Liver Transplant Patients in Canada, 1997-2003. *Transplantation Proceedings*, 40(5), pp.1466-1470.

Hryshchenko, O. and Yaremenko, V. (2021). A comparative analysis of text data classification accuracy and speed using neural networks, Bloom filter and naive Bayes. *Technology Audit and Production Reserves*, 5(2(61), pp.6-8.

Jain, V. (2017). Perspective analysis of telecommunication fraud detection using data stream analytics and neural network classification based data mining. *International Journal of Information Technology*, 9(3), p.1-8.

Jaouadi, Z., Abbas, T., Morgenthal, G. and Lahemer, T. (2020). Single and multi-objective shape optimization of streamlined bridge decks. *Structural and Multidisciplinary Optimization*, 61(4), pp.1495-1514.

Kan, M.Y. and Klavans, J.L., (2002). Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. *Proceedings of the ACM International Conference on Digital Libraries*, Wuhan, pp.36-45.

Kashyap, R., (2019). Big data analytics challenges and solutions. In: *Big Data Analytics for Intelligent Healthcare Management*, Academic Press, Cambridge, pp.19-41.

Khoshbakht, F., Shiranzaei, A. and Quadri, S.M.K., (2021). Role of the big data analytic framework in business intelligence and its impact : Need and benefits. *Turkish Journal of Computer and Mathematics Education*, 12(10), pp.560-566.

Kumar, D.U., Soon, T.K., Saad, M., Idna, I.M.Y., Mehdi, S. and Bend, H. (2018). Forecasting of photovoltaic power generation and model optimization : A review. *Renewable and Sustainable Energy Reviews*, 81, pp.912-928.

Kumar, J., Singh, A.K. and Buyya, R. (2020). Ensemble learning based predictive framework for virtual machine resource request prediction. *Neurocomputing*, 397, p.20-30.

Li, Z., Zhang, Z., Shi, J. and Wu, D., (2019). Prediction of surface roughness in extrusion-based additive manufacturing with machine learning. *Robotics and Computer Integrated Manufacturing*, 57, pp.488-495.

Liu, X., Liu, X., Zhu, Z. and Hu, L., (2020). An efficient multi-objective optimization method based on the adaptive approximation model of the radial basis function. *Structural and Multidisciplinary Optimization*, 63(4), p.1-19.

Luijken, K., Groenwold, R.H.H., Van Calster, B.E.W., Steyerberg, E.W. and Van Smeden, M. (2019). Impact of predictor measurement heterogeneity

across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*, 38(18), pp.3444-3459.

Luijken, K., Song, J. and Groenwold, R.H.H., (2022). Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. *Diagnostic and Prognostic Research*, 1, pp.1-11.

Luijken, K., Wynants, L., van Smeden, M., Van Calster, B., Steyerberg, E.W. and Groenwold, R.H.H., (2020). Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*, 119, pp.7-18.

Mariani, M. and Baggio, R. (2022). Big data and analytics in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management*, 34(1), pp.231-278.

McNamara, D.S., Allen, L.K., Crossley, S.A., Dascalu,M. and Perret, C.A., (2017). Natural language processing and learning analytics. In: *Handbook of Learning Analytics*. Ch. 8. Society for Learning Analytics Research, Alberta, pp.93-104.

Mishra, R. and Sharma, R. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives. *International Journal of Computer Science and Mobile Computing*, 46(6), pp.27-35.

Moharram, A., Altamimi, S. and Alshammari, R., (2021). Data Analytics and Predictive Modeling for Appointments No-show at a Tertiary Care Hospital. *2021 1st International Conference on Artificial Intelligence and Data Analytics*, *CAIDA 2021*, pp.275-277.

Nag, A.K. and Mitra, A., (2002). Forecasting daily foreign exchange rates using genetically optimized neural networks. *Journal of Forecasting*, 21(7), pp.501-511.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E., (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), pp.1-21.

Pajouheshnia, R., van Smeden, M., Peelen, L.M., Groenwold, R.H.H., (2019). How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology*, 105, pp.136-141.

Patel, A., Singh, N.M. and Kazi, F. (2017). *Internet of Things and Big Data Technologies for Next Generation Healthcare*. Springer Cham, Berlin.

Pourhomayoun, M., Alshurafa, N., Mortazavi, B., Ghasemzadeh, H., Sideris, K., Sadeghi, B., Ong, M., Evangelista, L., Romano, P., Auerbach, A., Kimchi, A. and Sarrafzadeh, M. (2014). Multiple model analytics for adverse event prediction in remote health monitoring systems. In: *2014 IEEE Healthcare Innovation Conference*, *HIC 2014*, pp.106-110.

Rajaraman, V., (2016). Big data analytics. *Resonance*, 21(8), pp.695-716.

Rajon, S.A.A., Shamim, A. and Arif, M., (2011). A Generic Framework for Implementing Electronic Commerce in Developing Countries. *International Journal of Computer and Information Technology*, 1(2).

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P. and Na, A.Y., (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv.

Rajput, A., (2019). Natural language processing, sentiment analysis, and clinical analytics. In: *Innovation in Health Informatics: A Smart Healthcare Primer*, Academic Press, USA, pp.79-97.

Rashid, T.A., Mustafa, A.M. and Saeed, A.M. (2018). Automatic kurdish text classification using KDC 4007 dataset. In: Advances in Internetworking, Data and Web Technologies, Barolli, L., Zhang, M. and Wang, Z., editors. *Lecture Notes on Data Engineering and Communications Technologies*. Vol. 6, Springer, Cham, Berlin, pp.187-198.

Rocha, J.L.M., Zela, M.A.C., Torres, N.I.V. and Medina, G.S. (2021). Analogy of the application of clustering and K-means techniques for the approximation of values of human development indicators. *International Journal of Advanced Computer Science and Applications*, 12(9), pp.526-532.

Russell, S. and Norvig, P. (2020). *Artificial Intelligence a Modern Approach*. 4th ed. Prentice Hall, Hoboken.

Saini, N., Saha, S., Chakraborty, D. and Bhattacharyya, B., (2019). Extractive single document summarization using binary differential evolution optimization of different sentence quality measures. *PLoS On*e, 14(11), p.e0223477. [Last accessed on 2022 Apr 01].

Salloum, S., Huang, J.Z., He, Y. and Chen, X. (2018). An asymptotic ensemble learning framework for big data analysis. *IEEE Access*, 7(c), pp.3675-3693.

Sanchez-Gomez, J.M., Vega-Rodriguez, M.A. and C., Perez, C.J. (2019). An Indicator-based Multi-objective optimizationapproach applied to extractive multi-documenttext summarization. *IEEE Latin America Transactions*, 17(8), pp.1291-1299.

Sanchez-Gomez, J.M., Vega-Rodríguez, M.A. and Pérez, C.J. (2018). Extractive multidocument text summarization using a multiobjective artificial bee colony optimization approach. *Knowledge-Based Systems*, 159, pp.1-8.

Schwarz, C., Schwarz, A. and Black, W.C., (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34(1), pp.1191-1208.

Shamsaldin, A., Rashid, T.A., Fattah, P. and Al-Salihi, N.K., (2019). A study of the convolutional neural networks applications. *UKH Journal of Science and Engineering*, 3(2), pp.31-40.

Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R. and Nagler, A., (2014). Application of machine learning algorithms for clinical predictive modeling: A data-mining approach in SCT. *Bone Marrow Transplantation*, 49(3), pp.332-337.

Smith, M., Szongott, C., Henne, B. and von Voigt, G., (2013). Big Data Privacy Issues in Public Social Media. *IEEE International Conference on Digital Ecosystems and Technologies* [Preprint].

Talasila, V., Madhubabu, K., Mahadasyam, M.C., Atchala, N.J. and Kande, L.S., (2020). The prediction of diseases using rough set theory with recurrent neural network in big data analytics. *International Journal of Intelligent Engineering and Systems*, 13(5), pp.10-18.

Tran, N.H., Bao, W., Zomaya, A., Nguyen Minh, N.H. and Hong, C.S., (2019). Federated Learning over Wireless Networks: Optimization Model Design and Analysis. *Proceedings-IEEE INFOCOM, 2019-April(1)*, pp.1387-1395.

Vargas-Calderón, V., Ochoa, A.M., Nieto, G.Y.C. and Camargo, J.E., (2021). Machine learning for assessing quality of service in the hospitality sector based on customer reviews. *Information Technology and Tourism*, 23(3), pp.351-379.

Verma, J.P. and Agrawal, S., Patel, P. and Patel, A., (2016). Big data analytics: challenges and applications for text, audio, video, and social media data. *International Journal on Soft Computing Artificial Intelligence and Applications*, 5(1), pp.41-51.

Vu, T., Belussi, A., Migliorini, S. and Eldway, A., (2021). Using deep learning for big spatial data partitioning. *ACM Transactions on Spatial Algorithms and Systems* (*TSAS*), 7(1), p.1-37.

Wang, J., Tang, Y., Nguyen, M. and Altintas, I., (2014). A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning. *Proceedings of the 2014 International Symposium on Big Data Computing*, *BDC 2014*, pp.16-25.

Wang, X.D., Hirsch, C., Kang, S. and Lacor, C., (2011). Multi-objective optimization of turbomachinery using improved NSGA-II and approximation model. *Computer Methods in Applied Mechanics and Engineering*, 200(9-12), pp.883-895.

Yadav, N. and Chatterjee, N., (2016). Text Summarization using Sentiment Analysis for DUC Data. *International Conference on Information Technology*, pp.5.

Yaremenko, V.S., Rogoza, W.S. and Spitkovskyi, V.I., (2021). Application of neural network algorithms and naive bayes for text classification. *Journal of*

Theoretical and Applied Information Technology, 99(1), pp.125-134.

Yu, C.H. Lee, H.S., Lara, E. and Gan, S., (2018). The ensemble and model comparison approaches for big data analytics in social sciences. *Practical Assessment Research and Evaluation*, 23(17).

Zaghloul, M., Salem, M. and Ali-Eldin, A., (2021). A new framework based on features modeling and ensemble learning to predict query performance. *PLoS One*, 16(10), pp.1-18.

Zarchi, M. and Attaran, B., (2019). Improved design of an active landing gear for a passenger aircraft using multi-objective optimization technique. *Structural and Multidisciplinary Optimization*, 59(5), pp.1813-1833.

Zheng, L. and Guo, L., (2020). Application of big data technology in insurance innovation. *Journal of Physics Conference Series*, 1682(1), pp.285-294.

Zhong, W., Yu, N. and Ai, C., (2020). Applying big data based deep learning system to intrusion detection. *Big Data Mining and Analytices*, 3(3), pp.181-195.

Zou, X., et al. (2019). Logistic Regression Model Optimization and Case Analysis, Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019, pp.135-139.