Evaluating the Impact of Uncertainty in Big Data Analytics: Case Studies and Challenges

Ahmed Khan
Department of Statistics and Data Science, Quaid-e-Azam University, Sargodha, Pakistan
ahmed.khan@quaidrural.pk

Rubab Naveed
Department of Computer Science and Big Data Analytics
rubabkhan193@gmail.com

Abstract

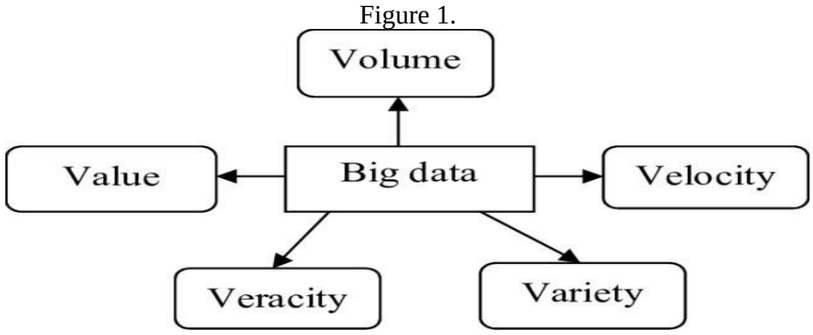
The booming field of big data analytics has attracted significant attention in both the academic and business worlds, driven by the need to extract useful insights from ever-growing and complicated data sets. The volume of data collected in numerous areas such as healthcare, social media, urban infrastructure, agriculture, finance, and education has increased due to developments in sensor networks, cyber-physical systems, and Internet of Things (IoT) technologies. However, these data are often questionable due to noise, incompleteness, and inconsistency, making their interpretation difficult. To systematically examine large amounts of data and enable highly accurate data-driven decision making, sophisticated analytical approaches are required. As data characteristics such as volume, variety and velocity increase, the level of inherent uncertainty also increases, reducing confidence in the subsequent analysis and decisions. Artificial intelligence (AI) approaches, which include machine learning, natural language processing, and computational intelligence, have demonstrated improved performance in terms of accuracy, speed, and scalability in big data analysis compared to traditional methods. Existing research largely focuses on specific approaches or areas; Nevertheless, there is a significant lack of studies addressing the difficulty of uncertainty in big data and the role of AI in minimizing this problem. The purpose of this article is to provide a complete review of the literature on big data analysis and outline the outstanding difficulties and future prospects for managing and mitigating uncertainty.

Indexing terms: Big Data, Uncertainty, Artificial intelligence, computational intelligence, Natural Language Processing

Introduction

The evaluation of big data analytics is becoming increasingly complex due to the inherent uncertainty in both data and models. Uncertainty can stem from various sources such as data inconsistency, imprecision in measurement, and ambiguity in data interpretation. These uncertainties have the potential to significantly affect the accuracy, reliability, and decision-making capability of big data analytical systems [1]. The integration of uncertainty into big data analytics is not straightforward and poses several challenges including computational complexity, scalability, and the need for real-time processing. This paper aims to provide a comprehensive review of the impact of uncertainty on big data analytics by presenting various case studies that illustrate the challenges and solutions. The objective is to shed light on the current methodologies that are capable of handling uncertainty in big data environments, and to highlight the areas that require further research for robust and reliable analytics [2].

Figure 1.



The amount of data generated in the digital age is nothing short of astonishing. According to the National Security Agency, the Internet processes a staggering 1826 petabytes of data per day [3]. To put this in perspective, in 2018, the world was

producing 2.5 quintillion bytes of data every single day. The exponential growth of data generation has defied earlier predictions, with 90% of all the data in the world being generated over just the last two years. Google alone processes an astonishing 3.5 billion searches per day, while Facebook users upload 300 million photos, 510,000 comments, and 293,000 status updates daily. Clearly, the world is awash with data, and this presents both an opportunity and a challenge [4].

To make sense of this data deluge, advanced data analysis techniques are crucial. These techniques are essential for transforming big data into smart data, which provides actionable insights and enhances decision-making capabilities for organizations and businesses [5]. For example, in healthcare, analyzing large datasets, such as Electronic Health Records, can enable practitioners to deliver more effective and affordable solutions by identifying trends in patient histories. Traditional data analytics struggle with big data due to its unique characteristics, known as the five V's: high volume, low veracity, high velocity, high variety, and high value. Additionally, big data presents other challenges like variability, viscosity, validity, and viability. This is where artificial intelligence (AI) techniques like machine learning, natural language processing, computational intelligence, and data mining come into play, offering faster, more accurate, and precise solutions for massive data volumes [6].

The primary aim of these advanced analytic techniques is to uncover hidden patterns, unknown correlations, and valuable information within vast datasets [7]. For instance, a detailed analysis of historical patient data can lead to early disease detection and better treatment plans. Furthermore, businesses can make informed decisions by leveraging simulations and predictive models. However, while AI-powered big data analytics holds great promise, it is not without its challenges. Uncertainty creeps in at various stages, often due to the inherent characteristics of big data. The V characteristics introduce uncertainty sources such as unstructured, incomplete, or noisy data. Dealing with incomplete and imprecise information poses significant challenges, and biased training data can lead to suboptimal results in machine learning algorithms [8]. As we scale these issues to the big data level, any errors or shortcomings in the analytics process are magnified. Therefore, addressing uncertainty in big data analytics is critical, as it can profoundly affect the accuracy of results.

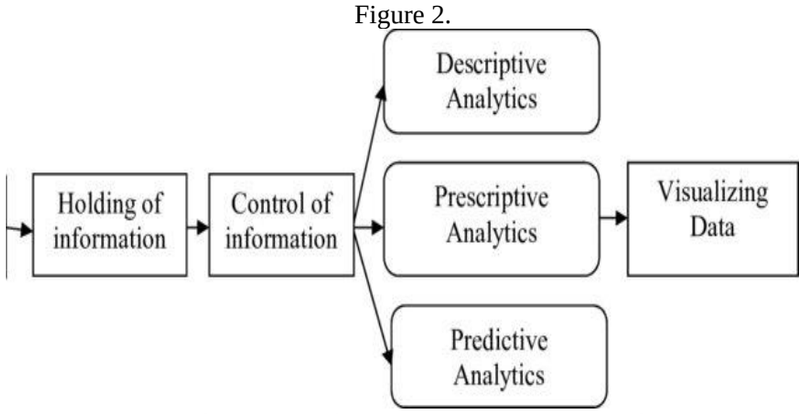
Surprisingly, despite the growing importance of uncertainty in big data analytics, there has been relatively little research dedicated to understanding its impact comprehensively. To fill this gap, this article provides an overview of existing AI techniques for big data analytics, including machine learning, natural language processing, and computational intelligence, from the perspective of uncertainty challenges [9]. It also outlines potential directions for future research in these areas. The contributions of this work include an examination of uncertainty challenges within each of the five V's, a review of techniques considering the impact of uncertainty, and a discussion of strategies to address these challenges. To the best of our knowledge, this is the first comprehensive survey of uncertainty in the context of big data analytics. The remainder of this paper is organized into sections covering background information, the perspective of uncertainty in different AI techniques, a summary of mitigation strategies, and a discussion of future research directions in this exciting and evolving field.

Big Data

The evolution of big data over the years has been nothing short of transformative. Back in 2011, it was heralded as the next frontier, promising unprecedented gains in productivity, innovation, and competitiveness. Fast forward to 2018, and the number of Internet users had surged to a staggering 3.7 billion, emphasizing the vastness of the digital realm. Within this digital universe, data was growing at an astonishing pace, ballooning from 1 zettabyte in 2010 to a staggering 7 zettabytes by 2014. This surge in data gave rise to the concept of the three V's - Volume, Velocity, and Variety, in 2001, defining the fundamental challenges of managing big data [10]. However, the complexity of big data didn't stop there; it expanded to include Value in 2011 and

Veracity in 2012, underlining the importance of quality and context in this data-rich landscape [11].

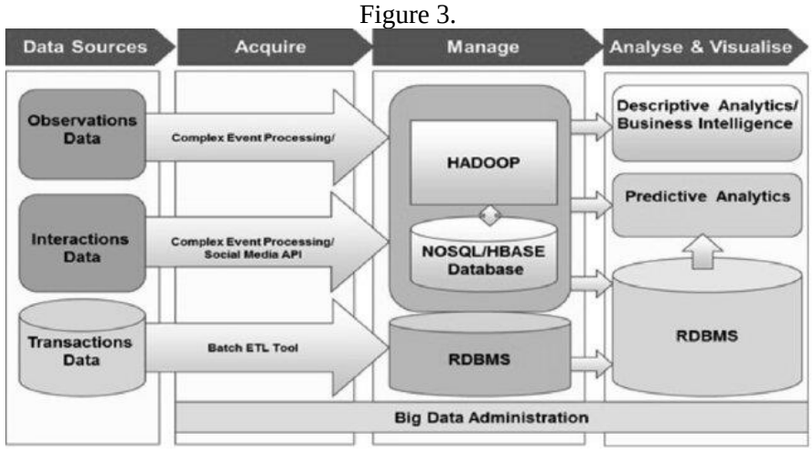
Figure 2.



Volume, in the context of big data, refers to the sheer magnitude of data being generated constantly. Defining a universal threshold for what constitutes a 'big dataset' is impractical, as it depends on various factors such as the time and type of data. While datasets in the exabyte (EB) or zettabyte (ZB) range are generally considered big data, challenges persist even with smaller datasets. For instance, Walmart collects a staggering 2.5 petabytes (PB) of data from over a million customers every hour, presenting scalability and uncertainty issues [12]. Many existing data analysis techniques struggle to handle such massive datasets, falling short when attempting to process and understand data at such a scale. Variety encompasses the diverse forms of data within a dataset, including structured, semi-structured, and unstructured data. Structured data, often stored in relational databases, is well-organized and easily sorted. In contrast, unstructured data, such as text and multimedia content, is random and challenging to analyze. Semi-structured data, like NoSQL databases, contains tags to separate data elements, but enforcing this structure is typically left to the database user. Uncertainty arises when converting between different data types or dealing with mixed data types, impacting the effectiveness of traditional big data analytics algorithms [13]. These algorithms, designed for well-formatted input data, may struggle with incomplete and diverse data formats, presenting a significant challenge in handling multi-modal, incomplete, and noisy data.

Efficiently analyzing unstructured and semi-structured data is particularly challenging due to the heterogeneous sources and data types involved. Real-world databases often suffer from inconsistencies, incompleteness, and noise. Data preprocessing techniques, including data cleaning, data integration, and data transformation, are employed to remove noise from data [14]. Data cleaning techniques address issues related to data quality and uncertainty stemming from data variety, such as noise and inconsistent data. By identifying and eliminating mislabeled training samples, data cleaning can significantly enhance the performance of data analysis, leading to improvements in classification accuracy, especially in machine learning.

Figure 3.



Velocity emphasizes the speed at which data is processed, categorized into batch, near-real-time, real-time, and streaming processing. It underscores the importance of processing data at a pace that matches its generation rate. For instance, Internet of Things (IoT) devices continuously generate vast amounts of sensor data. In the context of medical devices like pacemakers, any delays in processing and transmitting critical data to clinicians can have life-threatening consequences [15]. Similarly, in cyber-physical systems, where real-time operating systems enforce strict timing standards, any delays in delivering data from a big data application can lead to operational issues.

Veracity focuses on the quality of data, including its uncertainty and imprecision. Poor data quality can have significant economic consequences, with estimates suggesting that it costs the US economy billions of dollars annually. In a world where data can be inconsistent, noisy, ambiguous, or incomplete, establishing accuracy and trust in big data analytics becomes challenging. For instance, when analyzing health care records to detect disease trends, any ambiguities or inconsistencies in the dataset can interfere with the precision of the analytics process [16]. Similarly, the use of social media for both official and personal communication can introduce uncertainty when analyzing such data.

Value represents the context and usefulness of data for decision-making. While the previous Vs focus on the challenges of big data, value highlights the potential benefits. Companies like Facebook, Google, and Amazon have leveraged big data analytics to enhance their products and services. Amazon uses large datasets to provide product recommendations, Google utilizes location data for improved mapping services, and Facebook employs user activity data for targeted advertising and friend recommendations. These companies have harnessed the value of big data to make informed business decisions and achieve significant growth.

Uncertainty

Uncertainty is a pervasive challenge in the realm of big data analytics, stemming from a multitude of sources that span every stage of the data processing pipeline. From the inception of data collection, factors like variance in environmental conditions and issues related to sampling introduce ambiguity and imperfection. The very nature of the data, often characterized by multimodality and complexity, further complicates matters. For instance, in the realm of healthcare, patient health records amalgamate numerical, textual, and image data, making it inherently uncertain. Shockingly, a significant portion of attribute values pertinent to the timing of big data events are frequently missing due to noise and incompleteness. Social networks also suffer from missing links between data points, hovering around 80% to 90%, while patient reports transcribed from doctor diagnoses exhibit more than 90% of missing attribute values [17]. Industry experts, as indicated by IBM research in 2014, predicted that by 2015, an astonishing 80% of the world's data would be shrouded in uncertainty [18].

The multifaceted forms of uncertainty in the big data landscape have the potential to severely compromise the accuracy and effectiveness of analytical results. When training data is tainted by biases, incompleteness, or inaccurate sampling, learning algorithms operating on such compromised datasets inevitably produce flawed results. To address this pressing concern, meta-analysis studies that integrate uncertainty and data-driven learning have witnessed a surge in popularity. The management of uncertainty, woven into the fabric of the data analytics process, plays a pivotal role in the performance of big data learning. Notably, the challenges of multimodality and changed-uncertainty set big data apart from conventional small-size datasets, with the size of the dataset itself being positively correlated with the magnitude of uncertainty [19]. Techniques such as employing fuzzy sets to model uncertainty have gained traction, especially when dealing with vague or inaccurate information that may contain hidden relationships.

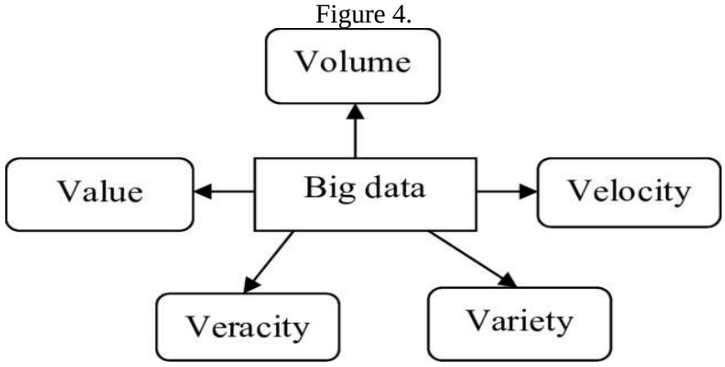
Evaluating uncertainty in big data poses a formidable challenge, especially when data collection methods introduce biases. To tackle the myriad forms of uncertainty, various theories and techniques have been developed. One such approach is Bayesian theory,

which interprets probability based on past events and prior knowledge. Belief function theory, on the other hand, aggregates imperfect data through an information fusion process when faced with uncertainty. Probability theory deals with the statistical characteristics of input data, incorporating randomness. Classification entropy measures ambiguity between classes and provides an index of confidence in classification. Fuzziness is employed to measure uncertainty in classes, particularly in human language. Fuzzy logic then handles uncertainty associated with human perception. Shannon's entropy quantifies information and its missing components in a variable. Rough set theory, with its upper and lower approximations, aids in reasoning with vague, uncertain, or incomplete information [20]. Probabilistic theory and Shannon's entropy are also frequently used to model imprecise, incomplete, and inaccurate data, providing a comprehensive toolkit to navigate the complex landscape of uncertainty in big data analytics.

Data Analytics

Big data analytics is a powerful approach for uncovering valuable insights from massive datasets, offering the capability to identify patterns, correlations, market trends, and user preferences that were previously beyond the reach of traditional analytical tools [21]. The advent of big data's five V characteristics necessitated a reevaluation of analysis techniques due to their limitations in processing time and space. In recent years, the opportunities presented by big data have grown exponentially, with a significant increase in the global adoption of big data technologies and services, accompanied by substantial growth in income generated from big data and business analytics. To effectively harness the potential of big data analytics, various advanced data analysis techniques and strategies have emerged. These include machine learning (ML), data mining, natural language processing (NLP), and computational intelligence (CI), among others. Additionally, strategies like parallelization, divide-and-conquer, incremental learning, sampling, granular computing, feature selection, and instance selection have proven instrumental in addressing the challenges posed by big data.

Figure 4.



Parallelization stands out as a technique that significantly reduces computation time by breaking down complex problems into smaller tasks that can be executed simultaneously. This approach optimizes processing power by distributing tasks across multiple threads, cores, or processors, thereby speeding up data analysis without reducing the overall workload. The divide-and-conquer strategy plays a vital role in handling big data. It involves breaking down a large problem into smaller, more manageable subproblems, solving them individually, and then integrating their solutions to address the overarching issue. This approach has been particularly effective in managing massive databases, where records are processed in smaller groups rather than all at once.

Incremental learning, on the other hand, is tailored for streaming data, adapting learning algorithms to incorporate new data as it arrives. It continually adjusts parameters based on incoming data, ensuring that each data point contributes to the model's ongoing refinement. Sampling offers a valuable data reduction method for big data analytics, enabling the extraction of patterns from large datasets by analyzing a carefully chosen

subset of the data. The effectiveness of this technique depends on the criteria used for data sampling. Granular computing simplifies the complexity of large datasets by grouping elements into subsets or granules. This approach is particularly useful for handling uncertainty in the search space, reducing large objects to a more manageable scale. Feature selection is a critical strategy in data mining, aiming to select a subset of relevant features for a more precise representation of the data. It plays a pivotal role in preparing high-scale data for analysis.

Instance selection is another practical technique commonly used in machine learning and data mining. It allows for the reduction of training sets and runtime in classification and training phases, thereby enhancing efficiency. The field of big data analytics continues to evolve, the application of advanced data analysis techniques and strategic approaches is essential for overcoming the challenges posed by massive datasets. These techniques not only enhance decision-making capabilities but also reduce costs and improve processing efficiency. However, the costs and challenges associated with uncertainty in big data analytics remain significant, prompting further exploration of this issue to ensure the development of robust and high-performing systems.

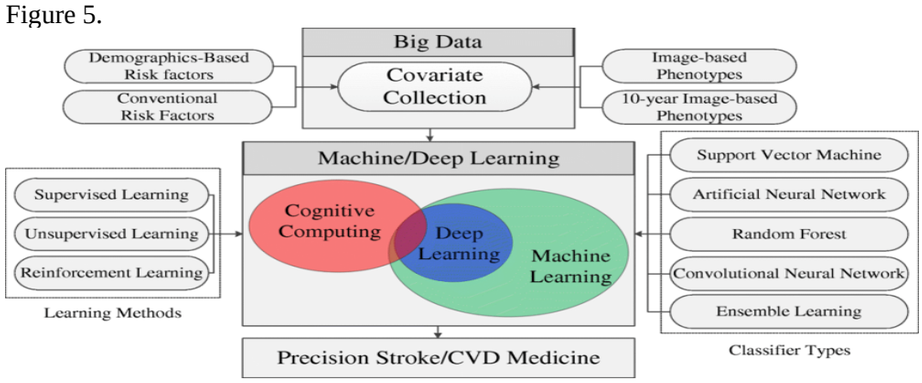
Addressing Uncertainty in Big Data Analytics Techniques

In this section, we delve into the effects of uncertainty on three pivotal AI methodologies applied in the realm of big data analytics: Machine Learning (ML), Natural Language Processing (NLP), and Computational Intelligence (CI). While numerous analytical techniques exist, we narrow our focus to these three and explore the inherent uncertainties associated with each, as well as strategies for mitigating them.

Machine Learning and Its Role in Big Data

In the domain of data analytics, Machine Learning (ML) stands as a vital tool for crafting predictive models and facilitating data-driven decision-making. Traditional ML methods, however, encounter challenges when dealing with the characteristics of big data—such as vast volumes, high velocity, diverse data types, low value density, and data incompleteness—coupled with uncertainties stemming from issues like biased training data and unexpected data types [22].

Figure 5.



Advanced ML techniques tailored for big data analysis include feature learning, deep learning, transfer learning, distributed learning, and active learning. Feature learning allows systems to autonomously uncover essential data representations for detection or classification from raw data. The choice of data representation significantly impacts ML algorithm performance. Deep learning, designed to extract meaningful knowledge from extensive and multi-source datasets, incurs substantial computational costs. Distributed learning combats scalability issues by distributing calculations across several workstations. Transfer learning enhances learners in one domain by transferring knowledge from related contexts [23]. Active learning, meanwhile, expedites ML tasks and mitigates labeling challenges by selectively collecting the most valuable data instances. Uncertainties in ML primarily arise from learning from data of low veracity

and low value. Active learning, deep learning, and fuzzy logic theory are particularly adept at addressing these uncertainties, offering greater flexibility and efficiency [24].

Natural Language Processing in the Big Data Landscape

Natural Language Processing (NLP), a facet of ML, empowers devices to interpret, analyze, and generate text. NLP and big data analytics specialize in processing vast amounts of textual data in real-time. Various NLP techniques, including lexical acquisition, word sense disambiguation, and part-of-speech tagging, have been harnessed for tasks like information extraction, topic modeling, text summarization, classification, clustering, question answering, and opinion mining [25]. For instance, NLP aids in fraud investigations by sifting through immense textual data, identifying criminal names, and analyzing bank records. Additionally, it plays a role in establishing traceability links among textual artifacts [26]. Nevertheless, uncertainty influences NLP, particularly in keyword searches, where documents containing keywords may not guarantee relevance. Ambiguities in word meanings and sentence structures pose challenges for automatic POS tagging. Although IBM Content Analytics (ICA) can alleviate some of these issues, large-scale data remains a concern. Furthermore, biomedical language introduces additional complexities to POS tagging. Integrating NLP techniques with uncertainty modeling like fuzzy and probabilistic sets holds promise for real-time textual data analysis, but further research is required in this domain [27].

Computational Intelligence for Big Data Analysis

Computational Intelligence (CI) encompasses nature-inspired computational methods that play a crucial role in addressing the complexities of big data analysis. These techniques, including evolutionary algorithms, artificial neural networks, and fuzzy logic, are instrumental in tackling high complexity, uncertainty, and data-related challenges that conventional methods struggle with. CI techniques are well-suited for scenarios involving substantial uncertainty, such as predicting user emotions based on extensive, fuzzy emotional data [28]. Challenges persist, especially concerning the veracity and value aspects of big data. New CI techniques are being developed to efficiently handle large datasets and adapt swiftly to data modifications. Swarm intelligence, AI, and ML algorithms are leveraged for tasks like predictive analysis, collaborative filtering, and building empirical statistical models to enhance big data analysis. Fuzzy logic, in particular, excels at modeling qualitative data for uncertainty challenges, making use of linguistic quantifiers and interpretable fuzzy rules for inference and decision-making [29]. EAs, mimicking the evolution process, offer efficient solutions to complex problems presented by big data. However, CI-based algorithms may still be affected by motion, noise, and unforeseen environmental factors, necessitating further research to address these multi-faceted challenges effectively [30].

Conclusion

This paper extensively examined various methodologies in the realm of big data analytics and assessed the impact of uncertainty on each of these approaches. The findings have been conveniently summarized in Table 2. To begin with, each AI technique has been categorized as ML, NLP, or CI. The next column sheds light on how uncertainty affects these techniques, encompassing both data-related and intrinsic uncertainties. Lastly, the third column encapsulates suggested measures to tackle the specific challenges posed by uncertainty in each case. For instance, the first entry in Table 2 highlights a scenario where uncertainty can be introduced in the domain of ML due to incomplete training data [31]. One potential remedy to address this particular form of uncertainty is the application of active learning techniques, which involve selecting a subset of data deemed most critical, thereby mitigating the challenge of limited available training data. It's noteworthy that each aspect of big data was dissected individually, but it's crucial to acknowledge that combining multiple characteristics of

big data introduces a significantly higher level of uncertainty, necessitating further in-depth investigation [32].

This paper delves into the influence of uncertainty on the realm of big data, encompassing its effects on both data analytics and the datasets themselves. The primary objective was to offer an extensive overview of contemporary big data analytics techniques, scrutinize the adverse repercussions of uncertainty on these methods, and elucidate the unresolved challenges that persist [33]. The paper diligently summarizes pertinent research pertaining to each prevalent technique, providing valuable insights for fellow members of the community engaged in the development of their own methodologies. While the discussion comprehensively addresses the quintessential five V's of big data, it acknowledges the existence of numerous other V's. Notably, the research predominantly concentrates on the dimensions of volume, variety, velocity, and veracity of data, with limited attention given to the aspect of value, specifically data associated with corporate interests and decision-making within distinct domains [34].

The passage discusses various avenues for future research in the field. Firstly, it emphasizes the need to investigate the interplay between different characteristics of big data since they coexist and interact in real-world scenarios. Secondly, there is a call for empirical assessments of the scalability and effectiveness of existing analytics techniques when applied to big data. Thirdly, it suggests the development of new techniques and algorithms in machine learning (ML) and natural language processing (NLP) to address the real-time decision-making demands posed by vast datasets [35]. Furthermore, the passage highlights the necessity of exploring efficient methods to model uncertainty in ML and NLP and to represent uncertainty arising from big data analytics. Lastly, it points out that while computational intelligence (CI) algorithms have been employed to address ML and uncertainty challenges in data analytics, there is a notable absence of CI metaheuristic algorithms specifically tailored for tackling uncertainty in the context of big data analytics.

References

[1] C. K. S. Leung, “Big data analysis and mining,” *architecture, mobile computing, and data analytics*, 2019.
[2] R. Ak and R. Bhinge, “Data analytics and uncertainty quantification for energy prediction in manufacturing,” *Conference on Big Data (Big Data)*, 2015.
[3] J. K.u and J. M.David, “Issues, Challenges and Solutions : Big Data Mining,” in *Computer Science & Information Technology (CS & IT)*, 2014.
[4] S. Fosso Wamba, A. Gunasekaran, and R. Dubey, “Big data analytics in operations and supply chain management,” *Ann. Oper. Res.*, 2018.
[5] J. Pei, “Some New Progress in Analyzing and Mining Uncertain and Probabilistic Data for Big Data Analytics,” in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, 2013, pp. 38–45.
[6] M. Muniswamaiah, T. Agerwala, and C. Tappert, “Data virtualization for analytics and business intelligence in big data,” in *CS & IT Conference Proceedings*, 2019, vol. 9.
[7] Y. Yang, S. Liu, and N. Xie, “Uncertainty and grey data analytics,” *Marine Economics and Management*, vol. 2, no. 2, pp. 73–86, Jan. 2019.
[8] W.-H. Weng, “Impacts of Competitive Uncertainty on Supply Chain Competence and Big Data Analytics Utilization: An Information Processing View,” 2020.
[9] B. Chin-Yee and R. Upshur, “Clinical judgement in the era of big data and predictive analytics,” *J. Eval. Clin. Pract.*, vol. 24, no. 3, pp. 638–645, Jun. 2018.
[10] A. Lieto, C. Battaglino, D. P. Radicioni, and M. Sanguinetti, “A Framework for Uncertainty-Aware Visual Analytics in Big Data,” *CEUR Workshop Proc.*, vol. 1510, pp. 146–155, Nov. 2015.
[11] M. Mohamed Nazief Haggag Kotb Kholaif, M. Xiao, and X. Tang, “Covid-19′s fear-uncertainty effect on renewable energy supply chain management and ecological sustainability performance; the moderate effect of big-data analytics,” *Sustain. Energy Technol. Assessments*, vol. 53, no. 102622, p. 102622, Oct. 2022.
[12] G. Cai and S. Mahadevan, “Big data analytics in uncertainty quantification: Application to structural diagnosis and prognosis,” *ASCE-ASME Journal of Risk and Uncertainty*, 2018.

[13] P. Braun, A. Cuzzocrea, F. Jiang, C. K.-S. Leung, and A. G. M. Pazdor, "MapReduce-Based Complex Big Data Analytics over Uncertain and Imprecise Social Networks," in *Big Data Analytics and Knowledge Discovery*, 2017, pp. 130–145.
[14] C. Shang and F. You, "Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era," *Proc. Est. Acad. Sci. Eng.*, vol. 5, no. 6, pp. 1010–1016, Dec. 2019.
[15] C. K.-S. Leung and Y. Hayduk, "Mining Frequent Patterns from Uncertain Data with MapReduce for Big Data Analytics," in *Database Systems for Advanced Applications*, 2013, pp. 440–455.
[16] W. Shi, A. Zhang, X. Zhou, and M. Zhang, "Challenges and Prospects of Uncertainties in Spatial Big Data Analytics," *Ann. Assoc. Am. Geogr.*, vol. 108, no. 6, pp. 1513–1520, Nov. 2018.
[17] M. Fahmideh and G. Beydoun, "Big data analytics architecture design—An application in manufacturing systems," *Comput. Ind. Eng.*, vol. 128, pp. 948–963, Feb. 2019.
[18] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5765–5767.
[19] A. Paul, A. Ahmad, M. M. Rathore, and S. Jabbar, "Smartbuddy: defining human behaviors using big data analytics in social internet of things," *IEEE Wirel. Commun.*, vol. 23, no. 5, pp. 68–74, Oct. 2016.
[20] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 1, p. 41, Jul. 2020.
[21] M. V. Ciasullo, R. Montera, and A. Douglas, "Building SMEs' resilience in times of uncertainty: the role of big data analytics capability and co-innovation," *Transform. Gov. People Proc.Policy*, vol. 16, no. 2, pp. 203–217, Apr. 2022.
[22] M. Kamal and T. A. Bablu, "Machine Learning Models for Predicting Click-through Rates on social media: Factors and Performance Analysis," *IJAMCA*, vol. 12, no. 4, pp. 1–14, Apr. 2022.
[23] M. Vartak *et al.*, "ModelDB: a system for machine learning model management," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, San Francisco, California, 2016, pp. 1–3.
[24] O. Kayode-Ajala, "Applying Machine Learning Algorithms for Detecting Phishing Websites: Applications of SVM, KNN, Decision Trees, and Random Forests," *International Journal of Information and Cybersecurity*, vol. 6, no. 1, pp. 43–61, 2022.
[25] X. Fang and T. Wang, "Using Natural Language Processing to Identify Effective Influencers," *International Journal of Market Research*, vol. 64, no. 5, pp. 611–629, Sep. 2022.
[26] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Federated query processing for big data in data science," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 6145–6147.
[27] D. B. Rawat, R. Doku, and M. Garuba, "Cybersecurity in big data era: From securing big data to data-driven security," *IEEE Trans. Serv. Comput.*, vol. 14, no. 6, pp. 2055–2072, Nov. 2021.
[28] M. Chessa *et al.*, "Three-dimensional printing, holograms, computational modelling, and artificial intelligence for adult congenital heart disease care: an exciting future," *Eur. Heart J.*, vol. 43, no. 28, pp. 2672–2684, Jul. 2022.
[29] K.-K. Mak and M. R. Pichika, "Artificial intelligence in drug development: present status and future prospects," *Drug Discov. Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019.
[30] P. Sundsøy, J. Bjelland, A. M. Iqbal, A. "sandy" Pentland, and Y.-A. de Montjoye, "Big data-driven marketing: How machine learning outperforms marketers' gut-feeling," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, Cham: Springer International Publishing, 2014, pp. 367–374.
[31] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," *Journal of Big Data*, vol. 3, no. 1, p. 25, Nov. 2016.
[32] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," *Proceedings of the 2017 ACM*, 2017.

[33] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Context-aware query performance optimization for big data analytics in healthcare," in *2019 IEEE High Performance Extreme Computing Conference (HPEC-2019)*, 2019, pp. 1–7.
[34] *Anomaly Detection in Network Intrusion Detection Systems Using Machine Learning and Dimensionality Reduction*. .
[35] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of Big Data Privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.