Artificial Intelligence Review (2024) 57:176
https://doi.org/10.1007/s10462-024-10811-5

Big data and predictive analytics: A systematic review of applications

Amirhossein Jamarani[1] · Saeid Haddadi[2] · Raheleh Sarvizadeh[3] · Mostafa Haghi Kashani[3] · Mohammad Akbari[4] · Saeed Moradi[5]

Published online: 17 June 2024

Abstract
Big data involves processing vast amounts of data using advanced techniques. Its potential is harnessed for predictive analytics, a sophisticated branch that anticipates unknown future events by discerning patterns observed in historical data. Various techniques obtained from modeling, data mining, statistics, artificial intelligence, and machine learning are employed to analyze available history to extract discriminative patterns for predictors. This study aims to analyze the main research approaches on Big Data Predictive Analytics (BDPA) based on very up-to-date published articles from 2014 to 2023. In this article, we fully concentrate on predictive analytics using big data mining techniques, where we perform a Systematic Literature Review (SLR) by reviewing 109 articles. Based on the application and content of current studies, we introduce taxonomy including seven major categories of industrial, e-commerce, smart healthcare, smart agriculture, smart city, Information and Communications Technologies (ICT), and weather. The benefits and weaknesses of each approach, potentially important changes, and open issues, in addition to future paths, are discussed. The compiled SLR not only extends on BDPA's strengths, open issues, and future works but also detects the need for optimizing the insufficient metrics in big data applications, such as timeliness, accuracy, and scalability, which would enable organizations to apply big data to shift from retrospective analytics to prospective predictive if fulfilled.

Keywords Big data · Predictive analytics · Big data applications · Systematic review

1 Introduction

Big data analytics refers to various techniques to analyze data, extract information, and gain insights from large-scale datasets with complex patterns because conventional data-processing views cannot be easily dealt with. Data having a huge number of samples have higher statistical power, as over-complicated data with great dimensional feature space might lead to a greater rate of false discovery (Breur 2016). Data capture, data storage, search, visualization, transfer, sharing, query, update, data analysis, information privacy, and data source are the most important challenges in big data (Anagnostopoulos

Extended author information available on the last page of the article

et al. 2016). The recent application of the term big data seems to refer to using user behavior analytics, predictive analytics, or different advanced data analytic methods that obtain value from data, and rarely to a data set of special size. In essence, available data quantities now are huge; however, it is not the most related feature of this data ecosystem (Rodríguez-Mazahua et al. 2016). Similarly, medical experts, researchers, business people, advertising, and governments usually have challenges dealing with large data sets in urban informatics, business informatics (Bhuimali et al. 2018), fintech (Thakuriah et al. 2017), and web surfing. Scientists face constraints in e-science activities, including meteorology (Fathi et al. 2021), biology (Gharajeh 2018), genomics (Wong 2016), complicated physics simulations, and environmental studies. Naturally, the big data ecosystem is explained by: Value, Veracity, Velocity, Variety, and Volume.

Value is the worth of the data being extracted. Data has no use or importance in itself, but it requires to be changed into a valuable to extract information. In addition, veracity defines data quality and value. It majorly affects the quality of captured data and the exact analysis. Velocity specifies the pace at which the data is generated and processed to satisfy the needs and challenges of the growth and development path. Variety elucidates the type and nature of the data. It assists those in analyzing it to use the resulting insight influentially. Lastly, Volume outlines the amount of generated and stored data. The data size defines the value and potential insight and whether or not it can be regarded as big data (Fathi et al. 2021).

With the emergence of systems of big data, predictive analytics has gained prominence. Enterprises have bigger and greater data pools in big data platforms, which has led to an increase in data mining opportunities to obtain predictive insights (Mohamed et al. 2020). The commercialization of machine learning tools has also expedited this trend, which results in emerging demands for predictive analytic services (Casado and Younas 2015). A large volume of techniques is used by predictive analytics to help organizations forecast outcomes, techniques that continue to develop with the expanding adoption of big data analytics. Big Data Predictive Analytics (BDPA) defines frameworks and systems that gather, analyze, and give an interpretation of great variety, volume, velocity, veracity, and value data to show patterns, trends, and relationships within data to find challenges and opportunities, to foresee future happenings, and direct decision making in contexts of applications.

Predictive analytics turns to statistical methods from forecasting modeling, machine learning, and data mining, which examine recent historical facts to foresee events or future that are unrevealed (Nyce 2007). In the business domain, predictive models apply the patterns that are found in transactional and historical data to find chances and dangers. Models find relations among lots of factors to permit assessing the risks or potentials inherent in a special set of conditions to direct decision-making for candidate transactions (Coker and Pulse 2014). In this functional effect, related to technical methods, an estimated score demonstrating a probability is provided by predictive analytics for each person, a customer, an employee, a healthcare patient, a vehicle to inform, determine, or influence organizational processes that are related to a wide range of people in healthcare, manufacturing, fraud detection, marketing, and the like. Predictive analytics is used in healthcare (Etemadi, et al. 2023), smart cities (Karimi et al. 2021), marketing (Miles et al. 2017), retail (Huang et al. 2019), actuarial science (Homer et al. 2017), social networking (Failed 2017a), financial services (Ouahilal et al. 2016), insurance (Longhi and Nanni 2019), telecommunication (Failed 2018a), mobility (Moreira-Matias et al. 2016), travel (Amirian et al. 2016), child protection (Russell 2015), pharmaceuticals (Sohrabi 2019), capacity planning (Delfmann et al. 2019), and other fields. This Systematic Literature Review (SLR) is arranged with the central aim of *recognition, taxonomic classification, comparison of the big data analytic*

approaches, and systematic contrast of the current articles that concentrate on planning, executing, and authentication of big data analytics. In accordance with the former purposes, we have made a determined effort to answer the following study questions: What are the fields of prediction analysis applications in big data? What are the evaluation metrics of predictive analytics using big data? What evaluation methods are used in BDPA? What are the tools and environments in BDPA? And, what are the challenges and future issues of BDPA?

We accompanied the recommendations in Kitchenham (2004); Brereton et al. 2007). Our main goal is to possess a systematic identification and assort classification of the current authenticated achievements on BDPA. Up to the present date, according to our inspections and studies, only the minority of SLRs investigated BDPA thoroughly. Another factor to consider is that none of them presented a complete and precise review article in the field of BDPA. Furthermore, since BDPA is an ultra-critical and sensitive field, it is necessary to provide a comprehensive study. Due to this, we have studied 109 articles to provide an exhaustive systematic review of predictive analytic methods that exploit big data for better performance. Although there are some reviews on various big data approaches, the vast majority of reviews do not focus on predictive analytic challenges, open issues, benefits, and drawbacks. To overcome this shortcoming, a systematic review of the literature and overview is presented for predictive analytics by using big data. This article assists researchers in gaining an overall understanding of various approaches that utilize big data for predictive analytics. The key contributions of this study are summarized as follows:

- Representing a systematic review of the predictive analytic approaches by utilizing big data
- Providing a technical and comprehensive taxonomy that categorizes various applications in BDPA as depicted in Fig. 1.
- Making a detailed comparison of the applied evaluation metrics and methods, tools, and the pros and cons of each study

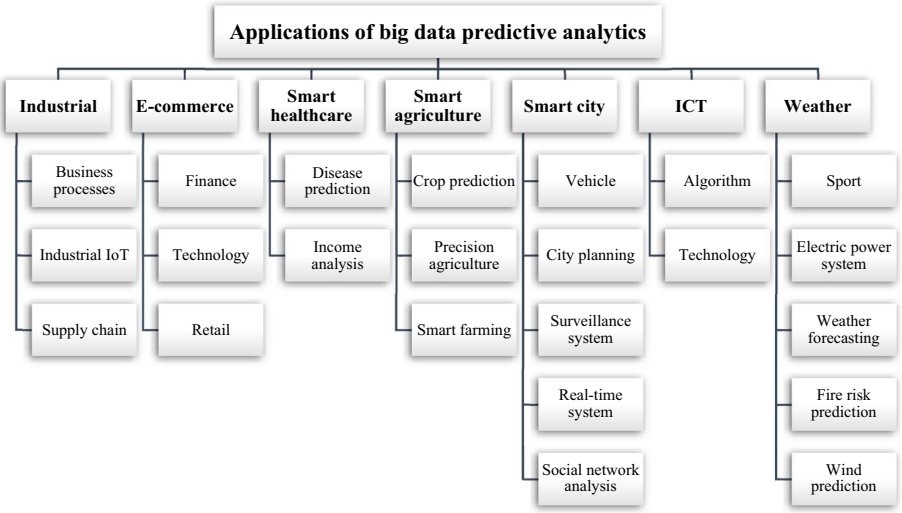


Fig. 1 Taxonomy of prediction analysis applications in big data

- Determining open challenges and future trends of predictive analytics by applying big data

The remaining parts of this article are structured as follows. Section 2 considers some relevant review works. Section 3 provides the research methodology, including research questions and the article selection process. Section 4 provides a classification for big data predictive analytics. Section 5 shows the comparisons and results of the articles reviewed. Section 4.6 refers to open issues and further studies. And, the final results are illustrated in Section 4.6.1.

2 Related work

Different kinds of review articles have been prepared in the field of BDPA. And, in this section, some of these articles are reviewed and compared with our work.

Philip Chen and Zhang (Philip Chen and Zhang 2014) surveyed big data applications, big data opportunities, and upcoming state-of-the-art methods and technologies that are being adopted to tackle the challenges of big data. To deal with problems, they provided techniques to tackle the pitfalls of big data, such as cloud computing, quantum computing, and biological computing. However, the article's procedure for selected articles was not provided, and it was not systematic.

Kumaresan and Rajakumar (Kumaresan and Rajakumar 2015) provided details on predictive analytics. The area of predictive analytics was discussed. The authors introduced several tools and techniques in predictive analytics. The given list of reviews was examined and deliberated upon to ascertain the utilization of predictive analytics by researchers in industrial and medical contexts. Various techniques and approaches were referenced in the process. This research discussed various issues and challenges of predictive analytics, available tools, applications, and modeling techniques in big data. However, no guideline for future research was suggested; recent year's published articles were not considered, no taxonomy was prepared, the article selection process was not transparent, and it was not an SLR.

Banumathi and Aloysius (Banumathi and Aloysius 2017) provided a review of different predictive analytic applications and approaches. Analytic methods, with dissimilar perspectives based on applications and data variety, were considered. Some of the applications discussed are big data in health care, hotel governance, consumer orientations, higher education, and data e-governance. The authors presented predictive approaches adapted for different applications with challenges and suggestions. The article specifically identified the main applications that depend thoroughly on BDPA solutions and already adopted themselves as one of the big data entities. Nevertheless, this study did not deliver an SLR, and the article selection process was not clear.

Poornima and Pushpalatha (Poornima and Pushpalatha 2018) introduced the thought of using predictive analytics and data mining methods on various medical datasets to foresee different illnesses with all advantages, disadvantages, and accuracy levels included that are related to future approaches to big data. The review list was discussed and presented to see how different authors applied predictive analytics for medicine and business and how they were regarded. The algorithms and techniques were also referred to while being applied to big data. However, this research did not represent an SLR, and no taxonomy was prepared. In addition, the article selection process is not clear. In other words, possible future studies

were not presented. Ghani, et al. (Ghani et al. 2019) survey looked at different angles in social media big data analytic topics. The authors arranged the survey based on different features. They provided a discussion on the applications of social media big data analytics by taking methods and quality tokens from various studies. Open research challenges and future works in big data analytics are introduced, but the authors did not consider noting the covered years of the articles reviewed, and it was not an SLR.

Kaffash, et al. (Kaffash et al. 2021) worked on a review taking into account the big data algorithms and applications on intelligent transportation. In this study, no taxonomy was organized. Mallika and Selvamuthukumaran (Mallika and Selvamuthukumaran 2022) provided a review of prospective precision medicine by utilizing big data. They illustrated the most frequently used tools and computational platforms on which precision medicine based its foundation on the role of big data; however, their review had no taxonomy nor any article collection process. Nobanee, et al. (Nobanee et al. 2022) only reviewed the applications of big data in the area of credit risk assessment. The authors described the notions of credit risk and types of credit risk and then connected the relation of big data with credit risk management. However, one of the weaknesses of their work was the limitation of their scope and the number of reviewed articles.

Ikegwu, et al. (Ikegwu et al. 2022) reviewed big data analytics in data-driven industries, covering tools, data sources, challenges, solutions, and research directions. The authors discussed different classification methods, data characteristics, and real-life applications across sectors. The study reviewed related studies to big data analytics, which were published only between 2013 and 2021. Also, the study is not a systematic literature. Himeur, et al. (Himeur, et al. 2023) provided an overview of the fall shorts of building automation and management systems (BAMSs) in terms of performance evaluation, energy consumption analysis, and security. The authors reviewed various AI-based tasks, presented existing frameworks, and discussed challenges of BAMSs performance in intelligent buildings.

In screening to manage diabetes long and short-term complications, predictive models were introduced by Cichosz, et al. (Cichosz et al. 2016). The authors also presented a systematic mapping study (SMS). These models have been created to manage diabetes and its related problems, and there has been a tremendous rise in the number of studies on these models recently. A linear regression or multiple logistics was applied to develop the prediction model, probably because of its clear functionality. Finally, in order to prove the usefulness of prediction models, they have to show their impact, or in other words, their application should yield more satisfactory outcomes in patients. Despite all efforts made to build these predictive models, a considerable scarcity in impact studies was observed. However, there was not a systematic review in this study, and the method of selecting articles was unclear. Newly published articles were also excluded.

To reach high-level comprehension in big data manufacturing, O'Donovan, et al. (O'Donovan et al. 2015) provided an SMS. Their contributions were some reports on the current state of work concerning big data approaches in assembling, such as methods of research taken into account, sectors in producing where big data exploration were concentrated, and results from big data research projects. The authors classified their study based on different research questions and answers. Nonetheless, their study did not provide any information regarding the covered years of the studied articles.

Different predictive models were classified by Muthukrishnan, et al. (Muthukrishnan et al. 2017), which were applied to monitor and improve the performances of students in educational settings similar to schools or universities. Within the educational data mining methodology, the whole areas were analyzed, two databases were selected, and systematic mapping research was conducted for this article. The main aim of the noted systematic

mapping study was to examine the current predictive analytic models within the educational environment of schools and other educational institutions. Due to the need to understand the functional applications linked to the approaches in healthcare, Mehta, et al. (Mehta et al. 2019) provided an SMS by considering artificial intelligence with big data. To examine the improvements in this field, the authors employed bubble plots to map the arrangements of publications. They categorized their reviewed article into different subclasses, which led to the creation of taxonomy; however, they did not review recently published articles.

Rahman and Reza (Rahman and Reza 2020) reviewed the non-functional requirements (NFRs) in big data. Afterward, they implemented a model to map the NFRs. The authors showed that some metrics, such as performance, scalability, and reliability are the most important factors in data-intensive systems. Biesialska, et al. (Biesialska et al. 2021) conducted an SMS to review agile software developments with the impact of big data. Their taken method of collecting articles was snowballing and manual search through the databases. In addition, articles were reviewed by the authors, in which their applications, company names, country, and per-industry usage were reviewed. Montero, et al. (Montero et al. 2021) took a systematic mapping approach to review big data quality models. The authors collected articles by providing an overview of their selection process, but they did not mention nor provide directions for future works on the quality models of big data.

The information system literature review was introduced by Ohiomah, et al. (Failed 2017b) on BDPA to find BDPA areas that were investigated before, but still needed greater focus. They suggested special research questions to be studied further and found out that big data arrival altered predictive analytic roles from such activities as generation and validation of theory to the more data-driven discovery of complicated patterns and relations among variables and evaluation of the probability of relationship occurrence in variables of a dataset. In this research, recently published articles were ignored. Mikalef, et al. (Mikalef et al. 2018) arranged an SLR on big data analytics to explain the system performance through which they should be leveraged to contribute to competitive productivity. They reviewed the research frameworks that were based on IT–business value, alongside the segments from strategic management. The authors focused on tools, technical methods, network analytics, and the infrastructure of big data analysis. Despite this, their review article did not pay attention to the recent date published articles.

Kolajo, et al. (Kolajo et al. 2019) tried to represent the flow of big data evaluation by providing a systematic review in order to recognize the tools and approaches. However, the authors did not provide a taxonomy for their study, and recently published articles were not included. Al-Sai, et al. (Al-Sai et al. 2020) provided an SLR to divide the schema and framework into five major groups of big data critical success factors, namely individuals, management, approaches, authorities, and companies. By answering three research questions during their survey, the authors tried to provide solutions to the key issues of big data analytics. Nonetheless, they did not research recently published articles to provide a more up-to-date SLR.

Rathore, et al. (Rathore et al. 2021) discussed the influencers on digital twinning. They identified research challenges and deficiencies that need to be worked on in the future to excel in digital twinning. The authors also divided their article into different sections, noting the scopes of manufacturing, medicine, transportation, education, business, and other industries in digital twinning. Naghib, et al. (Naghib et al. 2022) provided an SLR regarding the methods of how to manage big data in the Internet of Things (IoT). In their role as article organizers, they delineated four distinct categories: processes related to big data management (BDM), the BDM framework, quality attributes, and, finally, big data

analytics Georgiadis and Poels (Georgiadis and Poels 2022) came up to the conclusion that although there have been numerous studies in big data security assessment, there is still room and potential that needs to be fulfilled by more pertinent and methodological rules to lower data protection risks in systems that store big data analytic algorithms.

Acciarini, et al. (Acciarini et al. 2023) focused on reviewing the benefits of business model innovation with the use of big data to unleash companies to reach a comprehensive understanding of diverse applications. The authors offered guidance on harnessing the potential of big data in the industry. Shah, et al. (Shah et al. 2023) provided an SLR regarding the applications of BDPA in Supply Chain Risk Management (SCRM). The authors analyzed 68 selected articles, categorized them based on publication year, country, journal, application areas, and tools used. Singh, et al. (Singh et al. 2023) reviewed the prospective plus points and challenges of big data analytics (BDA) in the healthcare industry. The authors highlighted the increasing adoption of BDA in healthcare while addressing the associated challenges. Although the article was based on an SLR, it offered possible solutions for healthcare challenges.

The reviewed studies are divided into three categories: survey, systematic mapping study (SMS), and SLR, which are depicted in Table 1. Considering the previous points, neither of the SLRs (Failed 2017b; Mikalef et al. 2018; Kolajo et al. 2019; Al-Sai et al. 2020; Rathore et al. 2021; Naghib et al. 2022; Georgiadis and Poels 2022) has reviewed BDPA holistically. Ohiomah, et al. (Failed 2017b) only reviewed articles between 2006 and 2017. Kolajo, et al. (Kolajo et al. 2019) concentrated on big data stream examination covering years of which were between 2004 and 2018. Al-Sai, et al. (Al-Sai et al. 2020) reviewed big data's influential success elements between 2007 and 2019, which appeared to be lacking studies published after 2019; it would have been helpful to include more recent studies to ensure the findings were up to date The only article that is relatively close to our work is Mikalef, et al. (Mikalef et al. 2018), in which the authors only reviewed articles until 2018. Due to this, we state that the SLR that we have presented is the primary one trying to investigate BDPA thoroughly up to the 2023.

3 Research methodology

In the structured progression of this approach, we adhere to a three-step framework outlined as planning, execution, and documentation (Etemadi, et al. 2023; Brereton et al. 2007; Kitchenham et al. 2009), as illustrated in Fig. 2. The assessment is complemented by an external evaluation of the outcomes at each juncture. Initially, we discern the inquiries and motivations underlying this SLR during the planning phase. Subsequently, the selection of pertinent articles within this domain is based on predetermined inclusion/exclusion criteria (see Table 2) during the execution phase. Finally, within the documentation phase, observations are recorded, and the outcomes undergo analysis, comparison, and visualization, culminating in responses to the research queries, followed by the presentation of conclusive reports. This SLR adheres to a three-phase research methodology, detailed subsequently.

3.1 Planning the systematic review

Planning initiates with the identification of the research rationale for this SLR and concludes with the formulation of a review protocol, outlined as follows:

Table 1 Related studies in the field of BDPA

Review type	Ref	Main topic	Publication year	Article selection process	Taxonomy	Future work	Covered Years
Survey	Philip Chen and Zhang 2014)	Big data challenges and chances	2014	Not clear	Yes	Presented	Not mentioned
	Kumaresan and Rajakumar 2015)	Predictive analytics using big data	2015	Not clear	No	Not Presented	Not mentioned
	Banumathi and Aloysius 2017)	Predictive analytic concepts in big data	2017	Not clear	Yes	Presented	Not mentioned
	Poornima and Pushpalatha 2018)	BDPA with data mining	2018	Not clear	No	Not presented	Not mentioned
	Ghani et al. 2019)	Social media big data analytics utilizing machine learning approaches	2019	Clear	Yes	Presented	Not mentioned
	Kaffash et al. 2021)	Big data and intelligent transportation	2021	Clear	No	Presented	1997–2019
	Mallika and Selvamuthukumaran 2022)	Precision in medicine with the scope of big data	2022	Not clear	No	Not presented	Not mentioned
	Nobanee et al. 2022)	Credit risk management with big data applications	2022	Clear	No	Presented	2012–2021
	Ikegwu et al. 2022)	Analyzing big data for industry driven by data	2022	Clear	Yes	Presented	2013–2021
	Himeur, et al. 2023)	BAMSs	2023	Clear	Yes	Presented	2015–2022
SMS	Cichosz et al. 2016)	Big data and predictive models in the management of diabetes	2015	Not Clear	Yes	Presented	1990 – 2015
	O'Donovan et al. 2015)	Set out a review of big data approaches in manufacturing	2015	Clear	Yes	Presented	2009–2015
	Muthukrishnan et al. 2017)	Student's performance predictive analytics using big data	2017	Clear	Yes	Presented	2012 – 2016
	Mehta et al. 2019)	Progression of big data and artificial intelligence for smart healthcare	2019	Clear	Yes	Presented	2013–2019
	Rahman and Reza 2020)	No-functional needs in big data	2020	Not clear	No	Not presented	2012–2019
	Biesialska et al. 2021)	Agile systems with big data analytics	2021	Clear	No	Presented	2011–2019

Table 1 (continued)

Review type	Ref	Main topic	Publication year	Article selection process	Taxonomy	Future work	Covered Years
	Montero et al. 2021)	Quality models in big data	2021	Clear	Yes	Not presented	2010–2022
SLR	Failed 2017b)	BDPA in Information Systems	2017	Clear	Yes	Presented	2006 – 2017
	Mikalef et al. 2018)	The mechanisms through which big data analytics lead to advanced performance	2018	Clear	Yes	Presented	2010–2018
	Kolajo et al. 2019)	Big data stream analysis	2019	Clear	No	Presented	2004–2018
	Al-Sai et al. 2020)	Important success factors for big data	2020	Clear	No	Presented	2007–2019
	Rathore et al. 2021)	The impact of big data, artificial intelligence, and machine learning on digital twinning	2021	Clear	Yes	Presented	2015-Aug 2020
	Naghib et al. 2022)	Big data management	2022	Clear	Yes	Presented	2016–2022
	Georgiadis and Poels 2022)	Data protection rules using big data	2022	Clear	No	Not presented	1995-Jan 2021
	Acciarini et al. 2023)	Organizations can harness the power of big data to bring innovation in their business models	2023	Clear	No	Presented	2014–2022
	Shah et al. 2023)	BDPA and AI in SCRM	2023	Clear	Yes	Presented	2008–2022
	Singh et al. 2023)	BDA in healthcare	2023	Clear	No	Presented	1974–2021
	Our study	BDPA	2024	Clear	Yes	Presented	2014 –2023

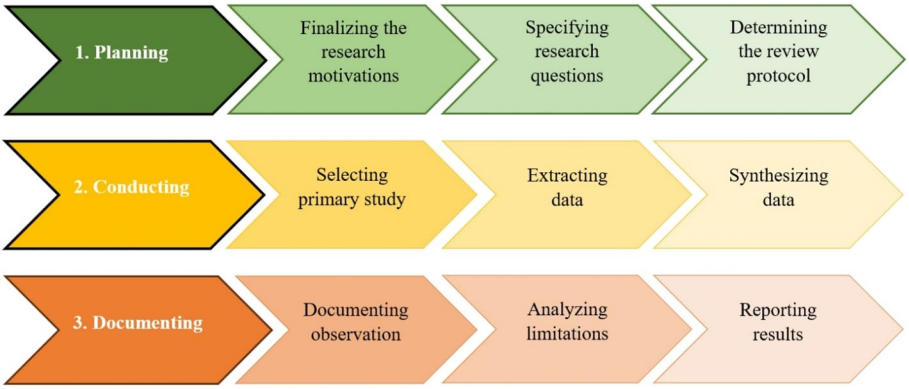


Fig. 2 Overview of research methodology

Table 2 Inclusion/Exclusion criteria

	Criteria	Justification
Inclusion	Research endeavors that concentrate on BDPA Articles published from 2014 to 2023	Gaining a comprehensive understanding of analytical methodologies for big data analytics Recent articles have referenced findings from classical and foundational literature in this particular field
Exclusion	Brief articles with a length of fewer than six pages Surveys and review articles that have not undergone evaluation or original articles are not written in English Chapters from books and academic theses	These studies lack sufficient information for inclusion in our research Due to reservations about the quality of unassessed articles and the inability to scrutinize non-English articles, these documents were excluded The outcomes of book chapters or theses are referenced in both journal and conference articles

Stage 1 - Clarifying the research motivation The initial stage involves specifying the research motivation, determined based on the contribution of this SLR, justified through a comparative analysis of existing reviews. The need for a systematic review leads to the *identification, classification, and comparison* of the recent studies concerned with BDPA. This work mainly focuses on a detailed comparison and classification of big data applications in several areas that are provided in Section 4. To claim that similar literature studies to ours have not yet been conducted, we browsed Google Scholar and famous publishers such as ScienceDirect, Springer, IEEE, ACM, Taylor & Francis, SAGE, World Scientific, Emerald, Wiley, and Hindawi with the following research string.

"Big data" < **AND** >
(survey < **OR** > review < **OR** > overview < **OR** > trends < **OR** > challenges < **OR** > "state of the art" < **OR** > study)

Initial results of the review articles with the above search terms were extracted from Google Scholar based on the titles of the articles. Then, we studied the abstract of the articles that discussed predictive analytics and the topics near it. Consequently, we selected our related works. Finally, we made a comparison of the related works to ours. However, none of the observed reviews mainly answered our proposed research questions (RQs) in Section 3.1. Since BDPA is quite a critical field of study, reinforcing and updating the current evidence on the applications of big data is necessary. So, the legitimate reason that motivated us to propose an SLR is to address all the aforementioned weaknesses. Table 1 presented a summary of the studied surveys that the parameters, such as review types, main topics, publication years, article selection processes, taxonomies, future works, and covered years of each study are depicted. It is clear that just ten articles have used the SLR method, and several articles have not mentioned their article selection processes. In contrast, in this research, our article selection process is completely clear; taxonomy is prepared, future works are explained, and recently published articles up to 2023 are included. Therefore, to fulfill the deficiencies mentioned above, we have conducted a very detailed and thorough study to cover the following demerits:

- Newly published articles, especially from 2020 to 2023, have been missed.
- The structure of most articles is not systematic since the article selection mechanism is not obvious.
- Many articles did not use appropriate classifications.
- The majority of reviews have ignored the evaluation parameters and tools.
- A notable number of previously reviewed articles did not provide a fixed structure and taxonomy.
- Previous articles at maximum reviewed a limited number of articles.

Stage 2—Formulating research questions In the second stage, aligned with the motivation of this article, the research questions are articulated to aid in the development and validation of the review protocol. The subsequent research questions, outlined below, seek to identify gaps in the current understanding of this subject. The resolution of these questions in the documenting phase can unveil new perspectives and ideas.

- RQ1: What are the fields of prediction analysis applications in big data?
- RQ2: What are the evaluation metrics of predictive analytics using big data?
- RQ3: What evaluation methods are used in BDPA?
- RQ4: What are the tools and environments in BDPA?
- RQ5: What are the challenges and future issues of BDPA?

Stage 3—Establishing the review protocol In line with the objectives of this SLR, the preceding stage involved the identification of research questions and the delineation of the review's scope to fine-tune search strings for literature extraction, as per (Brereton et al. 2007). Additionally, a protocol was formulated by drawing inspiration from the approach outlined by Brereton, et al. (Brereton et al. 2007) and our past involvement in SLRs (Bazzaz

Abkenar et al. 2021, 2023; Khoshniat et al. 2023; Songhorabadi et al. 2023, 2011; Kashani and Mahdipour 2023; Nikravan and Haghi Kashani 2022; Sheikh Sofla et al. 2022; Haghi Kashani et al. 2021, 2020; Ahmadi et al. 2021; Rahimi et al. 2020; Nemati et al. 2023; Abkenar et al. 2011; Kashani et al. 2011). To evaluate the formulated protocol prior to its implementation, external expertise was enlisted from a specialist with proficiency in conducting SLRs within this specific domain. The feedback received was incorporated into the refined protocol. A pilot study, covering approximately 20% of the included articles, was conducted to mitigate potential biases among researchers and to optimize the data extraction process. Further refinements were made to the review scope, search strategies, and inclusion/exclusion criteria during this pilot stage.

3.2 Conducting the systematic review

The conducting phase of the research methodology, which commences with the selection of articles and concludes with data extraction, constitutes the second stage. This section is dedicated to illustrating the procedures involved in searching and selecting articles undertaken during the second phase of the SLR.

Stage 1 – Selecting primary articles At this stage, we explored the following search string to collect primary articles:

("big data" < **OR** > "large data" < **OR** > Hadoop < **OR** > Spark < **OR** > Storm)
< **AND** >
(predictive < **OR** > forecasting < **OR** > prediction < **OR** > foresee)

We searched through famous databases, such as Google Scholar, ScienceDirect, Springer, IEEE, ACM, Taylor & Francis, SAGE, World Scientific, Emerald, Wiley, and Hindawi, based on the title, keywords, and the abstraction section of each article.

- *Initial selection:* This stage involves the screening of titles, abstracts, and keywords of potential primary articles. As a result, 1130 articles were primarily recorded from conference articles, journals, book chapters, and notes. The search string was applied to digital databases between 2014 to 2023.
- *Final selection:* After a full-text study all non-English articles, all editorial articles, all book chapters, all working articles, and all short articles, which were less than six pages, were omitted because they could not give us enough information. The extracted number of articles was 109 based on our inclusion and exclusion criteria in Table 2.

Stage 2 and 3—We acquired data from specified online search databases and structured the information based on aspects of characterization, following the guidelines outlined in Kitchenham (2004); Brereton et al. 2007). The scrutiny of these 109 articles forms the basis for our proposed classification of BDPA in Section 4, shedding light on both the advantages and drawbacks of these approaches in Section 5, and presenting future works and open issues in Section 4.6.

4 A classification for applications of big data predictive analytics

The main target of this section is to present a comprehensible trend of predictive analytics by using big data to examine all 109 selected articles. It is not easy to structure the related works on BDPA systematically because the literature is very diverse. The selected articles are classified into seven main groups in this study. According to the article domain, this part is categorized into seven application classes: Industrial, e-commerce, smart healthcare, smart agriculture, smart city, ICT, and weather. These seven categories are widespread among most researchers and authors because they scrutinize the problems and issues from various approaches.

Scrutinizing the articles shows that fifteen parameters exist in the evaluation of the obtained results, and every article might regard one parameter or more. The parameters are as follows:

- *Accuracy*: It refers to the assessment that is applied for setting the most qualified model at identifying relations in a dataset primarily based on input or training data. The following formula (Fawcett 2006) was used:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (N \rightarrow Negative, P \rightarrow Positive, F \rightarrow False, T \rightarrow True) \tag{1}$$

- *Timeliness*: It refers to data accessibility and availability in business decision-making. Clear, well-organized data makes intelligent decisions and leads to a better understanding of future expectations.
- *Cost*: The price that has to be paid totally by the one who requests the service to attain the highest composite service.
- *Scalability*: It refers to the measurement applied to see whether or not the algorithm/framework/platform accommodates fast alternation in data growth.
- *Reliability*: It refers to the probability that a system will be capable of performing a task designed or intended at a specific time and environment.
- *Performance*: The amount of useful work accomplished at a specified time.
- *Validity*: The measurement used to verify the suggested model. This factor tries to see whether or not models are executing as expected and in line with their design purposes and business applications.
- *Resource Utilization*: It refers to the time percentage that a component is occupied in comparison with the total time that the component is available to be applied.
- *Time:* Factors relevant to time, namely processing time, the overall time to provide an output, and performing time.
- *Energy:* The total sum of energy consumed to perform the applied requests.
- *Throughput:* The largest amount of processed data in a system at a particular time.
- *Sustainability:* The capability of the model to be maintained at a certain rate or level without the need for future updates.
- *Feasibility:* The possibility of a statement or model that can be conveniently done.
- *Security:* The degree of being free from threat or avoiding any irreplaceable consequences is generally critical in a smart city and smart healthcare.
- *Precision:* Quality, condition, or fact of being exact and precise while testing a model.

4.1 Industrial applications

We can classify these articles into three sub-classes. The first class has articles discussing business processes (Zhang et al. 2022a; Kong et al. 2022; Yang and Ge 2022; Shafi et al. 2021; Krumeich et al. 2014; Mishra 2019). The second class has articles focusing on the Industrial Internet of things (IIoT) (Failed 2017c, 2018b; Yu et al. 2020; Wang et al. 2020; Tryapkin and Shurova 2020; Lin et al. 2022; Kodidala et al. 2021; Rosati, et al. 2023). The third one deals with articles that talk about issues in supply-chain management (Hazen et al. 2014; Gunasekaran, et al. 2017; Dubey, et al. 2018; Dubey, et al. 2019; Dubey, et al. 2018; Jeble 2018; Nilashi, et al. 2023) and the pros of big data analysis and precise prediction in this domain. In Section 4.1.1, the selected articles are reviewed and then compared in Section 4.1.2.

4.1.1 Overview of industrial articles

Zhang, et al. (Zhang et al. 2022a) took five major factors into account to calculate the emission of carbon in this epoch. First, they calculated the energy infrastructure and energy intensity. Following the industrial structure, they looked up to employee scale and economic prosperity. With these tokens, they could deepen the calculation of emissions on a large scale. Kong, et al. (Kong et al. 2022) analyzed the specifications of the current industry and introduced some preliminaries using machine learning and deep learning. Afterward, they described the definitions of latent variable models (LVMs). Consequently, they advised about the important issues, future directions, and LVMs concepts. Yang and Ge (Yang and Ge 2022) made three contributions to the area of big data analysis in the industrial area. They first summarized the relationships of learning paradigms and then lifelong learning was expressed. Lastly, the authors shed light on the future directions and possible potential to excel the industrial applications with the use of big data.

Shafi, et al. (Shafi et al. 2021) mathematically modeled an IP system. Then, by the usage of proportional-integral-derivative, controlled utilizing of data. Consequently, a storming time series input–output data is driven. With the implementation of current neural network algorithms, the authors were able to control IP systems. The outcome of their research is the trainability of neural networks on random input/output data. Focusing on event-based predictions, Krumeich, et al. (Krumeich et al. 2014) used potentials via predictive analytics on big data to enable proactive control of processes in the business. Therefore, the article simply concentrates on processes in production in the manufacturing industry analytical processes and outlines—based on a case study related to a huge steel company in Germany, Saarstahl AG. In this company, data related to production are gathered to form a foundation to have exact predictions. Nevertheless, this sample cannot use available data potentials for proactive process control without considering big data analytics dedicated approaches.

A model was suggested by Mishra (Mishra 2019) for examining the way the deployment of information technology (IT) (i.e., the partnership of business–BDPA, strategic flexibility of IT, and alignment of business–BDPA) and capabilities of HR makes an influence on OP through BDPA. Structural equation modeling is used on survey data obtained from 159 companies in India in order to test the suggested hypotheses. Based on the results, the diffusion of BDPA would mediate the deployment impact of IT and HR capabilities upon OP. What is more, a direct influence of IT deployment and capabilities of HR upon diffusion of BDPA would be observable, and it also would have a direct relationship with OP. The

authors, in this study, showed that the deployment of IT and the capabilities of HR would influence OP indirectly through the diffusion of BDPA.

An open-source database was designed by Oyekanlu (Failed 2017c), which was properly compacted to be adjusted into edge analytic device memory by applying lightweight database software that is not limited by demanding a server, a client, password schemes, and further requirements similar to traditional database systems. The purpose was to have effective support for real-time computing for IIoT systems to send minimal reports regarding the stages of system conditions to the cloud. This way, by having this lightweight low-memory footprint, the database system at the edge would amplify the whole IIoT system's reliability.

Truong (Failed 2018b) described how they designed and augmented complex IoT and big data cloud systems for integrated analytics of IIoT predictive maintenance. The technique was supposed to locate different complicated interactions for tackling initial errors of the system, which were related to the results of fundamental analytics about the tools. Both Incidents of the system and results of critical analytics were addressed for the equipment integration for maintenance of IIoT predictive.

A manufacturing big data ecosystem was suggested by Yu, et al. (Yu et al. 2020) in order to deal with ingestion issues of big data, management, and analytics to detect faults of predictive maintenance in IoT-based smart factories. Compared to other similar studies that have made efforts to develop different techniques to detect anomalies via simulations, the suggested ECD architecture concentrated on a framework to guarantee the security of data and real-time data analytics by the deployment of a data lake, an encryption protocol, NoSQL database, and the like, on the Apache Spark platform. The MapReduce-based DPCA algorithm was introduced and defined for the fault detection model. In accordance with experimental results, the suggested big data ecosystem has the capability to alarm the system some days ahead of the real occurrence.

The causes and problems of the traditional remanufacturing mode were analyzed by Wang, et al. (Wang et al. 2020). Big data-driven hierarchical digital twin predictive remanufacturing (BDHDTPREMfg) concept and architecture, which was scalable, were suggested, and a big detailed data-driven control mechanism was presented. Afterward, a detailed paradigm scheme implementation in the application of AGV with SESD was presented. Interestingly, the application results validated the efficacy and feasibility of BDHDTPREMfg. Based on the above-mentioned analysis, the advantages of applying BDHDTPREMfg were shown.

The main challenges related to big data analytics were defined by Tryapkin and Shurova (Tryapkin and Shurova 2020). In order to save data storage resources and transportation devices, the best part of data has to be processed on the basis of real-time. On the other hand, a rich toolkit was permitted by the suggested tools to solve the problems of transport, which were associated with identifying the equipment operating modes and the predictive analytic tool deployment to evaluate anomaly of equipment status. Good examples of applying Big Data technologies were provided by this article to evaluate the conditions of existing infrastructures. Different systems of information, various systems of computerized monitoring, and microprocessor systems, which were set up at infrastructural facilities, had been implemented as record sources.

Lin, et al. (Lin et al. 2022) proposed a per-job framework in order to lower the costs and energy of data analysis. In aiming to reduce any possible loss of spot instances irregularities, the authors used a checkpointing mechanism. As for their future works, they have noted that they should optimize the cost when utilizing cloud infrastructures. Kodidala, et al. (Kodidala et al. 2021) tested their model with.NET Micro Systems to see

if they can improve the protection of data and privacy. Their main focus was on improving the security, reliability, and quality of the system. The test findings showed that the defined architecture offers valuable insights into IIoT's intelligent social communities.

Rosati, et al. (Rosati, et al. 2023) introduced a Decision Support System (DSS) for predictive maintenance (PdM) in industrial settings, leveraging IoT, Big Data, and Machine Learning. The DSS addresses the challenge of obtaining quality labeled data by employing a feature extraction strategy and ML prediction model based on specific topics collected from the production system. Experimental results demonstrate that this approach offers a good trade-off between predictive performance and computation effort. The data quality problem was introduced by Hazen, et al. (Hazen et al. 2014) in the supply chain management (SCM) context, and they suggested methods for data quality controlling and monitoring. Additionally, the authors highlighted interdisciplinary research topics on the basis of complementary theories. It was suggested that there is a need for continual improvement in the process of SCM data production and a framework, that is familiar, for the establishment of a quality control mechanism considering data quality. The SPC method application was mentioned but not theory-based topics.

Gunasekaran, et al. (Gunasekaran, et al. 2017) investigated the influence of BDPA assimilation on supply chain planning (SCP) and organizational performance (OP). This study worked on a resource-based view, and the authors considered three stages of assimilation; routinization, assimilation, and acceptance. And, tried to identify resource influence, information sharing, and connectivity, under the commitment influence of great management on big data assimilation capability, OP, and SCP. Based on the findings, information sharing and integration under the commitment influence of great management are relevant to BDPA confirmation positively, which is relevant to BDPA assimilation under BDPA routinization mediation influence and also positively relevant to OP and SCP.

Dubey, et al. (Dubey, et al. 2018) tested the BDPA role in collaborative performance (CP) among those involved in the sustainable development program to attain SCP purpose. The organization fit contingent influence upon BDPA impact on CP was investigated in this study. Variance-based structural equation modeling (PLS-SEM) was taken into account to examine the theory of this study by applying 190 respondents, as samples, who worked in an Indian auto-components manufacturing organization obtained from the ACMA, Bradstreet, and Dun databases. The findings suggested a considerable positive impact of BDPA upon the CP among partners, and it also indicated that resource complementarity and organizational compatibility had a beneficial moderating impact on joining CP and BDPA. There were several limitations to this study. First, the authors collected cross-sectional data. Secondly, it was confined to dyadic networks. In this article, the theoretical structure of the suggested framework was analyzed at the inter-organizational level. However, it was observed from the focal organization outlook only.

BDPA effects on social performance (SP) and environmental performance (EP) were investigated by Dubey, et al. (Dubey, et al. 2019) empirically by applying equation modeling of variance-based structural (i.e., PLS). It was found that BDPA had a considerable impact on SP/EP. However, any evidence for the moderating influence of flexible and control orientation in the links between SP/EP and BDPA was not found. The findings suggested a covert understanding of BDPA performance implications and addressed the vital questions of when and how BDPA is able to increase environmental/social sustainability in supply chains. This study collected data at one point in a time which is regarded as a limitation. It also concentrated on perceptions of the manager rather than his actual performance, which is another limitation. We can also refer to the authors' application of DCV

logic to define BDPA adoption and their research sample demographics that might limit the generalizability of findings as other limitations.

Dubey, et al. (Dubey et al. 2018) defined how big data and predictive analytics might refine coordination and visibility in humanitarian supply chains. A research model in a contingent resource-based view was conceptualized by the author. It was suggested that the capabilities of BDPA influenced coordination and visibility being influenced by the swift trust. Based on the results, BDPA has a considerable effect on coordination and visibility, and also, swift trust does not necessarily affect the relations between coordination, visibility, and BDPA. The application of cross-sectional survey data for testing the research speculation seems to be the main limitation of this article.

A theoretical model was developed by Jeble (Jeble 2018) to define the influence of big data and predictive analytics on an organization's maintainable business flourishing aim. This theoretical model was expanded by applying a resource-based view contingency theory, and logic. By applying PLS-SEM (partial least squares- structural equation modeling), the model was tested more. This study contributed a lot to supply chain management literature and operations. Empirically proven results and theory-driven results were provided by them, and then previous studies focusing on single performance measures (i.e., environmental and economic) were extended. The authors tried to answer some questions unsolved before by investigating BDPA's influence on performance measures.

Nilashi, et al. (Nilashi, et al. 2023) focused on a research gap by exploring the influence of BDPA on recycling and waste management in the food industry, particularly its impact on environmental and economic performance. The findings of their research highlighted the importance of employee knowledge and competitive pressure in driving BDPA adoption while emphasizing the significant role of BDPA in enhancing an organization's competitive advantage through improved environmental conditions.

4.1.2 Summary of industrial articles

In almost all articles, attempts have been made by the authors to improve environmental and organizational performance in addition to boosting their prediction accuracy. These improvements will positively affect resource reliability, utilization, and other related metrics. According to the reviewed and discussed industrial articles, the comparison of their specifications has been depicted in Table 3 which are divided into three categories: business processes (Zhang et al. 2022a; Kong et al. 2022; Yang and Ge 2022; Shafi et al. 2021; Krumeich et al. 2014; Mishra 2019); IIoT (Failed 2017c, 2018b; Yu et al. 2020; Wang et al. 2020; Tryapkin and Shurova 2020; Lin et al. 2022; Kodidala et al. 2021); Supply chain (Hazen et al. 2014; Gunasekaran, et al. 2017; Dubey, et al. 2018; Dubey, et al. 2019; Dubey, et al. 2018; Jeble 2018; Nilashi, et al. 2023). Table 3 shows the main ideas, evaluation methods, tools, advantages, and disadvantages of industrial articles. In addition, Table 4 depicts the improvement of different evaluation metrics in each study. These metrics include accuracy, timeliness, cost, scalability, reliability, performance, validity, and resource utilization.

4.2 E-commerce applications

Mainly the focus of these studies is on the commercial application of BDPA. They are classified into three sub-classes. The first category talks about finance and stock data (Failed 2018c; Haitao 2020; Chen 2018a; Saito and Gupta 2022; Han et al. 2023). And the second

Table 3 Comparison of industrial articles

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Business processes	Zhang et al. 2022a)	The calculation of carbon emission in the double carbon age	Prototype	Not mentioned	• High reliability • Low costs	• Not testing the model in a real condition
	Kong et al. 2022)	LVMs in big data	Simulation	Not mentioned	• High reliability • High accuracy	• Low security • Limited types of datasets
	Yang and Ge 2022)	The paradigm of big data analytics	Simulation	Not mentioned	• High accuracy • Low costs	• Not discussing fine-grained differences
	Shafi et al. 2021)	Intelligent neural network approach for assessing industrial applications of big data	Prototype	NARX, MATLAB	• High accuracy • High reliability	• Not testing the model in a real environment
	Krumeich et al. 2014)	Event-based prediction potentials for planning and controlling processes in business	Real testbed	Not mentioned	• High accuracy • High reliability	• Ambiguous suggestion for dedicated analytics
	Mishra 2019)	Capabilities of organizations that enable the diffusion of big data and predictive analytics and functioning as an organization	Real testbed	Not mentioned	• High organizational performance • High reliability	• Longitudinal data is not collected • Focused on limited organizational capabilities
IIoT	Failed 2017c)	Designation of a lightweight and small open-source database that is able to fit into IIoT edge analytic device memory	Design	Python SQLite	• High resource utilization • Low latency • Low cost	• Minimization of the sent data through a network is not done

Table 3 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Failed 2018b)	General Analysis of interactions between humans and software for predictive maintenance of IIoT to define the design of software and aspects of engineering	Prototype	RabbitMQ, Google Functions, RAHYMS	• High accuracy	• Lack of real-world experiments
	Yu et al. 2020)	A big data environment for the implementation of the mistaken identification and diagnosis in predictive maintenance with real industrial collection from wide-ranging global manufacturing plants	Real testbed	Apache Spark, Apache Drill, Apache Hive	• High performance • High scalability	• Limited number of variables for statistical analysis • More sophisticated data cleaning is needed
	Wang et al. 2020)	The remanufacturing paradigm to solve various bottlenecks	Design	Storm, TensorFlow, Hadoop	• High reliability • High accuracy • Low cost	• No implementation of multi-objective dynamic scheduling reconfiguration optimization
	Tryapkin and Shurova 2020)	Possibility of using the big data for the prediction of the voltage level in the middle of the inter-substation zone based on the authentications from a particular area	Real testbed	MATLAB	• High timeliness • High resource utilization	• Low scalability • Low accuracy

Table 3 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Lin et al. 2022)	Big data analysis for reducing costs in IIoT	Simulation	Not mentioned	• Low costs	• Low accuracy
	Kodidala et al. 2021)	IIoT management with big data	Simulation	Apache Spark, Hadoop, .NET Micro System	• High quality • High reliability	• Limited tools used
	Rosati, et al. 2023)	(DSS) for predictive maintenance (PdM)	Simulation	GUI-based data analytic interface	• High performance • High interpretability	• High processing time
Supply chain	Hazen et al. 2014)	Data quality problem in supply chain management	Real testbed	Control Charts Methods (histogram, fishbone diagram.)	• High accuracy • High timeliness • High consistency • High completeness	• No clear solution for enhancing data quality
	Gunasekaran, et al. 2017)	The information-sharing role and commitment of top management on the transformation of the supply chain and performance of a firm	Real testbed	Not mentioned	• High reliability • High validity • High unidimensionality	• Maturity of big data is not considered (sample was homogeneous)
	Dubey, et al. 2018)	BDPA's role in CP among the partners involved in the maintainable evolution of the program to attain SCP target	Real testbed	Not mentioned	• High organizational compatibility • High resource complementarity • High collaborative performance	• The study was confined to dyadic networks only • Perceptions were gathered from one side of a collaborative relationship
	Dubey, et al. 2019)	BDPA effects upon environmental and social performance by applying variance-based structural equation modeling	Design	Not mentioned	• High reliability • High validity	• No longitudinal study • No focus on actual performance

Table 3 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Dubey et al. 2018)	The way big data and predictive analytics are capable of improving coordination in chains of humanitarian supply	Real testbed	Not mentioned	• High coordination • High visibility	• Lack of multi-search methods
	Jeble 2018)	The influence of predictive analytics and big data capability upon the sustainability of the supply chain	Real testbed	Not mentioned	• High environmental performance • High social performance • High economic performance	• Common method bias (CMB) problem • Limited generalizability
	Nilashi, et al. 2023)	BDPA on recycling and waste management	Simulation	PLS-SEM	• low cost • low energy • competitiveness • High performance	• Not tested in a real-world environment

Table 4 Evaluation metrics in industrial articles

Scope	Ref	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Validity	Resource Utilization
Business processes	Zhang et al. 2022a)	✓	✓	✗	✗	✓	✓	✗	✗
	Kong et al. 2022)	✓	✓	✓	✗	✗	✓	✓	✓
	Yang and Ge 2022)	✓	✓	✓	✗	✗	✗	✓	✓
	Shafi et al. 2021)	✓	✓	✓	✓	✓	✓	✗	✓
	Krumeich et al. 2014)	✓	✗	✗	✓	✓	✗	✗	✗
	Mishra 2019)	✓	✗	✓	✗	✓	✓	✓	✗
IIoT	Failed 2017c)	✗	✓	✓	✗	✓	✗	✗	✓
	Failed 2018b)	✓	✗	✗	✓	✗	✗	✗	✓
	Yu et al. 2020)	✗	✗	✗	✓	✓	✓	✓	✗
	Wang et al. 2020)	✓	✗	✓	✓	✓	✓	✓	✗
	Tryapkin and Shurova 2020)	✗	✓	✗	✗	✗	✗	✗	✓
	Lin et al. 2022)	✓	✓	✓	✗	✓	✗	✓	✗
	Kodidala et al. 2021)	✓	✓	✓	✗	✓	✗	✗	✓
	Rosati, et al. 2023)	✗	✓	✓	✗	✗	✓	✓	✓
Supply chain	Hazen et al. 2014)	✓	✓	✗	✗	✗	✓	✗	✗
	Gunasekaran, et al. 2017)	✓	✗	✓	✗	✓	✗	✓	✗
	Dubey, et al. 2018)	✗	✓	✗	✓	✓	✓	✓	✗
	Dubey, et al. 2019)	✓	✗	✗	✗	✓	✓	✓	✗
	Dubey et al. 2018)	✓	✓	✗	✗	✓	✓	✓	✗
	Jeble 2018)	✓	✓	✗	✓	✗	✓	✓	✗
	Nilashi, et al. 2023)	✗	✗	✓	✗	✗	✓	✗	✓

class has articles discussing retail (Bradlow et al. 2017; Failed 2014; Lee 2017; Sasidhar and Mallikharjuna Rao 2019; Zheng et al. 2020; Chen 2021; Zhuang 2021; Alrumiah and Hadwan 2021; Li and Li 2022). The third class takes technology into account in the area of e-commerce: (Suguna et al. 2016; Failed 2015a; Xiao 2022). Almost all studies are related to different big data analytic solutions and retail to predict customers' behaviors. In Section 4.2, the selected e-commerce articles are reviewed and then compared in Section 4.2.2.

4.2.1 Overview of e-commerce articles

An algorithm of machine learning, a token-based ensemble specifically, was suggested by Morris, et al. (Failed 2018c), which used nonlinear and linear estimators for predicting big financial time-series data. The ensemble is composed of a long/short-term memory (LSTM) network, a traditional Kalman filter, and a traditional linear regression model. They found the adaptive features in short-term, high-risk trading when noisy data, like stock prices, were present and showed their ensemble the performance. In (Haitao 2020), the expanding situation of the targeted e-commerce supply chain management information system was proposed. The evaluation results indicated a large-scale reduction of costs and enhancements in the system's efficiency. However, the precision of the healthy development of the network loans should be investigated by both governments and industries.

Chen (Chen 2018a) produced a structural personalization regulation of e-commerce and text-matching algorithms. He stated that, by using the algorithms, not only the chances of the transaction increased but also the level of personalized service was elevated. However, the method lost its precision when it was tested in a commodity search. Saito and Gupta (Saito and Gupta 2022) came up with a quantitative model researching the impacts of social media on finances. They tested three distinct models, namely a revenue manager model, high-frequency trading equity, and interest rate framework. However, their models had some deficiencies as COVID-19 had lowered the travel frequency of people, in which affected one of their models in a hotel as they could not find adequate up-to-date data. Han, et al. (Han et al. 2023) proposed a method for risk analysis and supervision in internet finance using big data analysis, establishing a financial risk assessment model. The experiments conducted confirm the effectiveness of the approach in addressing the limitations of traditional models.

The opportunities in big data and also the possibilities resulting from big data in retailing were examined by Bradlow, et al. (Bradlow et al. 2017), especially along five main data dimensions—data related to channels, time customers, products, and geospatial locations. The rise in the quality of data and possibilities of application is the result of a combination of new sources of data -an intelligent application of domain knowledge and statistical tools mixed with theoretical insights. It is important that a theory can guide a systematic search to answer the retailing questions and can also streamline the analysis, and at the same time, remain intact. Big data and predictive analytic roles in repetition are becoming important since they are assisted by data sources and large-scale interconnected methods. All statistical issues, which were discussed in this study, concentrated on applications and relevance of Bayesian analysis techniques; data borrowing, hierarchical modeling, augmentation, and updating, a field experiment, and predictive analytics by applying big data in a retailing context.

A retail recommender model, which was based on a cooperating filter, was suggested by Sun, et al. (Failed 2014). They also designed an algorithm of corresponding distributed computing on MapReduce to execute a big data-based retail recommender system. This

big data procedure assisted the system in performing scalable data processing effortlessly. According to the outcomes of the experiment, the system was effective in estimating the retail sales for every product and store; therefore, this innovative method of precision marketing support benefited non-e-commerce enterprises.

In order to back anticipatory shipping, Lee (Lee 2017) suggested a model of genetic algorithm (GA)-based optimization. On the other hand, cloud computing was applied to store the big data that was generated from every channel. To predict the purchases of the future, based on If–Then prediction rules, and to understand the patterns of the purchases, cluster-based association rule mining was used. Afterward, a modified GA was applied to produce optimal anticipatory shipping plans. Besides the costs of transportation and shipping distance, in GA, the confidence of prediction rules was regarded. A huge number of numerical experiments were performed to illustrate the interchange between various elements in shipping, and the model optimization authentication was confirmed.

In (Sasidhar and Mallikharjuna Rao 2019), a model was presented to depict the widening area of cloud computing with big data in the retail market. The authors introduced a model that connects big data as a service with the cloud for the realistic measurement of a customer's behavior. However, the model's scalability suffered from narrowness, and there is still room for further studies to examine real data trends and network congestion of e-commerce retailers. Zheng, et al. (Zheng et al. 2020) presented a model using two methods, namely the analytic hierarchy process and the technique for order preference by similarity to an ideal solution to investigate logical distribution modes for the stores at JD. Their employed technique attracted the outcomes of subjective analysis, simultaneously giving full play to the advantages of measurable analysis.

Chen (Chen 2021) proposed an e-commerce method in accordance with the technology of network data, researching the game balance of four members of the supply chain to lower the cost and enhance the efficiency of customization. The author applied four elements of game equilibrium of supply: centralized decision-making and decentralized decision-making, C2B-dominated decision-making, and traditional enterprise-dominated business utilization. Zhuang (Zhuang 2021) analyzed the impacts of big data on e-commerce in the U.S. and China. The two major databases that the author used were the Web of Science and CNKI. He mostly tried to clarify any doubts about the development of big data on e-commerce in the noted nations. The author, however, only limited his case study to two countries.

Alrumiah and Hadwan (Alrumiah and Hadwan 2021) searched the vendors and also the customer's views on e-commerce. They reached the outcome that e-commerce has some negative affections on the customers, such as addiction and it is costly for vendors to be able to take advantage of big data analytics tools. Li and Li (Li and Li 2022) conducted research concerning big data mining tools via cellphone applications for e-commerce. They divided their work into two sides; theoretical and experimental. In the experimental sector, the authors came up with the appreciation of customers through promotional activities. And, e-commerce provides very convenient and high-quality atmospheres for both the buyers and sellers.

Suguna, et al. (Suguna et al. 2016) discussed the criticality of log files in e-commerce by analyzing the log files which were used for identifying the user's actions. Their employed model depicted how to process log files using MapReduce and how to utilize the Hadoop framework for parallel computation of log files. The author's approach decreased the response time, added to the functionality of the model, and provided accurate results in the appropriate mean of the response time. Aboutorabi[a], et al. (Failed 2015a) concentrated on variances between two NoSQL databases; MongoDB and Microsoft SQL Server. The

author's study bolded the big variances between the noted database management systems regarding the productivity of processing the queries. MongoDB produced better performance, flexibility, and reliability; however, there is a need to test the model in the real-world environment to have a more precise evaluation of the application of MongoDB and SQL. Xiao (Xiao 2022) mostly focused on big data killing occurrences in the e-commerce emergence. He modeled a four-party game model, which was evolutionary of the government departments, e-commerce companies and platforms, and the consumers. Although the study took big data killing into account in a detailed way, the author states that there is still room for perfection in this field as the four-party evolutionary model would not cover all the circumstances.

4.2.2 Summary of e-commerce articles

Most studies make efforts to reduce the cost and improve prediction accuracy. Such different tools as Hadoop and XLMiner have been applied. According to the reviewed and discussed e-commerce articles, the comparison of their specifications has been depicted in Table 5. Table 5 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each e-commerce article. In addition, Table 6 displays the improvement of different evaluation metrics in each study. These metrics include accuracy, timeliness, cost, scalability, reliability, performance, validity, resource utilization, time, and throughput.

4.3 Smart healthcare applications

These studies focus on predictions in order to help and protect individuals against diseases with the assistance of big data, which has been classified into two distinct groups. The first category discusses the methods for the prediction of disease: (Hendri and Sulaiman 2018; Venkatesh et al. 2019; Khatibi et al. 2019; Failed 2018d, 2017d, 2020; Souza et al. 2020; Gaedke Nomura et al. 2021; Safa et al. 2023). The second category focuses on the economic prosperity of smart healthcare (Weerakkody et al. 2021; Nallathamby et al. 2021; Chen 2018b; Awotunde et al. 2022; Ali et al. 2022; Das and Namasudra 2022; Zang and You 2022; Babar et al. 2022). In Section 4.2.2, the selected smart healthcare articles were reviewed. Finally, in Section 4.3.2, the discussed articles are compared and summarized.

4.3.1 Overview of smart healthcare articles

Findings from Malaysian healthcare facilities were reported by Hendri and Sulaiman (Hendri and Sulaiman 2018), who recorded 9,261 dengue patients in 2014. The research aimed to procure descriptive analysis and suggested techniques of big data analytic modeling to predict and define dengue patients' length of stay (LoS). Such demographic data as the date of discharge, age, admission, and gender have been considered as factors contributing to LoS prediction.

Naive Bayes (NB) machine learning Technique was proposed by Venkatesh, et al. (Venkatesh et al. 2019) for forecasting heart failure, which procures high accuracy. The heart disease data, obtained from the UCI machine learning repository, was trained by The Naive Bayes approach. This approach, then, predicted the test data in order to predict the classification. The suggested BPANB scheme applied Hadoop-Spark as a big data computing tool to attain important insights into healthcare data. To predict the future health conditions of various patients, experiments were conducted. The training

Table 5 Comparison of e-commerce articles

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Finance	Failed 2018c)	Token-based ensemble algorithm which uses both estimators, nonlinear and optimal linear, for predicting big financial time-series data	Simulation	LSTM, Kalman filter, and linear regression estimators	• High accuracy	• Selection of the best prediction algorithm is not efficient
	Haitao 2020)	The development situation of the objectivity of e-business supply chain management information systems	Real testbed	Not mentioned	• Low costs • High efficiency	• High technical and operational risks • Low accuracy
	Chen 2018a)	A personalized e-commerce recommendation system based on big data analysis	Real testbed	Not mentioned	• High accuracy	• Low scalability
	Saito and Gupta 2022)	Mathematical models and social media role in the management of finances	Formal	Not mentioned	• High accuracy • High validity	• Low scalability
	Han et al. 2023)	A financial risk assessment model utilizing BDA	Simulation	MATLAB	• High effectiveness • High performance	• Was not tested in a real-world case
Retail	Bradlow et al. 2017)	Big data role and predictive analytic role in retailing	Simulation	Not mentioned	• High accuracy	• Ethical and privacy issues should be considered
	Failed 2014)	Application of the recommender system technology in the non-e-commerce companies	Real testbed	Hadoop	• High scalability • High validity	• Performance is not considered

Table 5 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Lee 2017)	Propose an optimization model for regulating the allocation of products to different DCs based on multiple factors	Real testbed	Cloud computing, XLMiner	• High reliability • High timeliness	• Assumptions that made the model less complicated • Clusters were pre-defined according to geographical criterion
	Sasidhar and Mallikharjuna Rao 2019)	Unification of big data and public cloud	Prototype	Hadoop, Cloudera	• Low latency • Inexpensive and flexible storage • High throughput • High consistency	• Data tends were not realistic • Low scalability
	Zheng et al. 2020)	An approach for e-commerce companies for selecting logistics distribution modes based on big data	Simulation	Not mentioned	• Low costs • Low execution time • High efficiency	• To improve the timeliness, the system should be updated regularly
	Chen 2021)	The e-commerce approach according to methods of network data via researching the game balance	Simulation	Game theory tools, SQL	• Low costs • High efficiency	• Accuracy and precision were neglected
	Zhuang 2021)	Big data analytics role in e-commerce	Real testbed	CNKI	• High efficiency • Low costs	• Limited case studies
	Alrumiah and Hadwan 2021)	Big data; customers and vendors' outlook	Simulation	Not mentioned	• High efficiency	• High costs • Low security
	Li and Li 2022)	E-commerce with big data mining methods	Simulation	Not mentioned	• High accuracy • High quality • Easy to use	• Low scalability • Low security

Table 5 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Technology	Suguna et al. 2016)	Proposes a prediction for a prefetching system based on the processing of weblogs that takes Hadoop MapReduce into account	Simulation	Hadoop, MapReduce, HDFS	• High accuracy • Low execution time • High efficiency	• Not using correlation engines, such as HA cloud for more precise and scalable output
	Failed 2015a)	Evaluating the document-oriented MongoDB database with SQL	Simulation	MongoDB, SQL	• High scalability • High flexibility • High performance	• No real-time distributed environment for the evaluation of performance
	Xiao 2022)	Big data killing—price discrimination	Simulation	MATLAB2021	• High performance • High flexibility	• Not covering all the existing samples

Table 6 Evaluation metrics in e-commerce articles

Scope	Ref	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Validity	Resource Utilization	Time	Throughput
Finance	Failed 2018c)	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗
	Haitao 2020)	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
	Chen 2018a)	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓
	Saito and Gupta 2022)	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗
	Han et al. 2023)	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓
Retail	Bradlow et al. 2017)	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗
	Failed 2014)	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
	Lee 2017)	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗
	Sasidhar and Mallikharjuna Rao 2019)	✗	✗	✓	✗	✓	✗	✗	✗	✓	✓
	Zheng et al. 2020)	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗
	Chen 2021)	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓
	Zhuang 2021)	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗
	Alrumiah and Hadwan 2021)	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗
	Li and Li 2022)	✓	✗	✓	✓	✓	✗	✗	✓	✗	✓
Technology	Suguna et al. 2016)	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓
	Failed 2015a)	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗
	Xiao 2022)	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓

dataset estimated the health parameters that were required for classification, and the results defined the early detection of the disease to find out the future health of the patients.

To predict premature births and rank predictive features, Khatibi, et al. (Khatibi et al. 2019) suggested machine learning models for big data analytics. This model is capable of predicting premature births with 81% accuracy and 68% AUC. Findings suggested that pregnancy risk segments, gestational diabetes, heart-related problems, mother's age, underlying maternal diseases, the number of pregnancies, education level, prenatal gender, and city were high-ranked predictive features. To reduce the risk of premature births, there was a suggestion to manage and monitor the high-ranked features remotely with applications of smartphones and IoT gadgets at regular intervals.

An architecture of big-data-based predictive maintenance was suggested by Çoban, et al. (Failed 2018d) for biomedical devices in the medical domain. The data categorization was obtained from the alarms of devices, biomedical devices, and their health conditions in a real-time style; the suggested architecture made data management in this foreseeable maintenance model possible. To increase biomedical device performances, enhance the reliance on these devices, and reduce the employees' accidents that healthcare staff encounter, a model of predictive maintenance in big data has to be applied.

A clinical decision scheme based on RNN, called PCD, was suggested by Lin, et al. (Failed 2017d) to alert and predict disease before it happened while considering the patient's privacy. A homomorphic encryption scheme was designed by the authors not to disclose users' and data providers' privacy, so PCD could resist different security intimidations. In order to guarantee the improvement of accuracy prediction, the authors designed an averaged and sequential RNN model in real-time systems. As it can be seen from the experimental results, PCD could attain high efficiency in time and high accuracy in disease forecasting and at the same time keep the patients' and data providers' privacy. Souza, et al. (Souza et al. 2020) presented a framework that used three kernels, namely linear, polynomial, and RBF in a support vector machine (SVM) ensemble to forecast dengue cases. They illustrated the effectiveness and functionality of the proposed framework by conducting numerous case studies. Neither flexibility nor performance was taken into account in their approach. Nevertheless, their proposed method was relatively precise.

Gaedke Nomura, et al. (Gaedke Nomura et al. 2021) described the usage of a big data science framework to extend a pain information model and argued the prospects for its use in predictive modeling. Data for the proposed framework was extracted from a hospital. The framework was intended to customize remedies, improve health outcomes, and reduce costs. Christobel and Kamalakannan (Failed 2020) discussed the effects of diabetes—type 1, type 2, and GDM). The proposed framework employed the big PIMA dataset. Their presented algorithm used Hadoop/MapReduce, which helped the disease prediction and supported the production reports. It was observed that their method provided high performance.

Safa, et al. (Safa et al. 2023) proposed a healthcare big data analytics model (HCBDA) for disease prediction. The HCBDA model utilizes wireless sensor networks and IoT devices to monitor patients' biosignals and generate medical assistance. The proposed approach achieved a disease prediction accuracy of up to 96% and employed machine learning algorithms for classification and recommendation generation. The HCBDA paradigm clusters large healthcare data and utilizes a decision support system for analysis and prediction. Weerakkody, et al. (Weerakkody et al. 2021) employed a model for improving subject well-being. The authors used open data available from the national annual population survey, which limited the accessibility to detailed and personal data. However, the

presented model depicted that the data can be continuously collected without additional and exorbitant costs.

Nallathamby, et al. (Nallathamby et al. 2021) introduced a model to create an effective appointment-scheduling platform for outpatients. In their model, Hadoop and MapReduce were used, which led to low costs in healthcare and excellent flexibility for the presented model. Chen (Chen 2018b) established a systematic pathway to enable big data to become applicable in the medical sector by extracting data and analyzing them in a net relation map. The author created a framework of Taiwan's' healthcare industry to propose his methods and conduct the research. Awotunde, et al. (Awotunde et al. 2022) proposed a nonstop monitoring device with the usage of IoT health surveillance to measure the body's temperature, blood glucose, blood pressure, and other elements affecting the patient's overall health status. Ali, et al. (Ali et al. 2022) came up with a framework simulating the security release and sensitive data about the patients using online time slots to have an appointment. They filtered their data NoSQL and Redis cache to increase their accuracy while improving security.

Das and Namasudra (Das and Namasudra 2022) presented a scheme to make smart health care in an IoT-enabled environment more confidential and secure so as to only allow the authenticated user to gain access to a database. They analyzed the security and performance of the scheme, which satisfied the requirements, but the overall overhead of the model was a burden. Zang and You (Zang and You 2022) proposed a framework with three distinct modules, namely data preparation, model training, and data computation for evaluating a real-time efficiency process. They used machine learning techniques, such as regression and decision tree to train their model. Babar, et al. (Babar et al. 2022) developed a scheme with a pre-processing data module, data processing, and data ingestion module using Spark to analyze data properly. The reason that they used pre-processing data was to level up and speed up the real-time data processing time. However, their scheme had a lack of well-organized parallel data processing.

4.3.2 Summary of smart healthcare articles

These studies would like to focus on boosting performance, making more accurate predictions, and limiting costs. According to the reviewed and discussed healthcare articles, the comparison of their specifications has been depicted in Table 7. Table 7 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each healthcare article. In addition, Table 8 displays the improvement of different evaluation metrics in each study. These metrics include accuracy, timeliness, cost, scalability, reliability, performance, validity, resource utilization, time, precision, and energy.

4.4 Smart agriculture applications

These articles specifically aimed at enhancing the quality of various parameters in the field of smart agriculture, and they are divided into three sub-classes. The first group's articles zoom in on crop prediction (Tsouli Fathi et al. 2020; Velmurugan et al. 2021). The second group of articles takes a detailed look into precision agriculture (Bendre et al. 2016; Sabarina and Priya 2015; Keswani et al. 2020; Failed 2015b; Melgar-García, et al. 2022; Wang and Mu 2022). And, the third one discusses smart farming methods (Liu 2021; Roukh et al. 2020; Li and Niu 2020; Osinga et al. 2022; Shrivastava et al. 2023). In Section 4.3.1, the selected agriculture articles are reviewed and then compared in Section 4.4.2.

Table 7 Comparison of smart healthcare articles

Scope	Ref	Main idea	Evaluation method	**Tool**	Advantage	Disadvantage
Disease prediction	Hendri and Sulaiman 2018)	Recognition of big data analytics use for determining and predicting LoS of dengue patients from the electronic medical records of chosen Malaysian Healthcare premises	Real testbed	RapidMinder	• High accuracy	• Simplicity of the modeling technique
	Venkatesh et al. 2019)	Applying Naïve Bayes machine learning technique for predicting heart failure	Simulation	Apache Spark, Hadoop	• High accuracy • High resource utilization	• Limited to a homogeneous environment
	Khatibi et al. 2019)	Proposing some techniques for the prediction of provider-initiated preterm births and unplanned premature deliveries and ranking predictive features	Simulation	Not mentioned	• High accuracy • Low cost • High Performance	• Exclusion of some features for preterm birth occurrences • Resource Utilization was not considered
	Failed 2018d)	A scalable predictive maintenance architecture for the smart healthcare domain	Simulation	Apache Kafka, HDFS, Apache Storm, Apache Flink	• High reliability • High Scalability	• Not applied on a real-world use-case
	Failed 2017d)	A clinical decision scheme based on RNN, that could forecast and send warnings before diseases occur	Simulation	Amazon EC2 cloud, A java simulator	• High timeliness • Low Cost	• Scalability is not considered

Table 7 (continued)

Scope	Ref	Main idea	Evaluation method	**Tool**	Advantage	Disadvantage
	Souza et al. 2020)	A big data predictive analytic over hybrid sources	Simulation	SVM, Kernel	• High precision • High efficiency • High resource utilization	• Not flexible paradigms • Not considering privacy and performance
	Gaedke Nomura et al. 2021)	Developing a pain information model and discussing the potential for its use in predictive modeling	Real testbed	CRISP-DM, PostgreSQL, Python	• High recall • High security • Low costs	• The process of gathering data was time-consuming
	Failed 2020)	Discussing the impact of different types of diabetes	Simulation	Hadoop HDFS MapReduce	• High performance • High precision • High accuracy	• High execution time
	Safa et al. 2023)	HCBDA for disease prediction	Simulation	Python, DevKit, IOT-LDA	• High accuracy	• Low scalability
Income analysis	Weerakkody et al. 2021)	Researching the impact of big data by recognizing how it can be leveraged to influence well-being	Real testbed	Not mentioned	• High performance • Low costs • High efficiency	• Low scalability • Not using personal and detailed data • Low precision
	Nallathamby et al. 2021)	An effective appointment scheduling platform for outpatients	Real testbed	Hadoop, Hive, Python, MapReduce	• High efficiency • High speed • High flexibility • Low costs	• Not taking into account joining algorithm for better performance • Not having high coefficient clustering
	Chen 2018b)	Big data applications in the medical sector	Formal Now = Simulation	DEMATEL	• High scalability	• Low reliability • Low resource allocation

Table 7 (continued)

Scope	Ref	Main idea	Evaluation method	**Tool**	Advantage	Disadvantage
	Awotunde et al. 2022)	Smart healthcare analysis with the help of IoT monitoring and big data	Simulation	Hadoop	• Low costs • High efficiency	• Low confidentiality • Low security
	Ali et al. 2022)	Privacy-preserving in smart healthcare	Simulation	NoSQL, Redis	• High security	• Not testing the model in a real condition
	Das and Namasudra 2022)	Privacy-preserving in smart healthcare with a password mechanism	Simulation	AVISPA	• High security • Low costs	• The overall overhead of the proposed scheme was limited
	Zang and You 2022)	IIoT-enabled scheme for smart healthcare with the usage of big data and machine learning	Simulation	Machine learning algorithms, Spark	• Low costs • High efficiency	• lack of proper and accurate prediction
	Babar et al. 2022)	IoT-enabled environment helping teledentistry healthcare	Simulation	Hadoop, Apache Spark	• High efficiency	• Low parallel data loading • Low reliability

Table 8 Evaluation metrics in smart healthcare articles

Scope	Ref	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Validity	Resource Utilization	Time	precision	Energy
Disease prediction	Hendri and Sulaiman 2018)	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
	Venkatesh et al. 2019)	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗
	Khatibi et al. 2019)	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
	Failed 2018d)	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗
	Failed 2017d)	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
	Souza et al. 2020)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗
	Gaedke Nomura et al. 2021)	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓
	Failed 2020)	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗
	Safa et al. 2023)	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗
Income analysis	Weerakkody et al. 2021)	✓	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓
	Nallathamby et al. 2021)	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓
	Chen 2018b)	✓	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗
	Awotunde et al. 2022)	✓	✗	✓	✗	✓	✗	✓	✗	✓	✓	✓
	Ali et al. 2022)	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗
	Das and Namasudra 2022)	✓	✓	✗	✓	✓	✗	✓	✗	✗	✓	✗
	Zang and You 2022)	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗	✓
	Babar et al. 2022)	✓	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓

4.4.1 Overview of smart agriculture articles

According to a network of neurons, Tsouli Fathi, et al. (Tsouli Fathi et al. 2020) offered an algorithm to forecast the alterations in the climatic conditions influenced by the yields and production of crops in agriculture for the future and a defined area. The proposed ANN-based approach performance experimented with a 30-year meteorological dataset composed of 54,000 records having such features as humidity temperature, rainfall, wind velocity, and some agro-climatic data obtained from the climate rules, Köppen classification. The inaccuracy of prediction was very low, and the learning convergence was robust. According to predictive data mining, this study would discover the possibility of extracting useful knowledge and patterns out of a considerable amount of agro-climatic and meteorological data.

Velmurugan, et al. (Velmurugan et al. 2021) introduced a technique for consumption and processing strategies for evaluating the dataset. A highly technological methodology, fuzzy enumeration crop prediction algorithm (FECPA), was used for precise cultivation. The resources used in the FECPA system simulate Jupyter notebooks were shared with appropriate datasets and depicted the availability of fast branch-and-bound, naive Bayes, conventional neural network, and the presented system FECPA. An abstract idea was suggested by Bendre, et al. (Bendre et al. 2016), about big data in precision agriculture and about the manner of discovering the insights from big precision agriculture data via ICT resources for farming in the future. The authors also suggested an e-agriculture model that applies the services of ICT in the agricultural environment to collect big data. This article sorted out various big data sources in the precision ICT-based e-agriculture model, its challenges, and its future applications. In the end, they discussed the application of rainfall prediction by applying the unsupervised and supervised method to process and forecast data.

Sabarina and Priya (Sabarina and Priya 2015) focused on the manner of big data reduction size systematically by using a model of tensor-based feature reduction in favor of precision agriculture. With the help of the IHOSVD algorithm, the decomposition of data and the extraction of core values were done. Consequently, the total file size was reduced by deleting unnecessary data dimensions. The time spent on CPU usage and data analysis would be markedly decreased when dimensionality-reduced data were applied instead of raw unprocessed data. Keswani, et al. (Keswani et al. 2020) presented a real-time decision support system (DSS) to generate adequate valve control commands. Six different sensors were proposed to test the prediction techniques, such as deep neural networks (DNN) and random forests to forecast soil moisture content. It can be inferred that the DNN model was appropriate for a prediction-based smart irrigation scheme because DNN is an artificial neural network model including many hidden layers between inputs and outputs.

Bendre, et al. (Failed 2015b) employed an approach regarding unearthing extra perception from precise agriculture data by using big data. Their major aim in this article was to level up the accuracy of forecasting different weather parameters for future precision in agricultural areas. The outcomes are predicted using a regression model, and big data is handled by MapReduce. Melgar-García, et al. (Melgar-García, et al. 2022) took a triclustering approach in a field located in Portugal to examine the precision in agriculture. The main metrics that they tried to emphasize were scalability, performance, and reliability. However, they could not manage the anomalies with their proposed framework and algorithms.

Wang and Mu (Wang and Mu 2022) focused on big data management and risk monitoring in precision agriculture. They created an IoT-based and big data-based platform in

order to analyze how accurately the light wavelength would be transformed and received. Nevertheless, a more comprehensive study is needed to cover all types of plants and species as the authors only took wheat landfields into account. Liu (Liu 2021) used ZigBee wireless sensor network to cover all aspects of crops under the instruction of the concept of efficient agricultural technologies. The author initially introduced the multi-generation genetic algorithm back propagation model in the first layer of his model. Then, in the application layer, the hierarchical analytic process was proposed as the guidance mechanism of neural networks. In the end, he used mathematical statistics to test the presented model.

Roukh, et al. (Roukh et al. 2020) investigated the state-of-the-art platforms and big data architecture before proposing their solution. The authors introduced an overall big data architecture for smart farming. Their presented framework used the approach of Lambda architecture to address the issues of acquisition, execution and storing real-time big data. They tested their model in a real-time environment; however, their model was not tested in various agricultural lands to obtain more precise results. Li and Niu (Li and Niu 2020) used the K-means algorithm based on the farthest distance to research data mining in the agricultural production process. The tested outcomes illustrated that the improved K-means clustering method was constituted for the reduction of total time. The presented algorithm can be used in real-time data and does not lose its efficiency. However, the value of a large amount of data was not fully comprehended because of insufficient information to control the mining.

Osinga, et al. (Osinga et al. 2022) took a survey (n = 56) from stakeholders regarding livestock and fishing to food security. The method that they used was a mixed approach to conduct their study and they mainly focused on four perspectives, namely the elements changing the initiated drivers, distinguishing big data methods, the maturity status of technology, and the stakeholder's view. The authors in Shrivastava et al. (2023) adopted smart farming concepts, such as hydroponics with IoT platforms by eliminating the need for soil and optimizing resources. This vertical hydroponic system, aided by IoT sensors, allows continuous monitoring of crop health and supply of nutrients and water, resulting in increased productivity and reduced costs. The research article focused on the design and implementation of this automated vertical hydroponic farming method.

4.4.2 Summary of smart agriculture articles

This class of articles made use of their innovations in real-world cases and majorly concentrated on enhancing performance. According to the reviewed and discussed smart agriculture articles, they are divided into three categories, which have been depicted in Table 9. Table 9 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each smart agriculture article. In addition, Table 10 displays the improvement of different evaluation metrics in each study. These metrics include accuracy, timeliness, cost, scalability, reliability, performance, validity, and resource utilization.

4.5 Smart city applications

These articles particularly targeted big data for public transit to deter delays of public transportation, and improve the accuracy and changes in future intelligent transportation in smart cities. Due to this, we have arranged articles into five sub-classes. The first class has articles regarding vehicles in smart city (Balbin et al. 2020; Cui et al. 2019; Guo and Xu 2022). The second sub-class collects articles on city planning: (Failed 2017e, 2022;

Table 9 Comparison of smart agriculture articles

Scope	Ref	Main idea	Evaluation method	**Tool**	Advantage	Disadvantage
Crop prediction	Tsouli Fathi et al. 2020)	Appropriate use of the data mining methods for meteorological and agricultural data to help in the development of agriculture	Real testbed	Spark	• High accuracy	• Resource utilization and reliability are not considered
	Velmurugan et al. 2021)	To boost crop cultivation by using big data to create effective programs for data classification	Simulation	Python, Jupyter notebook	• High accuracy • High performance	• Scalability and cost are not taken into account
Precision agriculture	Bendre et al. 2016)	Big data in precision agriculture and how it discovers insights from big precision agriculture data	Real testbed	Not mentioned	• Low cost • High validity	• Scalability and timeliness are not considered
	Sabarina and Priya 2015)	How to systematically diminish the size of big agricultural data by applying a tensor-based feature reduction model	Real testbed	Not mentioned	• High performance	• Accuracy is not considered
	Keswani et al. 2020)	The efficiency control of farm irrigation by exploiting the abilities of the IoT and big data-based DSS to produce enough control commands	Real testbed	Hadoop, HDFS	• Low costs • accurate prediction • High efficiency	• Dependent on battery life • Scalability and timeliness are not considered

Table 9 (continued)

Scope	Ref	Main idea	Evaluation method	**Tool**	Advantage	Disadvantage
	Failed 2015b)	The usage of ICT services in agricultural big data	Simulation	MapReduce	• Low costs • High accuracy • High performance	• Low resource utilization
	Melgar-García, et al. 2022)	Precision agriculture using try clustering approach	Real testbed	BigTriGen, Evolutionary algorithms	• High precision • High efficiency	• Not detecting anomalies with algorithms
	Wang and Mu 2022)	Risk monitoring	Simulation	Not mentioned	• High precision	• Not testing the model in a real condition
Smart farming	Liu 2021)	Leveling up the operating outcome and practical impacts of the smart agricultural system	Simulation	Not mentioned	• High precision • High performance • High accuracy	• Time and cost are not considered
	Roukh et al. 2020)	Introduced a cloud-based intelligent farming management platform called Wall smart	Real testbed	Kafka, Strom, Hadoop,	• High productivity • High sustainability	• Low scalability • The user interface was not optimum
	Li and Niu 2020)	Studying the data storage, data processing, and data mining of big data sources in the agricultural production process	Simulation	MATLAB	• High performance • Low execution time	• Not using Neural network to predict the outcome
	Osinga et al. 2022)	Agriculture uses big data	Simulation	Not mentioned	• Low costs	• Big data cannot be used in the context of agriculture out-of-the-box
	Shrivastava et al. 2023)	Automated vertical hydroponic farming method	Design	Not mentioned	• Low costs • High effectiveness • Low complexity	• Low scalability

Table 10 Evaluation metrics in smart agriculture articles

Scope	Ref	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Validity	Resource Utilization
Crop prediction	Tsouli Fathi et al. 2020)	✓	✗	✗	✗	✗	✗	✗	✗
	Velmurugan et al. 2021)	✓	✗	✗	✗	✓	✓	✗	✗
Precision agriculture	Bendre et al. 2016)	✓	✗	✓	✗	✗	✓	✓	✗
	Sabarina and Priya 2015)	✗	✗	✓	✗	✗	✓	✗	✓
	Keswani et al. 2020)	✓	✗	✓	✗	✓	✗	✗	✗
	Failed 2015b)	✓	✓	✓	✗	✗	✓	✗	✗
	Melgar-García, et al. 2022)	✓	✓	✓	✓	✓	✓	✗	✗
	Wang and Mu 2022)	✓	✗	✓	✗	✓	✓	✗	✗
Smart farming	Liu 2021)	✓	✗	✗	✗	✓	✓	✓	✗
	Roukh et al. 2020)	✓	✗	✓	✗	✗	✓	✗	✓
	Li and Niu 2020)	✗	✓	✓	✗	✗	✓	✗	✗
	Osinga et al. 2022)	✗	✗	✓	✗	✓	✓	✗	✗
	Shrivastava et al. 2023)	✗	✗	✓	✓	✓	✗	✗	✓

Khan et al. 2017; Ng et al. 2017; Li et al. 2022; Chang 2021; Tong et al. 2022; Zhang et al. 2022b; Mortaheb and Jankowski 2023). The third category focuses on surveillance systems (Tian et al. 2018; Ramsahai, et al. 2023; Huang et al. 2023). The fourth group talks about the real-time system: (Nathali Silva et al. 2017; Rathore et al. 2018). The last category article concentrates on social network analysis (Azzaoui et al. 2021). In Section 4.4, the selected smart city articles are reviewed and then compared in Section 4.5.2.

4.5.1 Overview of smart city articles

F. Balbin, et al. (Balbin et al. 2020) proposed predictive analytics on an open data portal about bus performance. They concentrated on big data for public transit to deter delays in public transportation. The collected data from the analysis and real-life tests depicted that buses were on time, and the feasibility of analyzing big data was high. Cui, et al. (Cui et al. 2019) proposed a network calculus model to decrease the mean travel hour during rush hours. The results of their experiment demonstrated that the fleet management of autonomous vehicles in a smart city could significantly reduce travel time and energy consumption. Guo and Xu (Guo and Xu 2022) extracted data regarding traffic congestion. They aimed to study the mechanism and the function of diffusion of traffic congestion. As moved forward, they stated that there is no single technology or method to resolve the issue of traffic or even meet the specifications in researching the traffic flow.

Chin, et al. (Failed 2017e) looked for understanding the relationship between weather and not lengthy cycling behavior by utilizing four machine-learning classification algorithms. The outcomes were accurate and reliable. In addition, the results illustrated that the integration of ML, IoT, and big data paved the way for the feasibility of smart city technologies. Khan, et al. (Khan et al. 2017) designed a scheme for energy-aware communications in an IoT environment. Their architecture had four phases: identification of the energy needed for devices, deployment of sensors, scheduling, and information collection. The employed scheme optimized the energy consumption and balanced the load during rush hours.

Ng, et al. (Ng et al. 2017) stated that making use of the value of big infrastructure data is a challenge. In order to tackle that, the authors introduced a master data management (MDM) solution. In their study, the multi-domains master data objects were built using MDM tools, and the MDM was implemented with the registry style. To make a change in the smart city way of data analysis and to have efficient data processing, Li, et al. (Li et al. 2022) suggested a deep learning algorithm within uses big data analysis in addition to the convolutional neural network. The accuracy of their work is estimated to be above 97%, however, the energy consumption and resource utilization were higher than expected.

Chang (Chang 2021) introduced an ethical framework in regard to bid data applications in smart cities and city planning. The author mainly focused on raising public awareness of how to follow the instructions of smart cities. However, the point that was neglected in ethical issues of smart cities was the security and data release without authorization. Tong, et al. (Tong et al. 2022) conducted research using concurrent data driven by the government and national traffic noise to look further into sleep deprivation based on the structure of cities and homes in city planning. One of the highlights was that above 62% of the people had suffered from a lack of sleep due to traffic noise. Zhang, et al. (Zhang et al. 2022b) provided a case study with three phases of development. They studied the case in Wuhu; however, it was the sole case study in that the authors collected data. This limited their work as it had a limited and not enough statistical generalizability.

Lv, et al. (Failed 2022) with concern about the COVID-19 pandemic, tried to make smart city construction much safer and faster by introducing building information modeling, big data processing methods, and tools of digital tools. The authors in Mortaheb and Jankowski (2023) advocated for a reimagining of the smart city concept, emphasizing the crucial role of city planning and the integration of Geospatial Artificial Intelligence (GeoAI). By leveraging the synergies between city planning, big data, geographic information science and systems, and data science, the article proposes achieving policy goals such as enhancing urban services' efficiency, improving quality of life, addressing urban challenges, and generating valuable spatial data and knowledge. Tian, et al. (Tian et al. 2018) represented a block-level background modeling (BBM) algorithm to support long-term reference structure for efficient surveillance video coding and also developed a rate-distortion optimization for the surveillance source algorithm. Ramsahai, et al. (Ramsahai, et al. 2023) employed BDPA techniques such as exploratory data analysis, geocoding for hotspot mapping, and kernel density estimation, to analyze historical crime data and make crime predictions. Additionally, the study confirms the relevance of Twitter data in crime analysis, with the integration of Twitter data improving accuracy by 9%.

Huang, et al. (Huang et al. 2023) focused on the security threshold setting algorithm for a distributed optical fiber monitoring and sensing system based on big data. The components of the system are introduced, factors affecting system performance are analyzed, and methods for enhancing system performance are summarized. The proposed algorithms required less storage, powerful code sustainability, and high productivity. Nathali Silva, et al. (Nathali Silva et al. 2017) presented the design of a smart city based on big data analytics. Their model consisted of three levels: data generation and acquisition level, data management and processing level, and application level. The authors tested their model on the Hadoop environment with the datasets that were obtained from various authentic and reliable sources.

Rathore, et al. (Rathore et al. 2018) organized a system for smart digital cities to pave the way for acquiring information. The gathered data was processed in a real-time environment to reach the smart city by using Hadoop working under Apache Spark. The authors illustrated that the efficiency of the proposed system was enhanced when big data was using Apache Spark over Hadoop. El Azzaoui, et al. (Azzaoui et al. 2021) developed a big data analysis framework based on information shared openly on the social network service (SNS) platform to monitor, comprehend, and forecast the immediate future virus widespread besides controlling the infodemic and prevent biased or manipulated news from broadcasting. The authors proposed natural language processing (NLP) to obtain more precise and accurate data regarding the prediction of a virus outbreak.

4.5.2 Summary of smart city articles

The main goal of each reviewed article was to enhance the feasibility of smart cities by testing models in a real-time environment. According to the reviewed and discussed smart city articles, the comparison of their specifications has been depicted in Table 11. Table 11 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each smart city article. In addition, Table 12 displays the improvement of different evaluation metrics in each study. These metrics include accuracy, time, performance, reliability, energy, scalability, throughput, sustainability, feasibility, and security.

Table 11 Comparison of smart city articles

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Vehicle	Balbin et al. 2020)	Big data for public transit	Real testbed	Not mentioned	• High feasibility • High reliability • High accuracy	• High complexity
	Cui et al. 2019)	Developing intelligent management of autonomous vehicles in a smart city with the assessments of big vehicular data analytics	Prototype	Not mentioned	• Low execution time • Low energy • High efficiency	• High complexity
	Guo and Xu 2022)	Visualization of the traffic flow by the help of big data	Real testbed	Not mentioned	• High reliability • High accuracy	• Not having a simplified single method
City planning	Failed 2017e)	The ability of system information alongside artificial intelligence to customize the packages of smart city	Simulation	WEKA	• High security • High accuracy	• High execution time
	Khan et al. 2017)	Energy-aware communications in an IoT environment	Prototype	Hadoop	• Energy sufficient	• The system was not being tested in real sensors
	Ng et al. 2017)	Accessing to the value of big data infrastructure by master data management solution	Simulation	Not mentioned	• High accuracy • High consistency • High reliability	• Cannot maintain its efficiency when the number of users demands increases
	Li et al. 2022)	Big data analysis on digital twins of smart city	Simulation	Python, MATLAB	• High accuracy and consistency	• Had only produced an offline batch analysis

Table 11 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Chang 2021)	Ethical framework for the use of big data in city planning	Simulation	Not mentioned	• High accuracy	• Low scalability • Low security
	Tong et al. 2022)	Traffic noise affects people's sleeping cycles	Prototype	Not mentioned	• High efficiency	• Not exploring other impactors in the sleeping cycle of people, such as spatial correlation
	Zhang et al. 2022b)	A case study of smart city development	Prototype	Not mentioned	• High reliability • Low costs	• Limited case study; only one • Limited data collected
	Failed 2022)	The constriction of the smart city by digital twins	Simulation	Not mentioned	• Low time of execution • High performance	• No analysis of multi-source heterogeneity
	Mortaheb and Jankowski 2023)	GeoAI	Prototype	Spatial Decision Support Systems	• High efficiency	• Limited data collected
Surveillance system	Tian et al. 2018)	A block-level background algorithm to create a universal image for encoder reference	Simulation	BBM SRDO	• High compression efficiency • High performance • Less storage requirement	• The modeling strategy was not optimized • Not using content-based optimization
	Ramsahai, et al. 2023)	BDPA for exploring and predicting crime patterns	Simulation	NLP, exploratory data analysis	• High accuracy • High reliability	• High complexity
	Huang et al. 2023)	Security threshold setting algorithm for a distributed optical fiber monitoring and sensing system based on big data	Prototype	COMSOL	• High security • High performance	• Low scalability

Table 11 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Real-time system Social network analysis	Nathali Silva et al. 2017)	Proposing architecture of a smart city based on big data analytics	Real testbed	Hadoop HDFS HBASE HIVE Kalman filter	• High throughput • Low execution time	• Low scalability • The accuracy and efficiency of the proposed model is not stated, and it carried out a simulated experiment
	Rathore et al. 2018)	Setting an IoT-based smart city by using big data analytics while utilizing real-time data from the city	Real testbed	Hadoop Spark	• High efficiency • High throughput • Capable of working in a real-time environment	• High processing time
	Azzaoui et al. 2021)	Implementing a shared framework to predict the epidemic widespread and track its growth across the universe	Real testbed	SNS NLP	• High accuracy • High security • High performance	• SNS platforms could not be examined and have varied outcomes by different linguistics and geospatial options

Table 12 Evaluation metrics in smart city articles

Scope	Ref	Accuracy	Time	Performance	Reliability	Energy	Scalability	Throughput	Sustainability	Feasibility	Security
Vehicle	Balbin et al. 2020)	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗
	Cui et al. 2019)	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
	Guo and Xu 2022)	✓	✓	✓	✗	✗	✓	✗	✗	✗	✓
City planning	Failed 2017e)	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓
	Khan et al. 2017)	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗
	Ng et al. 2017)	✓	✗	✗	✓	✗	✗	✗	✓	✓	✓
	Li et al. 2022)	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
	Chang 2021)	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗
	Tong et al. 2022)	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗
	Zhang et al. 2022b)	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗
	Failed 2022)	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗
	Mortaheb and Jankowski 2023)	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓
Surveillance system	Tian et al. 2018)	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
	Ramsahai, et al. 2023)	✓	✗	✓	✓	✗	✗	✓	✗	✗	✓
Real-time system	Huang et al. 2023)	✓	✗	✓	✓	✗	✓	✗	✗	✗	✓
	Nathali Silva et al. 2017)	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
	Rathore et al. 2018)	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗
Social network analysis	Azzaoui et al. 2021)	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓

4.6 ICT applications

These articles are of different kinds and subjects and can be divided into two classes. The first class targets articles that are connected with algorithms or deal with machine learning approaches: (Nural et al. 2015; Failed 2017f; Oo and Thein 2019; Kannan et al. 2018; AlFarraj et al. 2019; Khine and Nyunt 2019; Arun Kumar and Venkatesulu 2019). The second class includes articles that point to technology: (Shenoy and Gorinevsky 2015; Su and Huang 2018; Mujeeb et al. 2019). The popular feature in all of these articles is combining big data and data mining techniques. In Section 4.4.2, the selected ICT articles were reviewed. Finally, in Section 4.6.2, the discussed articles are compared and summarized.

4.6.1 Overview of ICT articles

Applying semantic technology was suggested by Nural, et al. (Nural et al. 2015) to help data analysts and scientists in the selection of proper techniques of modeling and creation of special models, as well as the introduction of the rationale for the selected models and techniques. To describe models, modeling techniques, and results, the analytic ontology was developed by the authors to support inferencing for the selection of a semi-automated model. The ScalaTion framework is applied as a testbed to assess semantic technology use, and this framework supports more than thirty techniques of modeling for predictive big data analytics.

Putting special emphasis on modeling for predictive analytics, through meta-learning, with algorithms based on regression, Nural, et al. (Failed 2017f) tried to focus on and discuss the progress that was made in automated modeling. Besides a meta-learning system development, the authors introduced a wide range of meta-attributes to capture the features related to algorithms of regression. An influential system of predictive analytics was proposed by Oo and Thein (Oo and Thein 2019) for huge dimensional big data via an increasing scalable random forest (SRF) algorithm on the Apache Spark platform. Optimization of hyperparameters enhanced SRF, and the reduction of dimensions improved the prediction performance. The suggested system efficacy is tested on different real-world datasets. Based on the outcomes, the suggested approach attains a competitive performance in comparison with the RF algorithm that is implemented by Spark MLlib.

By applying the support vector machine approach (PAD-SVM), Kannan, et al. (Kannan et al. 2018) suggested a predictive analysis of demonetization data. This suggested that PAD-SVM had three steps: preprocessing, descriptive analysis, and prescriptive analysis. In the pre-processing step, the obtained data were cleaned, missing value treatment was performed, and the necessary data from the tweets were split. In the descriptive analysis step, the most effective individuals were found, this subject was regarded, and analytical functionalities were performed. Performing semantic analysis was to find users' sentiment values and to find each tweet's compound polarity. Predictive analysis was done to see people's present mindsets and the reaction of the community to current time issues. The purpose of performing this analysis was to find society's overall viewpoints and the views that might alter in the near future considering the demonetization scheme.

An optimized feature selection method and techniques of soft computing were introduced by AlFarraj, et al. (AlFarraj et al. 2019), to reduce dataset dimensionality. First, they collected data from different resources containing some inconsistent data, which reduced system efficiency. After that, they removed noise and inconsistency by using a normalized approach.

They also chose the optimized traits by applying the firefly's gravitational ant colony optimization method. The noted method of optimized feature selection could examine the features during the selection process. The chosen features were composed of details of special predictive analytics. The efficiency of the proposed system was assessed by applying various datasets.

Khine and Nyunt (Khine and Nyunt 2019) suggested Map Reduce on the basis of the multiple linear regression model that is appropriate for distributed and parallel execution, aiming at predictive analytics upon massive datasets. This QR decomposition-based model in the decomposition of big matrix training data unearths model coefficients from a huge amount of matrix data on the Map-Reduce framework extensively.

Arun Kumar and Venkatesulu (Arun Kumar and Venkatesulu 2019) suggested a Gramian symmetric data collection-based random forest bivariate regression and classification method to make improvements in the accuracy of prediction with less complication. Firstly, they collected a huge data volume. They used the Gramian symmetric matrix for storing the data volume in columns and rows of a matrix. Afterward, they classified and did the regression process by applying random decision forests in order to locate outcomes in the future. The relationship between independent variables (i.e., data) and a dependent variable (i.e., outcomes) was measured by the regression process via bivariate correlation. Random decision forests made some decision trees to classify on the basis of the correlation. In the end, it mixed some decision trees and used a voting scheme. Classification results vote was identified in order to attain great.

An influential computational methodology was presented by Shenoy and Gorinevsky (Shenoy and Gorinevsky 2015) for cross-sectional longitudinal analysis of ultimate event statistics in large data sets. The data that were analyzed were accessible through multiple periods of time and multiple individuals in a population, some of which may not have extreme events, and some may have data. They modeled utmost events with an exponential tail or Pareto distribution. The suggested technique is on the basis of non-parametric Bayesian formulation.

On the basis of Apache Spark, Su and Huang (Su and Huang 2018) decided to expand a real-time predictive maintenance system for detecting failures of imminent hard disk drives (HDD) in data centers. A real-time prediction system development was described, which was capable of assisting IT teams in having extensive storage systems by sending notifications for failures of the impending drive. A framework was described for failures of HDD predictive monitoring via analyzing files of machine logs rather than applying conventional methods of statistical prediction.

Mujeeb and Javaid (Mujeeb et al. 2019) regarded load relationships and price while proposing two multiple inputs multiple outputs deep recurrent neural network models for forecasting load and price. An efficient sparse autoencoder nonlinear autoregressive network with exogenous inputs as the first suggested model is composed of forecasting and feature engineering. They suggested ESAE for feature engineering and performed forecasting by applying NARX as an existing method. Differential evolution recurrent extreme learning machine (DE-RELM), as the second suggested model, is based on the meta-heuristic DE optimization technique and RELM model. The predictive and descriptive analysis was conducted on PJM and ISO NE, as two famous electricity markets' big data.

4.6.2 Summary of ICT articles

Most articles try to have predictions that are more accurate in their related domains. These studies, on the other hand, propose several beneficial tools (like Apache Spark, Hadoop.) to

be applied for data mining when facing big data. According to the reviewed and discussed ICT articles, the comparison of their specifications has been depicted in Table 13. Table 13 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each ICT article. In addition, Table 14 displays the improvement of different evaluation metrics in each study. These metrics include accuracy, timeliness, cost, scalability, reliability, performance, validity, and resource utilization.

4.7 Weather applications

This section represents the application of big data in weather-related activities, which we have divided into five sub-classes. The first class takes a look at the influence of weather on sport (Zhao et al. 2018; Abeza et al. 2022). The second sub-class discusses the employment of weather prediction in electric power systems (Kezunovic, et al. 2017). The third category mainly focuses on weather forecasting: (Aljawarneh et al. 2020; Alam and Amjad 2019; Failed 2018e, 2017g; Liu et al. 2015; Simpson and Nagarajan 2021; Reis et al. 2022; Roh 2022). The fourth category discusses the prediction of fire risk (Agarwal et al. 2020). The last sub-class explores the wind pace by utilizing weather prediction big data (Xu et al. 2020). In Section 4.5.1, the selected weather articles are reviewed and then compared in Section 4.7.2.

4.7.1 Overview of weather articles

Zhao, et al. (Zhao et al. 2018) proposed a real-time model to comprehend which weather alterations influenced cycling on an off-road trail and an on-road bridge cycling lane. They tested the model to calculate the proportion of cycling during different weather conditions. However, their model suffered from the limitation of varied weather conditions, such as snowfall. Abeza, et al. (Abeza et al. 2022) conducted semi-structural interviews with practitioners in four top leagues in the U.S. The four strategic factors that the authors considered were transactional, informational, strategic, and infrastructural. Kezunovic, et al. (Kezunovic, et al. 2017) introduced an application of big data to investigate climatic situations' effects on power system running time, outcome, and administration. They conducted a methodology that used the Spatio-temporal correlation between diverse data sets to create more appropriate decision-making tactics for smart distribution networks. Their prospect framework was called Gaussian conditional random fields, which was applied to two power system applications: Risk assessment and spatio-temporal solar generation.

Aljawarneh, et al. (Aljawarneh et al. 2020) presented a system that could cope with climatic variables which were related to big data. They designed three categories of experiments for testing: First, the execution of standard univariate analysis. Second, the performance of the multivariate analysis compared to univariate analysis, and third, the productivity of the neighbor-based analysis approach compared to univariate. The authors considered the local NoSQL database at different levels in order to execute a predictive analysis by utilizing univariate and multivariate solutions as well as forecasting based on training data from neighbor stations. Alam and Amjad (Alam and Amjad 2019) introduced weather data analysis by taking Hadoop with multiple node systems into account. They designed system architecture for weather prediction using a big data-based analytic approach in cloud circumstances to forecast the highest, lowest, average, and mild temperatures.

Table 13 Comparison of ICT articles

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Algorithm	Nural et al. 2015)	A framework to support selecting a semi-automated model and execution of the model to conduct predictive analytics	Simulation	Not mentioned	• High timeliness • High resource utilization	• Some concepts are not considered (skewness, kurtosis.)
	Failed 2017f)	Selection system of a meta-learning-based model and procuring a system assessment	Simulation	ScalaTion Framework	• High accuracy • High resource utilization	• Exclusion of metadata
	Oo and Thein 2019)	Suggesting a system of sufficient predictive analytics for high measurements of big data via increasing scalable random forest algorithm	Real testbed	Apache Spark	• High accuracy • High scalability	• A real-time PBA system for the high dimensional big data is excluded
	Kannan et al. 2018)	The use of BDPA supports vector machines on demonetization data	Simulation	Pydoop	• High accuracy	• Data source was limited (only Twitter)
	AlFarraj et al. 2019)	Optimized feature selection method on the basis of fireflies with gravitational ant colony algorithm for BDPA	Real testbed	MATLAB, Weka	• High accuracy	• High complexity
	Khine and Nyunt 2019)	MapReduce-based multiple linear regression model for parallel and distributed processing aiming at predictive analytics on massive datasets	Simulation	Hadoop	• High scalability • High flexibility	• No pre-processing provided

Table 13 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Arun Kumar and Venkatesulu 2019)	Introducing the efficient approach of Gramian symmetric data collection-based random forest bivariate regression and classification for healthcare BDPA	Simulation	Not mentioned	• High timeliness • High resource utilization	• Scalability was not considered
Technology	Shenoy and Gorinevsky 2015)	A computational methodology efficient for analyzing extreme event statistics longitudinally and cross-sectionally in large data sets	Real testbed	Not mentioned	• High accuracy • High scalability	• Methodology applied only on climate data and electrical power grid
	Su and Huang 2018)	Developing a maintenance system of real-time predictive for a hard disk drive	Simulation	Apache Spark	• High accuracy • High timeliness • High reliability	• High complexity
	Mujeeb et al. 2019)	Regarding price forecasting and electricity, load participate in the PJM and ISO NE markets for regulating the demand and price in the USA's power system	Simulation	MATLAB	• High accuracy • High performance	• No real-world scenario implementation of smart grid or microgrid

Table 14 Evaluation metrics in ICT Articles

Scope	Ref	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Validity	Resource Utilization
Algorithm	Nural et al. 2015)	✗	✗	✓	✗	✗	✓	✗	✓
	Failed 2017f)	✓	✗	✗	✓	✗	✓	✓	✓
	Oo and Thein 2019)	✓	✗	✗	✓	✓	✓	✗	✗
	Kannan et al. 2018)	✓	✗	✗	✓	✗	✓	✗	✗
	AlFarraj et al. 2019)	✓	✓	✓	✓	✗	✗	✗	✗
	Khine and Nyunt 2019)	✓	✓	✗	✓	✗	✓	✗	✗
	Arun Kumar and Venkatesulu 2019)	✓	✓	✗	✗	✗	✓	✗	✓
Technology	Shenoy and Gorinevsky 2015)	✓	✗	✗	✓	✗	✗	✗	✗
	Su and Huang 2018)	✓	✓	✓	✓	✓	✓	✗	✗
	Mujeeb et al. 2019)	✓	✗	✗	✗	✗	✓	✓	✗

Madan, et al. (Failed 2018e) explored an ongoing statistical linear regression and support machine learning, which could make constant types of information from equipment groups and weather forecasts. The results were originally accurate thus a method to level up the straight relapse was by collecting more information utilizing linear regression and supporting vector machine toward the sustainable and productive model. Onal, et al. (Failed 2017g) used big data IoT framework in a climate data analysis. They took into account weather clustering by utilizing a publicly available dataset. Their selected dataset was from library sci-kit-learn-based k-means clustering. In order to examine their use case, the authors provided the implementation details of each framework layer.

Liu, et al. (Liu et al. 2015) employed a computational intelligence technology named stacked auto-encoder to imitate climate data for three decades. They mainly introduced the greedy layer-wise unsupervised pre-training approach based on a deep neural network. Their proposed model levigated the criteria of the raw weather data layer by layer, and the test's outcomes depicted that the newly acquired features can enhance the performances of classical computational intelligence models. Simpson and Nagarajan (Simpson and Nagarajan 2021) presented a deep learning algorithm according to a stacked sparse autoencoder for forecasting the varied climatic conditions of a specific area. They imposed the principal component analysis to lower the extraction of criteria by dramatic variance. In addition, the authors presented an algorithm based on binary butterfly optimization algorithm along with a deep stack autoencoder to level up precision. Simulation results indicated that execution time and prospective errors decreased.

Reis, et al. (Reis et al. 2022) studied UHI in Lisbons' metropolitan neighborhood through the local weather types with the usage of the dataset available from Copernicus which was collecting data between 2008 and 2014. The outcomes of the analysis were a positive and correlative relationship between the modeled air temperature and the measurements, high precipitation rate, and the decomposition of UHI. Roh (Roh 2022) used the collected traffic data from five different WIM locations in Canada. The author divided each type of car into three categories on the road, in which he detected that winter traffic groups are more transferrable to homogeneous and heterogeneous road segments.

Agarwal, et al. (Agarwal et al. 2020) used two sets of big data, which were fire incident data from the National Fire Incident Reporting System and climatic data from the National Oceanic and Atmospheric Administration to achieve an overall record for forecasting and analyzing the fire risk. The authors described gradient-boosting trees and machine learning algorithms to precisely predict the incidents of a future fire. Xu, et al. (Xu et al. 2020) applied a framework to compute the wind pace forecasting based on the Apache Spark platform and using Python API for Spark. The authors proposed a synthesis computing framework applied to wind speed. The simulation results illustrated the proposed framework on Spark to foresee the wind's speed precisely and effectively.

4.7.2 Summary of weather articles

The studied articles mainly focused on the accuracy and costs of the presented models for forecasting climatic conditions by using various precise tools, such as Hadoop, GIS, and MongoDB. According to the reviewed and discussed weather articles, the comparison of their specifications is depicted in Table 15. Table 15 indicates the main ideas, evaluation methods, tools, advantages, and disadvantages of each weather article. In addition, Table 16 displays the improvement of different evaluation metrics in each

Table 15 Comparison of weather articles

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Sport	Zhao et al. 2018)	Targeted to note research voids through researching the effects of climate on cycling	Real testbed	Not mentioned	• High accuracy • High performance • High throughput • Low execution time	• Not being tested in a variety of climatic conditions, such as snowfall
	Abeza et al. 2022)	Using big data in professional sport	Simulation	Not mentioned	• High accuracy • High reliability	• Limited number of interviews • Low number of real datasets
Electric power system	Kezunovic, et al. 2017)	How big data can be implemented, spatiotemporally related, and tested in real-time in order to boost the abilities of modern electricity networks	Simulation	GIS	• Accurate prediction • Low cost	• Scalability and reliability are overlooked
Weather forecasting	Aljawarneh et al. 2020)	Proposing a visual big data system which is implemented to handle immense climate data	Prototype	NoSQL, MongoDB	• High precision • High prediction accuracy • High performance	• Low flexibility • Not measuring varied parameters • Low scalability
	Alam and Amjad 2019)	Designing architecture for parallel and distributed evaluation of big data in the cloud environment	Design	Hadoop, HDFS, MapReduce,	• High scalability	• High costs • Low accuracy • High execution time
	Failed 2018e)	Inspecting continuous statistical linear regression and assisting vector machine techniques	Simulation	SVM, MapReduce, Hadoop	• High accuracy • Valid and reliable forecast • High efficiency	• Not using Apache Spark for the proposed mode for synchronous forecasting

Table 15 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
	Failed 2017g)	Proposing a use case with a widened IoT framework that combines the data mining, execution time, and layers on weather data clustering analysis	Design	Node.js	• High resource utilization • High throughput • High scalability	• Possible errors occurred during using cluster method
	Liu et al. 2015)	A method that uses computational intelligence technology to process the immense volume of data	Simulation	SVM	• High accuracy • High performance	• Not testing the proposed model in varied climatic conditions
	Simpson and Nagarajan 2021)	The hybrid model for predicting the future weather condition	Simulation	Kernel	• High precision in prediction • High accuracy • Low error rate • Low computation time	• Not utilizing Hadoop and Spark for the verified climate prediction
	Reis et al. 2022)	Urban heat island effects measured with the help of big data	Simulation	Copernicus Climate Change Service Data Set	• High correlation • High reliability	• Not thoroughly computing the statistical differences between local weather types and climate zone
	Roh 2022)	Winter-level model created by big data	Prototype	Not mentioned	• High reliability • High performance	• Low security • High costs

Table 15 (continued)

Scope	Ref	Main idea	Evaluation method	Tool	Advantage	Disadvantage
Fire risk prediction	Agarwal et al. 2020)	Effects of climatic situations on the irreplaceable consequences of fire incidents	Simulation	Not mentioned	• High prediction accuracy • High precision	• Performance and energy are not considered
Wind prediction	Xu et al. 2020)	A hybrid administrator for calculating the wind pace big data prediction	Simulation	Apache Spark, Hadoop, Python Spark	• High accuracy • High performance • Low execution time	• Not paying attention to costs and scalability

Table 16 Evaluation metrics in weather articles

Scope	Ref	Accuracy	Time	Performance	Reliability	Energy	Scalability	Throughput	Sustainability	Feasibility	Cost
Sport	Zhao et al. 2018)	✓	✓	✓	✗	✗	✗	✓	✗	✓	✓
	Abeza et al. 2022)	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Electric power system	Kezunovic, et al. 2017)	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓
Weather forecasting	Aljawarneh et al. 2020)	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓
	Alam and Amjad 2019)	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓
	Failed 2018e)	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗
	Failed 2017g)	✗	✗	✓	✗	✗	✓	✓	✗	✓	✗
	Liu et al. 2015)	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗
	Simpson and Nagarajan 2021)	✓	✓	✓	✓	✗	✗	✓	✗	✗	✓
	Reis et al. 2022)	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
	Roh 2022)	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓
Fire risk prediction	Agarwal et al. 2020)	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓
Wind prediction	Xu et al. 2020)	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗

study. These metrics include accuracy, time, performance, reliability, energy, scalability, throughput, sustainability, feasibility, and cost.

5 Discussion

Previous sections described the review process of the selected articles in BDPA in seven groups. Here in this section, the authors deliberate a statistical analysis of the reviewed on the basis of different attributes for answering the RQs in Section 3.1:

5.1 Overview of the selected studies

To examine the current state of research on BDPA, the following supplemental questions are considered:

- Which years had seen the most number of published articles in the area of BDPA?
- In which representatives did the researchers provide their results?

The articles were classified based on years of publication, from 2014 to 2023, which is illustrated in Fig. 3. We had the highest number of published articles in 2021 and 2022. Figure 4 depicts the classification of the studied articles over time as per journals, including ScienceDirect, Emerald, IEEE, Springer, Taylor & Francis, Hindawi, Wiley, SAGE, and ACM. Research on BDPA is currently in a progressive state, with researchers actively exploring methodologies, algorithms, and applications for handling and analyzing large datasets. There is also a focus on generative AI and machine learning techniques while addressing ethical and privacy concerns associated with BDPA. Figure 5 illustrates the classification among nine publishers, where 34% of the total articles belong to ScienceDirect, 25% to Springer, 21% to IEEE, 6% to Taylor & Francis, 6% to Hindawi, 3% to Emerald, 3% to Wiley, and the least proportion is constituted by SAGE and ACM at 1% each.

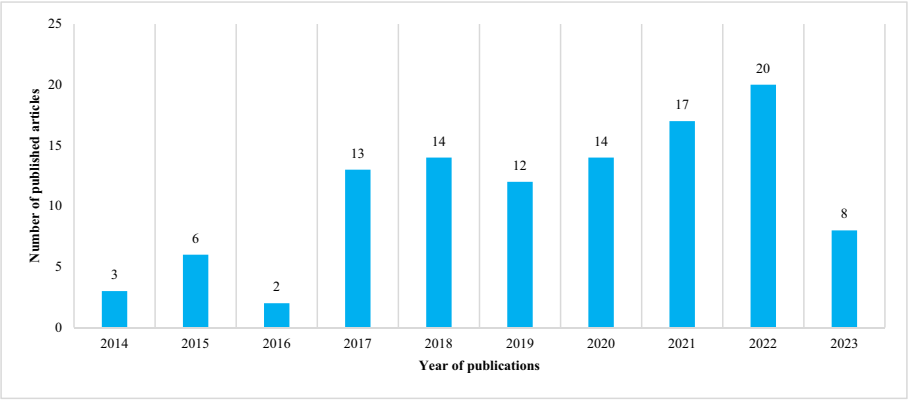


Fig. 3 Distribution of articles by publication year

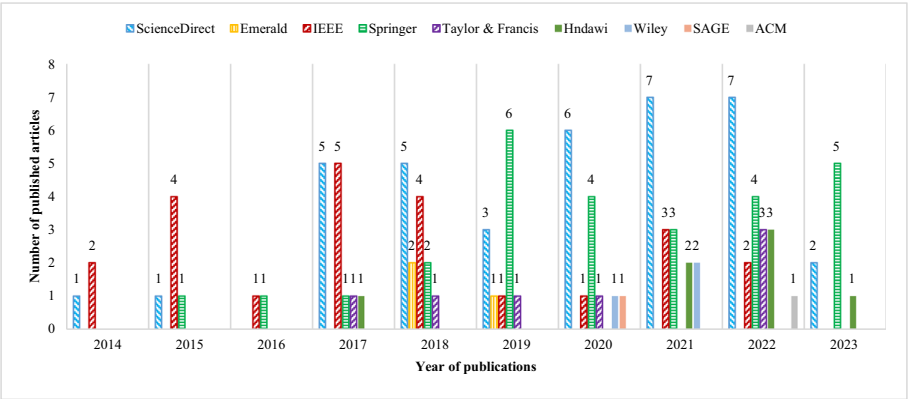
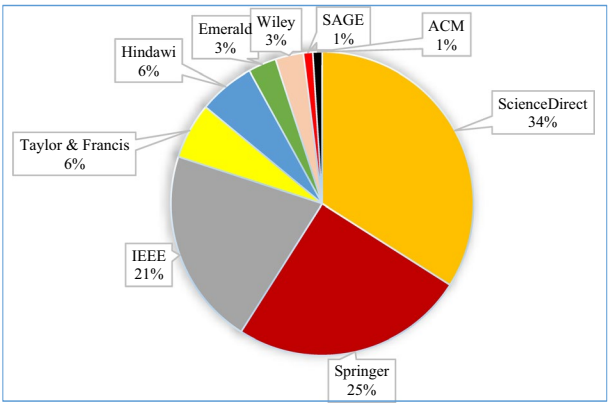


Fig. 4 Distribution of articles over time for each publisher

Fig. 5 Percentage of articles by different publishers



5.2 Research aims, methods, and evaluation metrics

This section aims to provide clear answers to the stated RQs 1 to 4, in Section 3.1, based on the collected statistical data.

- RQ1: What are the fields of prediction analysis applications in big data?

Figure 6 presents the comparison side of the big data predictive analytic applications. The best parameters to be used for categorization are the domains and main topics of articles that establish a logical relationship between them. Articles were divided into seven different categories based on their application. Industrial articles comprised 19% and smart city 17%. Smart healthcare and e-commerce articles comprised 32%, collectively, and smart agriculture made up 12%. Weather and ICT had the lowest number of the articles, with 11% and 9%, respectively. So, industrial articles own a great number of articles in BDPA. However, ICT and weather possessed a few numbers of articles in BDPA. Table 17 represents a summary of the benefits and limitations of the discussed

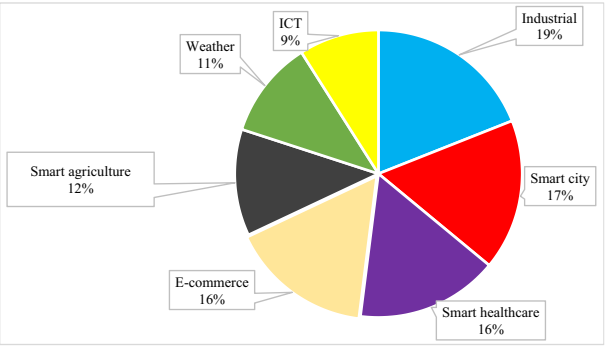


Fig. 6 Percentage of big data predictive analytic approaches

Table 17 The main pros and cons of the discussed classification(add q)

Application	Pros	Cons
Industrial	• Greater accuracy • Greater reliability • Greater performance • Greater validity	• Unsatisfactory resource utilization • Unsatisfactory cost
E-commerce	• Greater accuracy • Decreasing cost	• Unsatisfactory reliability • Unsatisfactory resource utilization
Smart healthcare	• Greater accuracy • Decreasing cost	• Unsatisfactory validity
Smart agriculture	• Greater accuracy • Decreasing cost • Greater performance	• Unsatisfactory timeliness • Unsatisfactory scalability • Unsatisfactory reliability
Smart city	• Greater accuracy • Greater feasibility	• Unsatisfactory throughput • Unsatisfactory sustainability
ICT	• Greater accuracy • Greater performance • Greater scalability	• Unsatisfactory reliability • Unsatisfactory validity
Weather	• Greater accuracy • Greater performance	• Unsatisfactory scalability • Unsatisfactory sustainability

groups. It indicates that all applications have distinguishing features such as better accuracy and better performance.

•RQ2: What are the evaluation metrics of predictive analytics using big data?

There were several metrics to evaluate the predictive analytics, but some of them, including accuracy, timeliness, cost, scalability, reliability, performance, and time were more popular among authors. According to Fig. 7, it is observed that most of the articles focused on accuracy by 18% while resource utilization was considered in only 5% of the reviewed papers. Performance and others (energy, throughput, feasibility, security, precision, and sustainability) constituted 14% each. Likewise, cost and reliability were the main consideration of 22% of articles, collectively. Figure 8 specifies the importance of each evaluation metric in different categories. By applying (2), to calculate the

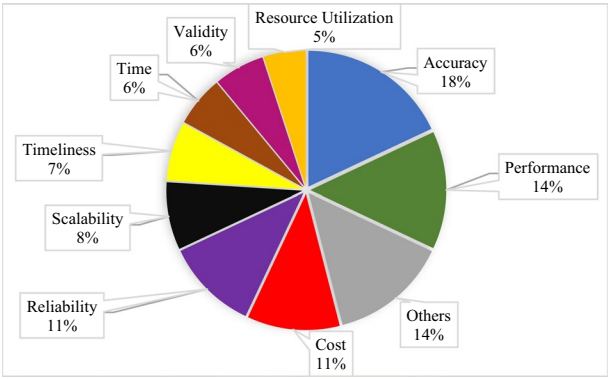
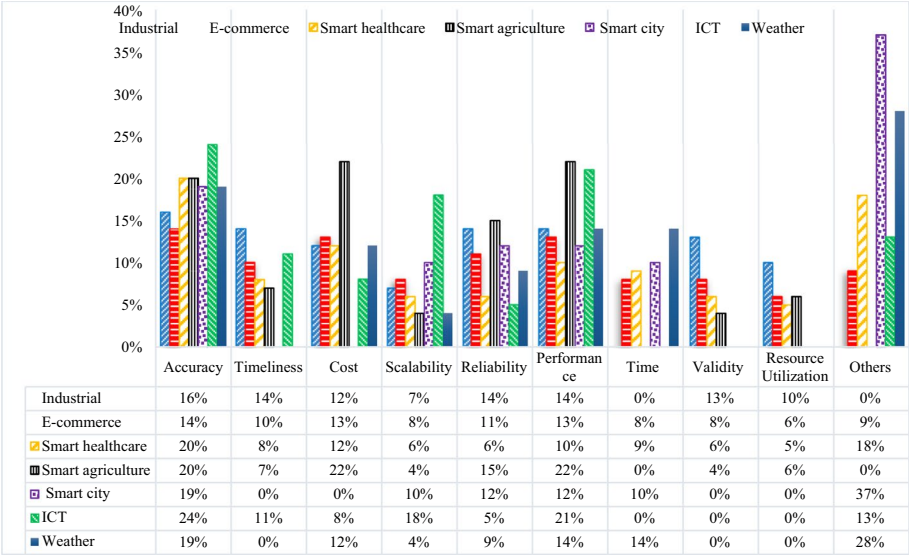


Fig. 7 Percentage of evaluation metrics in BDPA



	Accuracy	Timeliness	Cost	Scalability	Reliability	Performance	Time	Validity	Resource Utilization	Others
Industrial	16%	14%	12%	7%	14%	14%	0%	13%	10%	0%
E-commerce	14%	10%	13%	8%	11%	13%	8%	8%	6%	9%
Smart healthcare	20%	8%	12%	6%	6%	10%	9%	6%	5%	18%
Smart agriculture	20%	7%	22%	4%	15%	22%	0%	4%	6%	0%
Smart city	19%	0%	0%	10%	12%	12%	10%	0%	0%	37%
ICT	24%	11%	8%	18%	5%	21%	0%	0%	0%	13%
Weather	19%	0%	12%	4%	9%	14%	14%	0%	0%	28%

Fig. 8 Percentage of evaluation metrics in each category in BDPA

percentages in each category, the number of a *metric* has been counted separately and divided by the sum of the number of all metrics.

$$Imp_{percentage}(i) = \frac{metric(i)}{\sum_{j=1}^{n} metric(j)} \tag{2}$$

- RQ3: What evaluation methods are used in BDPA?

Figure 9 depicts the comparison results in Tables 3, 5, 7, 9, 11, 13, and 15. According to Fig. 9, it was observed that 52% of case studies put into work the simulation evaluations, 30% considered a real testbed environment, 11% of case studies used to

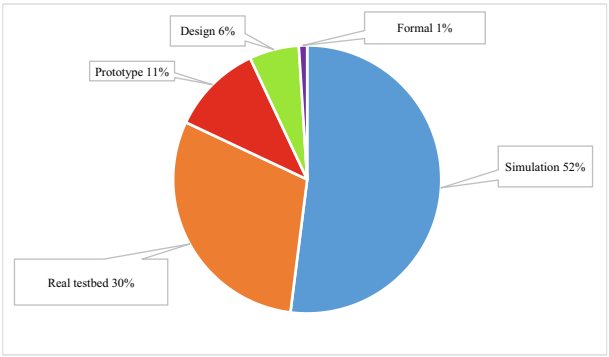


Fig. 9 Percentage of evaluation methods in BDPA

prototype their study, 6% of them just designed an algorithm or framework, and other 1% used formal methods.

- RQ4: What are the tools and environments in BDPA?

Based on Tables 3, 5, 7, 9, 11, 13, and 15, it was observed that various tools and modeling environments were used in case studies, such as Hadoop, Apache Spark, and MATLAB. However, many articles did not mention their tools.

6 Open issues and future trends

In this section, the following research question is addressed:

- RQ5: What are the challenges and future issues of BDPA?

To answer RQ5, major challenges that impact BDPA are briefly discussed. The main challenges for exploiting big data for predictive analysis are streaming data sources, user privacy, multiple information sources, vertical domain applications, scalability, structured data vs. unstructured, incompleteness, leadership, adoption, and trust. Then each of these challenges is discussed, respectively.

- *Streaming data sources*

  The recent explosion of social media services like Facebook, Instagram, and Twitter has led to a significant interest in social media predictive analytics and has attracted many researchers – automatically inferring hidden information from a large number of freely available content. It has a large number of diverse applications, including personalized marketing, real-time healthcare analytics (Balduini et al. 2014), online targeted advertising (Kuang et al. 2018), politics (Bianchi 2019), personalized recommendation systems and searches (Cheung et al. 2018), large-scale passive polling, and real-time live polling. One of the most noteworthy topics in social media predictive analytics is sentiment analytics (Park et al. 2018; Nguyen et al. 2014; You et al. 2015), which is the technique of figuring out if a posted report is negative, positive, or neutral. Sentiment analytics helps data analysts within large enterprises conduct nuanced market research, and public opinion, understand customer experiences, and monitor brand and product

reputation (Hajiali 2020). Despite its value and importance, building predictive models for social media data is a challenging and overwhelming task as social data is large in volume, fast in pace, and heterogeneous in content. How to tackle all these challenges simultaneously is still an open problem.

- *Privacy of data*

  Rigorous ethical authorizations are required to collect data that may include the personal information of users and process the collected data for downstream tasks (Scotti 2017; Leary 2015). However, the big data era now makes such data available to organizations to explore without a crystal-clear ethical policy in place. Even though anonymization of data is common to protect user's privacy, there remains a potential risk of eliciting critical information form a big pile of data. In addition, predictive analytics can cause privacy concerns, especially when sensitive information is used (Gong et al. 2016). Predicting which employees are likely to quit their jobs and delivering that information to their manager is an example. Hence, despite previous studies, it is still a very interesting subject to understand the ethical and privacy implications of BDPA.
- *Multiple information sources*

  Different pieces of data are often housed in different systems. This may lead to incomplete or inaccurate analysis (Jiang et al. 2016). Combining data manually is time-consuming and can limit insights into what is easily viewed. Many problems that are related to incorrect insights can be terraced back to the way data is gathered, verified, stored, and used. When one works with data-sensitive/insensitive industries, the tiniest error seems to be critical to the overall process success (Zhao et al. 2019; Li et al. 2016). Furthermore, it is common to have inconsistent information coming from different sources. Thus, how to integrate data from different sources to develop better predictive models remains an open issue.
- *Vertical domain applications*

  With regard to applications, predictive analytics on big data was applied to distinct domains, including e-commerce and marketing intelligence (Das et al. 2017; Tuladhar et al. 2018), healthcare (Harris et al. 2016; Belle et al. 2015), financial (Ravi and Kamaruddin 2017), security (Shao et al. 2018), public safety (Turet and Costa 2018), and utility. For example, online retailers such as eBay, Amazon, and Alibaba use BDPA to collect insights, predict consumer behavior and operational efficiency, and improve their customer relationship management initiatives, decision-making, and marketing campaigns. Although some studies have been done in various industries, it is still an attractive open issue.
- *Scalability*

  One of the significant evaluation metrics in BDPA, according to reviewed articles, is the scalability of algorithms and platforms (Hu et al. 2014). Possessing a huge amount of data is totally beneficial for database systems. As social networks are popular, collections of huge and extended databases have been developed. No need to say that it will be essential to set some limitations via system scalability. Scaling methods are so challenging that communication and synchronization overheads rise. Famous and successful organizations try to scale their overhead capabilities, specifically in the case of predictive analytics and data mining, to make an improvement in business performance and to decrease fraud (Sun et al. 2019). However, it is complicated to scale conventional approaches to predictive analytic projects and to execute them in a real-time environment which is needed in the architecture of modern enterprises. Therefore, scaling data and scalable analytic algorithms to generate real-time results can be a direction for future work.

- *Structured data vs. unstructured data*
  With recent data mining statistical methods, the analysis of the structural data is not difficult; however, in doing so, such techniques as natural language processing (Quan et al. 2019), multimedia analytics (Fiadino et al. 2016), and text analytics need to be improved. It is also costly to process unstructured data for analysis as conducting a predictive analytic project (Seng and Ang 2019). It is time-consuming, challenging, and boring to select, clean, and transform the relevant data.
- *Incompleteness*
  As was observed in the reviewed articles, the most important evaluation metric for researchers is the accuracy of the predicted model. The accuracy of models is restricted by the exhaustivity and accuracy of the data being used. Because the analytical algorithms try to build models based on the current data, the inadequacy in the records may lead to deficiencies in the model (Akbari et al. 2017). Equally, the evolved version of the model may not embody sufficient statistics to be able to spot adequate precious sentinel predictive patterns. Reducing incompleteness can be considered as an interesting open issue for future research.
- *Leadership*
  To manage challenges, successful enterprises in a data-driven era have made teams set goals, modulate attainments, and ask appropriate questions that can be answered by data insights. The big data power, besides its technical approach, can use human sight and vision. Having the vision and capability of talking about future opportunities and trends, leaders will be capable of acting and motivating their teams to attain their goals effectively (Shroff 2017; Courtney 2018). So, considering the role of enterprisers, leadership in BDPA is another direction for future works.
- *Adoption*
  Clearly, the harder a technology is to be applied, the less likely it is to be adopted by the end-users. No need to say, to satisfy this challenge, using predictive analytic solutions is demanding, as they are standalone tools. On the other hand, to apply it, users are forced to alter from their initial business applications to predictive analytic solutions. Besides, scaling and deploying traditional predictive tools is difficult, making updating a painstaking process (Al-Qirim et al. 2017; Raguseo 2018). Therefore, this can be a challenge and an opportunity for the next studies.
- *Trust*
  In industry, a lack of trust in big data seems to be a critical case. The willingness to put reliance on another one is trust. The building block of trust formation has many metrics: subjective reasons, like predispositions of an individual, and objective reasons. In the big data domain, trust can be related to the quality of data that is raised by the processes of data quality assurance. Predictive analytics that is based on low-quality data will not be reliable (Rubin et al. 2017). Thus, increasing the quality of data is a challenge and can be another path for further research.

7 Conclusion and limitation

This article provided a systematic review of BDPA. Predictive analytics and big data were investigated, and the relationship between them was elaborated. The research methodology was described, and 109 principal studies were chosen out of 1130 primary articles from our search query, which were published between the years 2014 and 2023. According to

this SLR, most articles were published in 2021 and 2022, and the least of them were published in 2016. ScienceDirect, with publishing 34%, outnumbered other journals in publishing BDPA articles. However, SAGE and ACM with 1% comprised the least number of published articles. 109 articles were sorted into seven categories in accordance with their applications, namely industrial, e-commerce, smart healthcare, smart agriculture, smart city, ICT, and weather. For each of these classes, numerous characteristics were reviewed and compared. Accuracy was the main concern of the researchers because it had the highest percentage, 19%, among evaluation metrics. Time, validity, and resource utilization had the least importance in the reviewed articles. For the evaluation method, 52% of studies implemented a simulation, 30% of studies used a real testbed environment, 11% brought a prototype up, 6% of them designed a new model, and 1% of studies used formal methods. Moreover, to design and develop more efficient architectures, frameworks, and algorithms in BDPA in the future, a detailed description of the open issues and challenges of BDPA was presented. Ultimately, considering RQ5, to create a more functional BDPA, some open issues and future challenges, such as streaming data sources, user privacy, multiple information sources, vertical domain applications, scalability, structured data vs. unstructured, incompleteness, leadership, adoption, and trust ought to be addressed. Moreover, the practical implications of this SLR extend to its role as a comprehensive director in the domain of BDPA. This work serves as a stimulus for researchers, markets, and industry to have BDPA implemented in their plans and improve the accuracy of their work. We hope that, in the domain of BDPA researchers cooperate to make progress and investigate this field further. This SLR provided significant insights into ongoing research trends, industry integration, and policy implications. Through delineating key research domains, showcasing potential applications in diverse sectors, and delineating necessary technological innovations to tackle current obstacles, this review served as a valuable resource for stakeholders. Furthermore, it emphasized the ethical imperatives and cautious implementation of BDPA tools, aiming to facilitate a shift towards proactive predictive analytics in tandem with adherence to regulatory standards and protection of individual privacy. Nevertheless, the thorough exploration of BDPA outlined in this study is accompanied by a few limitations, including the following:

- *Language*: Non-English articles have been omitted.
- *Research domain*: Different sources have covered BDPA. JCR-indexed journals and famous conferences have been included to attain competency fully. The nationally published articles are removed. In addition, book chapters, survey articles, and editorial articles have not been considered.
- *Study and publication bias*: Google Scholar, Springer, IEEE, ScienceDirect, SAGE, ACM, World Scientific, Emerald, Wiley, Hindawi, and Taylor & Francis were selected as electronic databases. The statistics show that these electronic databases supply the foremost connected and valid articles. Nevertheless, the choice of all applicable studies cannot be warranted. Some proper articles were excluded due to the mentioned processes.
- *Taxonomy*: The articles are classified into seven categories based on the application: Industrial, e-commerce, smart healthcare, smart agriculture, smart city, ICT, and weather. However, it can be classified in broader terms.
- *Study queries*: In order to expand this study, five questions are chosen. However, other questions may be regarded.

Author Contributions *Amirhossein Jamarani:* Conceptualization, Writing—Original Draft, Investigation, Visualization. *Saeid Haddadi*: Conceptualization, Methodology, Review & Editing, Investigation, Visualization. *Raheleh Sarvizadeh*: Investigation, Visualization. *Mostafa Haghi Kashani*: Supervision, Project administration, Investigation, Validation. *Mohammad Akbari*: Validation, Writing—Review & Editing. *Saeed Moradi*: Validation, Writing—Review & Editing.

Funding No funding is provided for the preparation of the manuscript.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Competing interests The authors have no relevant financial or nonfinancial interests to disclose.

Ethics approval This article does not contain any studies with human participants.

References

Abeza G, O'Reilly N, Nadeau J, Abdourazakou Y (2022) Big data in professional sport: the perspective of practitioners in the NFL, MLB, NBA, and NHL," Journal of Strategic Marketing, pp. 1–21, 2022, https://doi.org/10.1080/0965254X.2022.2108881

Abkenar SB, Kashani MH, Akbari M, Mahdipour E (2020) Twitter spam detection: A systematic review. arXiv preprint arXiv:2011.14754

Aboutorabi SH, Rezapour M, Moradi M, Ghadiri N (2015) Performance evaluation of SQL and MongoDB databases for big e-commerce data," in 2015 International Symposium on Computer Science and Software Engineering (CSSE), 18–19 Aug. 2015 2015, pp. 1–7, https://doi.org/10.1109/CSICSSE.2015.7369245

Acciarini C, Cappa F, Boccardelli P, Oriani R (2023) How can organizations leverage big data to innovate their business models? A systematic literature review,". Technovation 123:102713. https://doi.org/10.1016/j.technovation.2023.102713

Agarwal P, Tang J, Narayanan ANL, Zhuang J (2020) Big Data and Predictive Analytics in Fire Risk Using Weather Data,". Risk Anal. https://doi.org/10.1111/risa.1348040(7),pp.1438-1449

Ahmadi Z, Haghi Kashani M, Nikravan M, Mahdipour E (2021) Fog-based healthcare systems: A systematic review,". Multimed Tools Appl 80(30):36361–36400. https://doi.org/10.1007/s11042-021-11227-x

Akbari M, Hu X, Wang F, Chua T (2017) Wellness Representation of Users in Social Media: Towards Joint Modelling of Heterogeneity and Temporality. IEEE Trans Knowl Data Eng 29(10):2360–2373. https://doi.org/10.1109/TKDE.2017.2722411

Alam M, Amjad M (2019) Weather forecasting using parallel and distributed analytics approaches on big data clouds,". J Statistics Manag Syst 22(4):791–799. https://doi.org/10.1080/09720510.2019.1609559

AlFarraj O, AlZubi A, Tolba A (2019) Optimized feature selection algorithm based on fireflies with gravitational ant colony algorithm for big data predictive analytics,". Neural Comput Appl 31(5):1391–1403. https://doi.org/10.1007/s00521-018-3612-0

Ali A, Pasha MF, Fang OH, Khan R, Almaiah MA, Al Hwaitat AK (2022) Big Data Based Smart Blockchain for Information Retrieval in Privacy-Preserving Healthcare System," in Big Data Intelligence for Smart Applications, Y. Baddi, Y. Gahi, Y. Maleh, M. Alazab, and L. Tawalbeh Eds. Cham: Springer International Publishing, 2022, pp. 279–296. https://doi.org/10.1007/978-3-030-87954-9_13

Aljawarneh S, Lara JA, Yassein MB (2020) A visual big data system for the prediction of weather-related variables: Jordan-Spain case study," Multimed Tools Appl, 2020/10/01 2020, https://doi.org/10.1007/s11042-020-09848-9

Al-Qirim N, Tarhini A, Rouibah K (2017) Determinants of big data adoption and success. In Proceedings of the 1st International Conference on Algorithms, Computing and Systems, pp. 88-92

Alrumiah SS, Hadwan M (2021) Implementing Big Data Analytics in E-Commerce: Vendor and Customer View. IEEE Access 9:37281–37286. https://doi.org/10.1109/ACCESS.2021.3063615

Al-Sai ZA, Abdullah R, Husin MH (2020) Critical Success Factors for Big Data: A Systematic Literature Review. IEEE Access 8:118940–118956. https://doi.org/10.1109/ACCESS.2020.3005461

Amirian P, Basiri A, Morley J (2016) Predictive analytics for enhancing travel time estimation in navigation apps of Apple, Google, and Microsoft. In Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science, pp. 31–36

Anagnostopoulos I, Zeadally S, Exposito E (2016) Handling big data: research challenges and future directions. The J Supercomputing 72(4):1494–1516. https://doi.org/10.1007/s11227-016-1677-z

Arun Kumar S, Venkatesulu M (2019) Gramian matrix data collection-based random forest classification for predictive analytics with big data,". Soft Comput 23(18):8621–8631. https://doi.org/10.1007/s00500-019-04014-2

Awotunde JB, Jimoh RG, Ogundokun RO, Misra S, Abikoye OC (2022) Big data analytics of iot-based cloud system framework: Smart healthcare monitoring systems. In Artificial intelligence for cloud and edge computing. Springer International Publishing, Cham, pp 181–208

Babar M, Tariq MU, Alshehri MD, Ullah F, Uddin MI (2022) Smart teledentistry healthcare architecture for medical big data analysis using IoT-enabled environment,". Sustainable Computing: Informatics and Systems 35:100719. https://doi.org/10.1016/j.suscom.2022.100719

Balbin PPF, Barker JCR, Leung CK, Tran M, Wall RP, Cuzzocrea A (2020) Predictive analytics on open big data for supporting smart transportation services,". Procedia Computer Science 176:3009–3018. https://doi.org/10.1016/j.procs.2020.09.202

Balduini M, Bozzon A, Valle ED, Huang Y, Houben G (2014) Recommending Venues Using Continuous Predictive Social Media Analytics. IEEE Internet Comput 18(5):28–35. https://doi.org/10.1109/MIC.2014.84

Banumathi S, Aloysius A (2017) Predictive analytics concepts in big data- a survey. Int J Adv Res Computer Sci, Big Data, Predictive Analytics, Big Data Applications, Predictive Approaches, Challenges 8(8):4. https://doi.org/10.26483/ijarcs.v8i8.4628

Bazzaz Abkenar S, Haghi Kashani M, Mahdipour E, Jameii SM (2021) Big data analytics meets social media: A systematic review of techniques, open issues, and future directions,". Telematics Informatics 57:101517. https://doi.org/10.1016/j.tele.2020.101517

Bazzaz Abkenar S, Haghi Kashani M, Akbari M, Mahdipour E (2023) Learning textual features for Twitter spam detection: A systematic literature review,". Expert Syst Appl 228:120366. https://doi.org/10.1016/j.eswa.2023.120366

Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K (2015) Big Data Analytics in Healthcare,". Biomed Res Int 2015:16. https://doi.org/10.1155/2015/370194

Bendre MR, Thool RC, Thool VR (2015) Big data in precision agriculture: Weather forecasting for future farming," in 2015 1st International Conference on Next Generation Computing Technologies (NGCT), 4–5 Sept. 2015 2015, pp. 744–750, https://doi.org/10.1109/NGCT.2015.7375220

Bendre MR, Thool RC, Thool VR (2016) Big Data in Precision Agriculture Through ICT: Rainfall Prediction Using Neural Network Approach. In: Satapathy S, Bhatt Y, Joshi A, Mishra D (eds) Proceedings of the International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing, vol 438. Springer, Singapore. https://doi.org/10.1007/978-981-10-0767-5_19

Bhuimali A, Aithal S, Paul PK (2018) Business Informatics: With Special Reference To Big Data As An Emerging Area: A Basic Review. Int J Recent Researches inScience, Eng Technol, vol. 6, https://doi.org/10.5281/zenodo.1249786

Bianchi DG (2019) Politics and big data. Nowcasting and forecasting elections with social media," Contemporary Italian Politics, pp. 1–2, 2019, https://doi.org/10.1080/23248823.2019.1619298

Biesialska K, Franch X, Muntés-Mulero V (2021) Big Data analytics in Agile software development: A systematic mapping study,". Inf Software Technol 132:106448. https://doi.org/10.1016/j.infsof.2020.106448

Bradlow ET, Gangwar M, Kopalle P, Voleti S (2017) The Role of Big Data and Predictive Analytics in Retailing,". J Retail 93(1):79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain,". J Syst Software 80(4):571–583. https://doi.org/10.1016/j.jss.2006.07.009

Breur T (2016) Statistical Power Analysis and the contemporary "crisis" in social sciences. J Marketing Analytics 4(2):61–65. https://doi.org/10.1057/s41270-016-0001-3

Casado R, Younas M (2015) Emerging trends and technologies in big data processing,". Concurrency and Computation: Practice and Experience 27(8):2078–2091. https://doi.org/10.1002/cpe.3398

Chang V (2021) An ethical framework for big data and smart cities,". Technol Forecast Soc Chang 165:120559. https://doi.org/10.1016/j.techfore.2020.120559

Chen P-T (2018) Medical big data applications: Intertwined effects and effective resource allocation strategies identified through IRA-NRM analysis,". Technol Forecasting Social Change 130:150–164. https://doi.org/10.1016/j.techfore.2018.01.033

Chen H (2018a) Personalized recommendation system of e-commerce based on big data analysis. J Interdisciplinary Mathematics 21(5):1243–1247

Chen S (2021) Analysis of Customization Strategy for E-Commerce Operation Based on Big Data,". Wireless Commun Mobile Computing 2021:6626480. https://doi.org/10.1155/2021/6626480

Cheung M, She J, Wang N (2018) Characterizing User Connections in Social Media through User-Shared Images. IEEE Transactions on Big Data 4(4):447–458. https://doi.org/10.1109/TBDATA.2017.2762719

Chin J, Callaghan V, Lam I (2017) Understanding and personalising smart city services using machine learning, The Internet-of-Things and Big Data," in 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), 19–21 June 2017 2017, pp. 2050–2055, https://doi.org/10.1109/ISIE.2017.8001570

Christobel TP, Kamalakannan T (2020) Predictive analysis in Gestational Diabetic Mellitus (GDM) using HCNN-LSTM/DPNN (Big Data)," in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 3–5 Dec. 2020 2020, pp. 407–413, https://doi.org/10.1109/ICISS49785.2020.9315888

Cichosz SL, Johansen MD, Hejlesen O (2016) Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications,". J Diabetes Sci Technol 10(1):27–34. https://doi.org/10.1177/1932296815611680

Çoban S, Gökalp MO, Gökalp E, Eren PE, Koçyiğit A (2018) [WiP] Predictive maintenance in healthcare services with big data technologies. In2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA). IEEE, pp. 93–98

Coker F (2015) Pulse: Understanding the vital signs of your business. BookBaby

Courtney SJ (2018) Privatising educational leadership through technology in the Trumpian era,". J Educ Administration History 50(1):23–31. https://doi.org/10.1080/00220620.2017.1395826

Cui Q et al (2019) Big Data Analytics and Network Calculus Enabling Intelligent Management of Autonomous Vehicles in a Smart City. IEEE Internet Things J 6(2):2021–2034. https://doi.org/10.1109/JIOT.2018.2872442

Das S, Singh P, Puri G (2017) A Predictive Analytics Model for Maximising Profit in e-commerce Companies,". E-Commerce for Future & Trends, STM J 4:19–32

Das S, Namasudra S (2022) A Lightweight and Anonymous Mutual Authentication Scheme for Medical Big Data in Distributed Smart Healthcare Systems," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–12, 2022, https://doi.org/10.1109/TCBB.2022.3230053

De S, Maity A, Goel V, Shitole S, Bhattacharya A (2017) Predicting the popularity of instagram posts for a lifestyle magazine using deep learning," in 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), 7–8 April 2017 2017, pp. 174–177, https://doi.org/10.1109/CSCITA.2017.8066548

Delfmann P (2019) Predictive Analytics of Winter Sports Processes Using Probabilistic Finite Automata. In: Bergener K, Räckers M, Stein A (eds) The Art of Structuring: Bridging the Gap Between Information Systems Research and Practice. Springer International Publishing, Cham, pp 179–189

Dubey R, Luo Z, Gunasekaran A, Akter S, Hazen BT, Douglas MA (2018) Big data and predictive analytics in humanitarian supply chains. The Int J Logistics Manag 29(2):485–512. https://doi.org/10.1108/IJLM-02-2017-0039

Dubey R et al (2018) Examining the role of big data and predictive analytics on collaborative performance in context to sustainable consumption and production behaviour,". J Cleaner Production 196:1508–2152. https://doi.org/10.1016/j.jclepro.2018.06.097

Dubey R et al (2019) Can big data and predictive analytics improve social and environmental sustainability?,". Technol Forecasting Social Change 144:534–545. https://doi.org/10.1016/j.techfore.2017.06.020

El Azzaoui A, Singh SK, Park JH (2021) SNS Big Data Analysis Framework for COVID-19 Outbreak Prediction in Smart Healthy City,". Sustainable Cities Soc 71:102993. https://doi.org/10.1016/j.scs.2021.102993

Etemadi M et al (2023) A systematic review of healthcare recommender systems: Open issues, challenges, and techniques,". Expert Systems with App 213:118823. https://doi.org/10.1016/j.eswa.2022.118823

Fathi M, Haghi Kashani M, Jameii SM, Mahdipour E (2021) Big Data Analytics in Weather Forecasting: A Systematic Review, Archives of Computational Methods in Engineering, https://doi.org/10.1007/s11831-021-09616-4

Fawcett T (2006) An introduction to ROC analysis,". Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fiadino P, Casas P, D'Alconzo A, Schiavone M, Baer A (2016) Grasping Popular Applications in Cellular Networks With Big Data Analytics Platforms. IEEE Trans Netw Serv Manage 13(3):681–695. https://doi.org/10.1109/TNSM.2016.2558839

Gaedke Nomura AT, de Abreu Almeida M, Johnson S, Pruinelli L (2021) Pain Information Model and Its Potential for Predictive Analytics: Applicability of a Big Data Science Framework,". J Nursing Scholarship 53(3):315–322. https://doi.org/10.1111/jnu.12648

Georgiadis G, Poels G (2022) Towards a privacy impact assessment methodology to support the requirements of the general data protection regulation in a big data analytics context: A systematic literature review,". Comput Law Secur Rev 44:105640. https://doi.org/10.1016/j.clsr.2021.105640

Ghani NA, Hamid S, Targio Hashem IA, Ahmed E (2019) Social media big data analytics: A survey,". Comput Human Behav 101:417–428. https://doi.org/10.1016/j.chb.2018.08.039

Gharajeh MS (2018) Chapter Eight - Biological Big Data Analytics," in *Advances in Computers*, vol. 109, P. Raj and G. C. Deka Eds.: Elsevier, pp. 321–355. https://doi.org/10.1016/bs.adcom.2017.08.002

Gong Y, Fang Y, Guo Y (2016) Private Data Analytics on Biomedical Sensing Data via Distributed Computation. IEEE/ACM Trans Comput Biol Bioinf 13(3):431–444. https://doi.org/10.1109/TCBB.2016.2515610

Gunasekaran A et al (2017) Big data and predictive analytics for supply chain and organizational performance,". J Business Res 70:308–317. https://doi.org/10.1016/j.jbusres.2016.08.004

Guo H, Xu L (2022) Research on the application of big data visualization technology in urban road congestion," Eur J Remote Sensing, pp. 1–12, 2022, https://doi.org/10.1080/22797254.2022.2147448

Haghi Kashani M, Rahmani AM, Jafari Navimipour N (2020) Quality of service-aware approaches in fog computing,". Int J Communication Syst 33(8):e4340. https://doi.org/10.1002/dac.4340

Haghi Kashani M, Madanipour M, Nikravan M, Asghari P, Mahdipour E (2021) A systematic review of IoT in healthcare: Applications, techniques, and trends,". J Network Comput Appl 192:103164. https://doi.org/10.1016/j.jnca.2021.103164

Haitao S (2020) Big data analysis of e-commerce loan risk of college students in the context of network finance,". Inf Syst e-Business Manag 18(3):439–454. https://doi.org/10.1007/s10257-019-00424-9

Hajiali M (2020) Big data and sentiment analysis: A comprehensive and systematic literature review," Concurrency and Computation: Practice and Experience, vol. n/a, no. n/a, p. e5671, 2020/04/19 2020, https://doi.org/10.1002/cpe.5671

Han Q, Liu D, Hu C (2023) Risk Analysis and Establishment of Supervision System of Internet Finance Based on Big Data Era,". Wireless Commun Mobile Computing 2023:5134720. https://doi.org/10.1155/2023/5134720

Harris SL, May JH, Vargas LG (2016) Predictive analytics model for healthcare planning and scheduling,". Eur J Operational Res 253(1):121–131. https://doi.org/10.1016/j.ejor.2016.02.017

Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA (2014) Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications,". Int J Product Econ 154:72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

Hendri HJ, Sulaiman H (2018) Predictive Modeling for Dengue Patient's Length of Stay (LoS) Using Big Data Analytics (BDA). InRecent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT 2017). Springer International Publishing. pp. 12–19

Himeur Y et al (2023) AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives,". Artif Intell Rev 56(6):4929–5021. https://doi.org/10.1007/s10462-022-10286-2

Homer ML, Palmer NP, Fox KP, Armstrong J, Mandl KD (2017) Predicting Falls in People Aged 65 Years and Older from Insurance Claims,". The Am J Med 130(6):744.e17-744.e23. https://doi.org/10.1016/j.amjmed.2017.01.003

Hu H, Wen Y, Chua T, Li X (2014) Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access 2:652–687. https://doi.org/10.1109/ACCESS.2014.2332453

Huang T, Bergman D, Gopal R (2019) Predictive and Prescriptive Analytics for Location Selection of Add-on Retail Products,". Prod Oper Manag 28(7):1858–1877. https://doi.org/10.1111/poms.13018
Huang Z, Mao C, Guan S, Tang H, Chen G, Liu Z (2023) Security threshold setting algorithm of distributed optical fiber monitoring and sensing system based on big data in smart city,". Soft Comput 27(8):5147–5157. https://doi.org/10.1007/s00500-021-06212-3
Ikegwu AC, Nweke HF, Anikwe CV, Alo UR, Okonkwo OR (2022) Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions,". Clust Comput 25(5):3343–3387. https://doi.org/10.1007/s10586-022-03568-5
Jeble S (2018) Impact of big data and predictive analytics capability on supply chain sustainability. The Int J Logistics Manag 29(2):513–538. https://doi.org/10.1108/IJLM-05-2017-0134
Jiang S, Qian X, Mei T, Fu Y (2016) Personalized Travel Sequence Recommendation on Multi-Source Big Social Media. IEEE Transactions on Big Data 2(1):43–56. https://doi.org/10.1109/TBDATA.2016.2541160
Kaffash S, Nguyen AT, Zhu J (2021) Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis,". Int J Product Economics 231:107868. https://doi.org/10.1016/j.ijpe.2020.107868
Kannan N, Sivasubramanian S, Kaliappan M, Vimal S, Suresh A (2018) Predictive big data analytic on demonetization data using support vector machine," Cluster Computing, 2018/03/14 2018, https://doi.org/10.1007/s10586-018-2384-8
Karimi Y, Haghi Kashani M, Akbari M, Mahdipour E (2021) Leveraging big data in smart cities: A systematic review. Concurrency and Computation: Practice and Experience 33(21):e6379. https://doi.org/10.1002/cpe.6379
Kashani MH, Ahmadzadeh A, Mahdipour E (2020) Load balancing mechanisms in fog computing: A systematic review. arXiv preprint arXiv:2011.14706
Kashani MH, Mahdipour E (2023) Load Balancing Algorithms in Fog Computing. IEEE Trans Serv Comput 16(2):1505–1521. https://doi.org/10.1109/TSC.2022.3174475
Keswani B, Mohapatra AG, Keswani P, Khanna A, Gupta D, Rodrigues J (2020) Improving weather dependent zone specific irrigation control scheme in IoT and big data enabled self driven precision agriculture mechanism,". Enterprise Information Systems 14(9–10):1494–1515. https://doi.org/10.1080/17517575.2020.1713406
Kezunovic M et al. (2017) Predicting spatiotemporal impacts of weather on power systems using big data science," in Data Science and Big Data: An Environment of Computational Intelligence: Springer, 2017, pp. 265–299. https://doi.org/10.1007/978-3-319-53474-9_12
Khan M, Babar M, Ahmed SH, Shah SC, Han K (2017) Smart city designing and planning based on big data analytics,". Sustain Cities Soc 35:271–279. https://doi.org/10.1016/j.scs.2017.07.012
Khatibi T, Kheyrikoochaksarayee N, Sepehri MM (2019) Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features,". Arch Gynecol Obstet 300(6):1565–1582. https://doi.org/10.1007/s00404-019-05325-3
Khine KLL, Nyunt TTS (2019) Predictive Big Data Analytics Using Multiple Linear Regression Model," in Big Data Analysis and Deep Learning Applications, Singapore, T. T. Zin and J. C.-W. Lin, Eds., 2019// 2019: Springer Singapore, pp. 9–19. https://doi.org/10.1007/978-981-13-0869-7_2
Khoshniat N, Jamarani A, Ahmadzadeh A, Haghi Kashani M, Mahdipour E (2023) Nature-inspired metaheuristic methods in software testing," Soft Computing, 2023/06/08 2023, https://doi.org/10.1007/s00500-023-08382-8
Kitchenham B (2004) Procedures for performing systematic reviews. Keele, UK, Keele University 33(2004):1–26
Kitchenham B, Pearl Brereton O, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering – A systematic literature review,". Inf Software Technol 51(1):7–15. https://doi.org/10.1016/j.infsof.2008.09.009
Kodidala VSSJ, Akkala S, Madupoju SK, Dasara VSST, Juvvadi M, Thangadurai N (2021) Big Data analysis of demand side management for Industrial IOT applications,". Materials Today: Proceedings 46:8313–8319. https://doi.org/10.1016/j.matpr.2021.03.301
Kolajo T, Daramola O, Adebiyi A (2019) Big data stream analysis: a systematic literature review,". J Big Data 6(1):47. https://doi.org/10.1186/s40537-019-0210-7
Kong X, Jiang X, Zhang B, Yuan J, Ge Z (2022) Latent variable models in the era of industrial big data: Extension and beyond,". Annu Rev Control 54:167–199. https://doi.org/10.1016/j.arcontrol.2022.09.005
Krumeich J, Jacobi S, Werth D, Loos P (2014) Big Data Analytics for Predictive Manufacturing Control - A Case Study from Process Industry," in 2014 IEEE International Congress on Big Data, 27 June-2 July 2014 2014, pp. 530–537, https://doi.org/10.1109/BigData.Congress.2014.83

Kuang K, Jiang M, Cui P, Luo H, Yang S (2018) Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach. IEEE Transactions on Big Data 4(4):487–501. https://doi.org/10.1109/TBDATA.2017.2734102

Kumaresan G, Rajakumar P (2015) Predictive Analytics Using Big Data: A Survey. Int J Manag Inf Technol Eng 3:61–68

Leary DEO (2015) Big Data and Privacy: Emerging Issues. IEEE Intell Syst 30(6):92–96. https://doi.org/10.1109/MIS.2015.110

Lee CKH (2017) A GA-based optimisation model for big data analytics supporting anticipatory shipping in Retail 4.0,". Int J Product Res 55(2):593–605. https://doi.org/10.1080/00207543.2016.1221162

Li F, Li Y (2022) Big Data Mining Method of E-Commerce Consumption Pattern Based on Mobile Platform,". Security Communication Networks 2022:3991135. https://doi.org/10.1155/2022/3991135

Li C, Niu B (2020) Design of smart agriculture based on big data and Internet of things. Int J Distrib Sens Netw 16(5):1550147720917065

Li Y et al (2016) Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery. IEEE Trans Knowl Data Eng 28(8):1986–1999. https://doi.org/10.1109/TKDE.2016.2559481

Li X, Liu H, Wang W, Zheng Y, Lv H, Lv Z (2022) Big data analysis of the Internet of Things in the digital twins of smart city based on deep learning,". Futur Gener Comput Syst 128:167–177. https://doi.org/10.1016/j.future.2021.10.006

Lin L, Pan L, Liu S (2022) A Cost-Effective Framework for Running Industrial Big Data Analysis Applications in Public Clouds. IEEE Internet Things J 9(13):10554–10562. https://doi.org/10.1109/JIOT.2021.3122196

Lin J, Niu J, Li H (2017) PCD: A privacy-preserving predictive clinical decision scheme with E-health big data based on RNN. In2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, pp. 808–813

Liu JNK, Hu Y, He Y, Chan PW, Lai L (2015) Deep Neural Network Modeling for Big Data Weather Forecasting. In: Pedrycz W, Chen S-M (eds) Information Granularity, Big Data, and Computational Intelligence. Springer International Publishing, Cham, pp 389–408

Liu W (2021) Smart sensors, sensing mechanisms and platforms of sustainable smart agriculture realized through the big data analysis," Cluster Computing, 2021/05/12 2021, https://doi.org/10.1007/s10586-021-03295-3

Longhi L, Nanni M (2019) Car telematics big data analytics for insurance and innovative mobility services. J Ambient Intell Humanized Computing, 2019/12/14 2019, https://doi.org/10.1007/s12652-019-01632-4

Lv Z, Chen D, Lv H (2022) Smart city construction and management by digital twins and BIM big data in COVID-19 scenario," ACM Transactions on Multimidia Computing Communications and Applications, 2022 https://doi.org/10.1145/3529395

Madan S, Kumar P, Rawat S, Choudhury T (2018) Analysis of Weather Prediction using Machine Learning & Big Data," in 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 22–23 June 2018 2018, pp. 259–264, https://doi.org/10.1109/ICACCE.2018.8441679

Mallika C, Selvamuthukumaran S (2022) Technological perspective on precision medicine in the context of big data—a review. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1. Springer Nature Singapore, Singapore, pp. 553–564

Mehta N, Pandit A, Shukla S (2019) Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study,". J Biomed Informatics 100:103311. https://doi.org/10.1016/j.jbi.2019.103311

Melgar-García L et al (2022) A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture,". Neurocomputing 500:268–278. https://doi.org/10.1016/j.neucom.2021.06.101

Mikalef P, Pappas IO, Krogstie J, Giannakos M (2018) Big data analytics capabilities: a systematic literature review and research agenda,". Inf Syst e-Business Manag 16(3):547–578. https://doi.org/10.1007/s10257-017-0362-y

Miles DA (2017) Market Research and Predictive Analytics: Using Analytics to Measure Customer and Marketing Behavior in Business Ventures. In: Carayannis EG, Sindakis S (eds) Analytics, Innovation, and Excellence-Driven Enterprise Sustainability. Palgrave Macmillan US, New York, pp 77–108

Mishra D (2019) Organizational capabilities that enable big data and predictive analytics diffusion and organizational performance. Manag Decis 57(8):1734–1755. https://doi.org/10.1108/MD-03-2018-0324

Mohamed A, Najafabadi MK, Wah YB, Zaman EAK, Maskat R (2020) The state of the art and taxonomy of big data analytics: view from new big data framework,". Artif Intell Rev 53(2):989–1037. https://doi.org/10.1007/s10462-019-09685-9

Montero O, Crespo Y, Piatini M (2021) Big data quality models: a systematic mapping study. In Quality of Information and Communications Technology: 14th International Conference, QUATIC 2021. Proceedings 14 2021. Algarve, Portugal, Springer International Publishing. pp. 416–430

Moreira-Matias L, Gama J, Ferreira M, Mendes-Moreira J, Damas L (2016) Time-evolving O-D matrix estimation using high-speed GPS data streams,". Expert Syst Appl 44:275–288. https://doi.org/10.1016/j.eswa.2015.08.048

Morris KJ, Egan SD, Linsangan JL, Leung CK, Cuzzocrea A, Hoi CSH (2018) Token-Based Adaptive Time-Series Prediction by Ensembling Linear and Non-linear Estimators: A Machine Learning Approach for Predictive Analytics on big Stock Data," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 17–20 Dec. 2018 2018, pp. 1486–1491, https://doi.org/10.1109/ICMLA.2018.00242

Mortaheb R, Jankowski P (2023) Smart city re-imagined: City planning and GeoAI in the age of big data,". J Urban Management 12(1):4–15. https://doi.org/10.1016/j.jum.2022.08.001

Mujeeb S, Javaid N (2019) ESAENARX and DE-RELM: Novel schemes for big data predictive analytics of electricity load and price,". Sustain Cities Soc 51:101642. https://doi.org/10.1016/j.scs.2019.101642

Muthukrishnan S, Govindasamy M, Mustapha M (2017) Systematic mapping review on student's performance analysis using big data predictive model. J Fundamental Appl Sci 9(4S):730–758

Naghib A, Jafari Navimipour N, Hosseinzadeh M, Sharifi A (2022) A comprehensive and systematic literature review on the big data management techniques in the internet of things," Wireless Networks, 2022/11/15 2022, https://doi.org/10.1007/s11276-022-03177-5

Nallathamby R, Robin R, Miriam D (2021) Optimizing appointment scheduling for out patients and income analysis for hospitals using big data predictive analytics," J Ambient Intell Human Comput, vol. 12, 06/01 2021, https://doi.org/10.1007/s12652-020-02118-4

Nathali Silva B, Khan M, Han K (2017) Big Data Analytics Embedded Smart City Architecture for Performance Enhancement through Real-Time Data Processing and Decision-Making,". Wireless Commun Mobile Computing 2017:9429676. https://doi.org/10.1155/2017/9429676

Nemati S, Haghi Kashani M, Faghih Mirzaee R (2023) Comprehensive survey of ternary full adders: Statistics, corrections, and assessments,". IET Circuits Devices Syst 17(3):111–134. https://doi.org/10.1049/cds2.12152

Nestor DMJ, Ogudo KA (2018) Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time," in 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 6–7 Aug. 2018 2018, pp. 1–10, https://doi.org/10.1109/ICABCD.2018.8465476

Ng ST, Xu FJ, Yang Y, Lu M (2017) A Master Data Management Solution to Unlock the Value of Big Infrastructure Data for Smart, Sustainable and Resilient City Planning,". Procedia Engineering 196:939–947. https://doi.org/10.1016/j.proeng.2017.08.034

Nguyen T, Phung D, Dao B, Venkatesh S, Berk M (2014) Affective and Content Analysis of Online Depression Communities. IEEE Trans Affect Comput 5(3):217–226. https://doi.org/10.1109/TAFFC.2014.2315623

Nikravan M, Haghi Kashani M (2022) A review on trust management in fog/edge computing: Techniques, trends, and challenges,". J Network Comput Appl 204:103402. https://doi.org/10.1016/j.jnca.2022.103402

Nilashi M et al. (2023) How can big data and predictive analytics impact the performance and competitive advantage of the food waste and recycling industry?," Annals of Operations Research, 2023/03/25 2023, https://doi.org/10.1007/s10479-023-05272-y

Nobanee H, Shanti H, Aldhanhani H, Alblooshi A, Alali E (2022) Big data and credit risk assessment: a bibliometric review, current streams, and directions for future research. Cogent Economics & Finance 10(1):2132638. https://doi.org/10.1080/23322039.2022.2132638

Nural MV, Cotterell ME, Miller JA (2015) Using Semantics in Predictive Big Data Analytics," in 2015 IEEE International Congress on Big Data, 27 June-2 July 2015 2015, pp. 254–261, https://doi.org/10.1109/BigDataCongress.2015.43

Nural MV, Peng H, Miller JA (2017) Using meta-learning for model type selection in predictive big data analytics," in 2017 IEEE International Conference on Big Data (Big Data), 11–14 Dec. 2017 2017, pp. 2027–2036, https://doi.org/10.1109/BigData.2017.8258149

Nyce C (2007) "Predictive Analytics White Paper (PDF)", American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America

O'Donovan P, Leahy K, Bruton K, O'Sullivan DTJ (2015) Big data in manufacturing: a systematic mapping study,". J Big Data 2(1):20. https://doi.org/10.1186/s40537-015-0028-x
Ohiomah A, Andreev P, Benyoucef M (2017) A review of big data predictive analytics in information systems research. In Proceedings of the Conference on Information Systems Applied Research ISSN, Vol. 2167, p. 1508
Onal AC, Sezer OB, Ozbayoglu M, Dogdu E (2017) Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," in 2017 IEEE International Conference on Big Data (Big Data), 11–14 Dec. 2017 2017, pp. 2037–2046, https://doi.org/10.1109/BigData.2017.8258150
Oo MCM, Thein T (2019) An efficient predictive analytics system for high dimensional big data," J King Saud Univ – Comput Inf Sci, 2019/09/07/ 2019, https://doi.org/10.1016/j.jksuci.2019.09.001
Osinga SA, Paudel D, Mouzakitis SA, Athanasiadis IN (2022) Big data in agriculture: Between opportunity and solution,". Agric Syst 195:103298. https://doi.org/10.1016/j.agsy.2021.103298
Ouahilal M, Mohajir ME, Chahhou M, Mohajir BEE (2016) A comparative study of predictive algorithms for business analytics and decision support systems: Finance as a case study," in 2016 International Conference on Information Technology for Organizations Development (IT4OD), 30 March-1 April 2016 2016, pp. 1–6, https://doi.org/10.1109/IT4OD.2016.7479258
Oyekanlu E (2017) Predictive edge computing for time series of industrial IoT and large scale critical infrastructure based on open-source software analytic of big data," in 2017 IEEE International Conference on Big Data (Big Data), 11–14 Dec. 2017 2017, pp. 1663–1669, https://doi.org/10.1109/BigData.2017.8258103
Park D, Kim S, Lee J, Choo J, Diakopoulos N, Elmqvist N (2018) ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. IEEE Trans Visual Comput Graphics 24(1):361–370. https://doi.org/10.1109/TVCG.2017.2744478
Philip Chen CL, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,". Inf Sci 275:314–347. https://doi.org/10.1016/j.ins.2014.01.015
Poornima S, Pushpalatha M (2018) A survey of predictive analytics using big data with data mining. Int J Bioinform Res Appl 14(3):269–282. https://doi.org/10.1504/ijbra.2018.092697
Quan Z, Wang Z, Le Y, Yao B, Li K, Yin J (2019) An Efficient Framework for Sentence Similarity Modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(4):853–865. https://doi.org/10.1109/TASLP.2019.2899494
Raguseo E (2018) Big data technologies: An empirical investigation on their adoption, benefits and risks for companies,". Int J Inf Manag 38(1):187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008
Rahimi M, Songhorabadi M, Kashani MH (2020) Fog-based smart homes: A systematic review,". J Network Comput Appl 153:102531. https://doi.org/10.1016/j.jnca.2020.102531
Rahman MS, Reza H (2020) Systematic Mapping Study of Non-Functional Requirements in Big Data System," in 2020 IEEE International Conference on Electro Information Technology (EIT), 31 July-1 Aug. 2020 2020, pp. 025–031, https://doi.org/10.1109/EIT48999.2020.9208288
Ramsahai E et al (2023) Crime prediction in Trinidad and Tobago using big data analytics,". Int J Data Sci Anal 15(4):421–432. https://doi.org/10.1007/s41060-023-00386-9
Rathore MM, Paul A, Hong W-H, Seo H, Awan I, Saeed S (2018) Exploiting IoT and big data analytics: Defining Smart Digital City using real-time urban data,". Sustain Cities Soc 40:600–610. https://doi.org/10.1016/j.scs.2017.12.022
Rathore MM, Shah SA, Shukla D, Bentafat E, Bakiras S (2021) The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities. IEEE Access 9:32030–32052. https://doi.org/10.1109/ACCESS.2021.3060863
Ravi V, Kamaruddin S (2017) Big data analytics enabled smart financial services: opportunities and challenges. In Big Data Analytics: 5th International Conference, BDA 2017, Hyderabad, India. Proceedings 5 2017. Springer International Publishing, pp. 15–39
Reis C, Lopes A, Nouri AS (2022) Assessing urban heat island effects through local weather types in Lisbon's Metropolitan Area using big data from the Copernicus service,". Urban Climate 43:101168. https://doi.org/10.1016/j.uclim.2022.101168
Rodríguez-Mazahua L, Rodríguez-Enríquez C-A, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G (2015) A general perspective of Big Data: applications, tools, challenges and trends,". The J Supercomputing 72(8):3073–3113. https://doi.org/10.1007/s11227-015-1501-1
Roh H-J (2022) A study on securing model usefulness through geographical scalability testing of winter weather model developed with big traffic data,". Transp Plan Technol 45(6):473–497. https://doi.org/10.1080/03081060.2022.2132947

Rosati R et al (2023) From knowledge-based to big data analytic model: a novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0,". J Intell Manufacturing 34(1):107–121. https://doi.org/10.1007/s10845-022-01960-x

Roukh A, Fote FN, Mahmoudi SA, Mahmoudi S (2020) Big Data Processing Architecture for Smart Farming,". Procedia Computer Science 177:78–85. https://doi.org/10.1016/j.procs.2020.10.014

Rubin E, Argyris YA, Benbasat I (2017) Consumers' trust in price-forecasting recommendation agents. InHCI in Business, Government and Organizations. Supporting Business: 4th International Conference, HCIBGO 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada. Proceedings, Part II 4 2017. Springer International Publishing, pp. 71–80

Russell J (2015) Predictive analytics and child protection: Constraints and opportunities,". Child Abuse Neglect 46:182–189. https://doi.org/10.1016/j.chiabu.2015.05.022

Sabarina K, Priya N (2015) Lowering Data Dimensionality in Big Data for the Benefit of Precision Agriculture,". Procedia Comput Sci 48:548–554. https://doi.org/10.1016/j.procs.2015.04.134

Safa M, Pandian A, Gururaj HL, Ravi V, Krichen M (2023) Real time health care big data analytics model for improved QoS in cardiac disease prediction with IoT devices,". Health and Technol 13(3):473–483. https://doi.org/10.1007/s12553-023-00747-1

Saito T, Gupta S (2022) Big data applications with theoretical models and social media in financial management," Annals of Operations Research, https://doi.org/10.1007/s10479-022-05136-x

Sasidhar C, Mallikharjuna Rao N (2020) Integrated Big Data with Public Cloud: A Model for E-Commerce Retailer Market. In: Kumar A, Paprzycki M, Gunjan V (eds) ICDSMLA 2019. Lecture Notes in Electrical Engineering, vol 601. Springer, Singapore. https://doi.org/10.1007/978-981-15-1420-3_198

Scotti V (2017) Big data or big (privacy) problem? IEEE Instrum Meas Mag 20(5):23–26. https://doi.org/10.1109/MIM.2017.8036692

Seng JKP, Ang KL (2019) Multimodal Emotion and Sentiment Modeling From Unstructured Big Data: Challenges, Architecture, & Techniques. IEEE Access 7:90982–90998. https://doi.org/10.1109/ACCESS.2019.2926751

Shafi I, Malik Z, Din S, Jeon G, Ahmad J (2021) A computationally intelligent neural network-based nonlinear autoregressive exogenous balancing approach for real-time processing in industrial applications using big data,". Concurrency and Computation: Practice and Experience 33(22):e6382. https://doi.org/10.1002/cpe.6382

Shah HM, Gardas BB, Narwane VS, Mehta HS (2023) The contemporary state of big data analytics and artificial intelligence towards intelligent supply chain risk management: a comprehensive review. Kybernetes 52(5):1643–1697. https://doi.org/10.1108/K-05-2021-0423

Shao Z, Cai J, Wang Z (2018) Smart Monitoring Cameras Driven Intelligent Processing to Big Surveillance Video Data. IEEE Transactions on Big Data 4(1):105–116. https://doi.org/10.1109/TBDATA.2017.2715815

Sheikh Sofla M, Haghi Kashani M, Mahdipour E, Faghih Mirzaee R (2022) Towards effective offloading mechanisms in fog computing,". Multimed Tools Appl 81(2):1997–2042. https://doi.org/10.1007/s11042-021-11423-9

Shenoy S, Gorinevsky D (2015) Predictive Analytics for Extreme Events in Big Data," in 2015 IEEE First International Conference on Big Data Computing Service and Applications, 30 March-2 April 2015 2015, pp. 184–193, https://doi.org/10.1109/BigDataService.2015.66

Shrivastava A, Nayak CK, Dilip R, Samal SR, Rout S, Ashfaque SM (2023) Automatic robotic system design and development for vertical hydroponic farming using IoT and big data analysis,". Materials Today: Proceedings 80:3546–3553. https://doi.org/10.1016/j.matpr.2021.07.294

Shroff R (2017) Predictive Analytics for City Agencies: Lessons from Children's Services,". Big Data 5(3):189–196. https://doi.org/10.1089/big.2016.0052

Simpson SV, Nagarajan G (2021) An edge based trustworthy environment establishment for internet of things: an approach for smart cities," Wireless Networks, 2021/06/04 2021, https://doi.org/10.1007/s11276-021-02667-2

Singh RK, Agrawal S, Sahu A, Kazancoglu Y (2023) Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda. The TQM J 35(1):262–291. https://doi.org/10.1108/TQM-02-2021-0051

Sohrabi B (2019) A predictive analytics of physicians prescription and pharmacies sales correlation using data mining. Int J Pharma Healthcare Marketing 13(3):346–363. https://doi.org/10.1108/IJPHM-11-2017-0066

Songhorabadi M, Rahimi M, MoghadamFarid A, Haghi Kashani M (2023) Fog computing approaches in IoT-enabled smart cities,". J Netw Comput Appl 211:103557. https://doi.org/10.1016/j.jnca.2022.103557

Songhorabadi M, Rahimi M, Farid AM, Kashani MH (2020) Fog computing approaches in smart cities: a state-of-the-art review. arXiv preprint arXiv:2011.14732

Souza J, Leung CK, Cuzzocrea A (2020) "An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics," (in eng). Adv Inf Network Appl 1151:669–680. https://doi.org/10.1007/978-3-030-44041-1_59

Su C-J, Huang S-F (2018) Real-time big data analytics for hard disk drive predictive maintenance,". Comput Electric Eng 71:93–101. https://doi.org/10.1016/j.compeleceng.2018.07.025

Suguna S, Vithya M, Eunaicy JIC (2016) Big data analysis in e-commerce system using HadoopMapReduce," in 2016 International Conference on Inventive Computation Technologies (ICICT), 26–27 Aug. 2016 2016, vol. 2, pp. 1–6, https://doi.org/10.1109/INVENTIVE.2016.7824798

Sun J et al (2019) An Efficient and Scalable Framework for Processing Remotely Sensed Big Data in Cloud Computing Environments. IEEE Trans Geosci Remote Sens 57(7):4294–4308. https://doi.org/10.1109/TGRS.2018.2890513

Sun C, Gao R, Xi H (2014) Big data based retail recommender system of non E-commerce. In Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, pp. 1–7

Thakuriah P, Tilahun NY, Zellner M (2017) Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In: Thakuriah P, Tilahun N, Zellner M (eds) Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics. Springer International Publishing, Cham, pp 11–45

Tian L, Wang H, Zhou Y, Peng C (2018) Video big data in smart city: Background construction and optimization for surveillance video processing,". Futur Gener Comput Syst 86:1371–1382. https://doi.org/10.1016/j.future.2017.12.065

Tong H, Warren JL, Kang J, Li M (2022) Using multi-sourced big data to correlate sleep deprivation and road traffic noise: A US county-level ecological study," Environmental Research, p. 115029, 2022/12/08/ 2022, https://doi.org/10.1016/j.envres.2022.115029

Truong H (2018) Integrated Analytics for IIoT Predictive Maintenance Using IoT Big Data Cloud Systems," in 2018 IEEE International Conference on Industrial Internet (ICII), 21–23 Oct. 2018 2018, pp. 109–118, https://doi.org/10.1109/ICII.2018.00020

Tryapkin E, Shurova N (2020) The Use of Technology 'Big Data' and 'Predictive Analytics' in the Power Supply System of Railways," in VIII International Scientific Siberian Transport Forum, Cham, Z. Popovic, A. Manakov, and V. Breskich, Eds., 2020// 2020: Springer International Publishing, pp. 60–68. https://doi.org/10.1007/978-3-030-37916-2_7

Tsouli Fathi M, Ezziyyani M, Ezziyyani M, El Mamoune S (2020) Crop yield prediction using deep learning in Mediterranean Region. In Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health. Springer International Publishing, pp. 106–114

Tuladhar JG, Gupta A, Shrestha S, Bania UM, Bhargavi K (2018) Predictive analysis of e-commerce products. InIntelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017. Springer Singapore, pp. 279–289

Turet JG, Costa AP (2018) Big data analytics to improve the decision-making process in public safety: a case study in Northeast Brazil. In Decision Support Systems VIII: Sustainable Data-Driven and Evidence-Based Decision Support: 4th International Conference, ICDSST 2018, Proceedings 4 2018. Heraklion, Springer International Publishing, pp. 76–87

Velmurugan P, Kannagi A, Varsha M (2021) Superior fuzzy enumeration crop prediction algorithm for big data agriculture applications," Materials Today: Proceedings, 2021/03/13/ 2021, https://doi.org/10.1016/j.matpr.2021.02.578

Venkatesh R, Balasubramanian C, Kaliappan M (2019) Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique,". J Med Syst 43(8):272. https://doi.org/10.1007/s10916-019-1398-y

Wang Q, Mu Z (2022) Risk monitoring model of intelligent agriculture Internet of Things based on big data,". Sustainable Energy Technol Assess 53:102654. https://doi.org/10.1016/j.seta.2022.102654

Wang Y, Wang S, Yang B, Zhu L, Liu F (2020) Big data driven Hierarchical Digital Twin Predictive Remanufacturing paradigm: Architecture, control mechanism, application scenario and benefits,". J Cleaner Production 248:119299. https://doi.org/10.1016/j.jclepro.2019.119299

Weerakkody V, Sivarajah U, Mahroof K, Maruyama T, Lu S (2021) Influencing subjective well-being for business and sustainable development using big data and predictive regression analysis,". J Business Res 131:520–538. https://doi.org/10.1016/j.jbusres.2020.07.038

Wong KC (2016) Big Data Analytics in Genomics. Springer International Publishing. https://doi.org/10.1007/978-3-319-41279-5

Xiao M (2022) Supervision Strategy Analysis on Price Discrimination of E-Commerce Company in the Context of Big Data Based on Four-Party Evolutionary Game,". Computat Intell Neurosci 2022:2900286. https://doi.org/10.1155/2022/2900286

Xu Y, Liu H, Long Z (2020) A distributed computing framework for wind speed big data forecasting on Apache Spark,". Sustainable Energy Technol Assess 37:100582. https://doi.org/10.1016/j.seta.2019.100582

Yang Z, Ge Z (2022) On Paradigm of Industrial Big Data Analytics: From Evolution to Revolution. IEEE Trans Industr Inf 18(12):8373–8388. https://doi.org/10.1109/TII.2022.3190394

You Q, Cao L, Cong Y, Zhang X, Luo J (2015) A Multifaceted Approach to Social Multimedia-Based Prediction of Elections. IEEE Trans Multimedia 17(12):2271–2280. https://doi.org/10.1109/TMM.2015.2487863

Yu W, Dillon T, Mostafa F, Rahayu W, Liu Y (2020) A Global Manufacturing Big Data Ecosystem for Fault Detection in Predictive Maintenance. IEEE Trans Industr Inf 16(1):183–192. https://doi.org/10.1109/TII.2019.2915846

Zang J, You P (2022) An industrial IoT-enabled smart healthcare system using big data mining and machine learning," Wireless Networks, 2022/11/09 2022, https://doi.org/10.1007/s11276-022-03129-z

Zhang L, Yan Y, Xu W, Sun J, Zhang Y (2022) Carbon Emission Calculation and Influencing Factor Analysis Based on Industrial Big Data in the "Double Carbon" Era,". Computational Intell Neurosci 2022:2815940. https://doi.org/10.1155/2022/2815940

Zhang D, Pee LG, Pan SL, Cui L (2022b) Big data analytics, resource orchestration, and digital sustainability: A case study of smart city development,". Gov Inf Q 39(1):101626. https://doi.org/10.1016/j.giq.2021.101626

Zhao J, Wang J, Xing Z, Luan X, Jiang Y (2018) Weather and cycling: Mining big data to have an in-depth understanding of the association of weather variability with cycling on an off-road trail and an on-road bike lane,". Transportation Research Part a: Policy and Practice 111:119–135. https://doi.org/10.1016/j.tra.2018.03.001

Zhao W, Han S, Meng W, Sun D, Hu RQ (2019) BSDP: Big Sensor Data Preprocessing in Multi-Source Fusion Positioning System Using Compressive Sensing. IEEE Trans Veh Technol 68(9):8866–8880. https://doi.org/10.1109/TVT.2019.2929560

Zheng K, Zhang Z, Song B (2020) E-commerce logistics distribution mode in big-data context: A case analysis of JD.COM,". Ind Mark Manage 86:154–162. https://doi.org/10.1016/j.indmarman.2019.10.009

Zhuang W (2021) The Influence of Big Data Analytics on E-Commerce: Case Study of the US and China. Wireless Commun Mobile Computing 2021:2888673. https://doi.org/10.1155/2021/2888673

Authors and Affiliations

Amirhossein Jamarani[1] · Saeid Haddadi[2] · Raheleh Sarvizadeh[3] · Mostafa Haghi Kashani[3] · Mohammad Akbari[4] · Saeed Moradi[5]

✉ Mostafa Haghi Kashani
mh.kashani@qodsiau.ac.ir

Amirhossein Jamarani
C00550518@louisiana.edu

Saeid Haddadi
saeid.haddadi@srbiau.ac.ir

Raheleh Sarvizadeh
r.savizadeh@gmail.com

Mohammad Akbari
akbari.ma@aut.ac.ir

Saeed Moradi
saeed-moradi@srbiau.ac.ir

[1] The Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, LA, USA

[2] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[3] Department of Computer Engineering, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran

[4] Department of Computer Science, Amirkabir University of Technology, Tehran, Iran

[5] Department of Computer Engineering, Maku Branch, Islamic Azad University, Maku, Iran