Data Analytics and Techniques: A Review

Safa S. Abdul-Jabbar¹ and Alaa K. Farhan²

¹Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq
²Department 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 findings in the published paper. Finally, Sections IV and V integrate them into a model according to the task at hand.

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 (Khoshbakhht, 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
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. In effect, 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.
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):

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).

---

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

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

(Contd...)

ARO p-ISSN: 2410-9355, e-ISSN: 2307-549X
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 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 turnover 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>
<thead>
<tr>
<th>References</th>
<th>Research scope</th>
<th>Main research issues</th>
<th>Main research techniques</th>
<th>Main research findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luijken, Song and Groenwold, 2022</td>
<td>Healthcare systems</td>
<td>Analyze the effect predictor measurement heterogeneity</td>
<td>Quantitative prediction error analysis</td>
<td>Proves that the increasing random predictor measurement heterogeneity will decrease the model discrimination</td>
</tr>
<tr>
<td>Rajon, Shamim and Arif, 2011</td>
<td>Electronic commerce</td>
<td>Design a general prototype for the e-marketplace</td>
<td>General framework designing tools</td>
<td>They are creating and developing an e-market prototype that is product-agnostic</td>
</tr>
<tr>
<td>Salloum, et al., 2018</td>
<td>Big data</td>
<td>Making the extensive data analysis is feasible when the volume of data exceeds the available computation power</td>
<td>ML algorithms</td>
<td>Enhance the efficiency of extensive data analysis</td>
</tr>
<tr>
<td>Kumar, Singh, and Buyya, 2020</td>
<td>Cloud systems</td>
<td>These systems must allocate and deallocate resources with low operational cost and maintain the quality of services</td>
<td>ML algorithms</td>
<td>Provide accurate predictions by reducing the error prediction</td>
</tr>
<tr>
<td>Zaghoul, Salem, and Ali-Eldin, 2021</td>
<td>Query optimizer</td>
<td>Attempt to predict a performance metric</td>
<td>ML algorithms</td>
<td>Provide effective query performance metric</td>
</tr>
<tr>
<td>Jain, 2017</td>
<td>Fraud detection</td>
<td>Provide an efficient method for fraud detection using a self-coded</td>
<td>ML algorithms</td>
<td>Provide accurate and efficient cloud analytics</td>
</tr>
<tr>
<td>Talasila, et al., 2020</td>
<td>Healthcare systems</td>
<td>Provide an accurate method for diseases prediction</td>
<td>ML algorithms</td>
<td>Provide prediction accuracy up to 98.57%</td>
</tr>
<tr>
<td>Zhong, Yu, and Ai, 2020</td>
<td>Intrusion detection</td>
<td>Intrusion detection depends on data analytics</td>
<td>ML algorithms</td>
<td>Boost the detection rate of intrusive attacks and reduce the time spent significantly</td>
</tr>
</tbody>
</table>

ML: Machine learning

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


A RO p-ISSN: 2410-9355, e-ISSN: 2307-549X


