

Assessing Big Data's Impact of Uncertainty from the Perspective of Core Education

Amitava Sarder¹, Somsubhra Gupta² and Ranjan Kumar Mondal³

¹School of CS, Swami Vivekananda University, Barrackpore, West Bengal, India

Email: sarder.amitava@gmail.com

²Department of Science, Swami Vivekananda University, Barrackpore, West Bengal, India

³CSE, Swami Vivekananda University, Barrackpore, West Bengal, India

Emails: dean.science@svu.ac.in, ranjankm@svu.ac.in

Abstract— With organizations like Shodhganga managing sizable academic data repositories, this study examines the growing problem of uncertainty in big data analytics from a regulatory and educational standpoint. Numerous industries, including healthcare, urban infrastructure, social media, finance, and education, are generating huge quantities of data as a result of the quick development of sensor networks, cyber-physical systems, and the Internet of Things (IoT). However, the accuracy and comprehension of this data are often compromised by noise, incompleteness, and inconsistencies. Accurate decision-making becomes even more challenging as data attributes—volume, variety, and velocity—increase due to the associated uncertainty. Artificial intelligence (AI) approaches, especially machine learning, natural language processing, and computational intelligence, have demonstrated better performance in terms of scalability and precision than traditional deterministic methods, which falter under such complexity. The study suggests a hybrid, multi-layered analytical framework to address these issues. This method combines neural networks for semantic-rich prediction, wrapper-based feature selection to reduce dimensionality, probabilistic modeling for informed classification, and fuzzy logic for handling subtle uncertainties. The real-world student-por.csv dataset is used to test this framework, and the results show that it can accurately classify educational outcomes with high F1-scores (85%), particularly for high-performing students. The study discusses advancement in theory and practice by providing an innovative approach which is sensitive to uncertainty and turns noisy educational datasets into intelligent perception for directing institutional policy and individualized learning plans

Index Terms— Big Data Analytics, Artificial Intelligence, Uncertainty, Wrapper Feature Selection, Educational Data, Neural Networks, Fuzzy Logic.

I. INTRODUCTION

Big data analytics (BDA) faces increasing difficulties for evaluation due to the uncertain nature of the data and models. Ambiguous data interpretation, inconsistent data and imprecise measurements are just a few of the elements which may be responsible for such uncertainty. Important data analytical systems face remarkable hindrance when they try to make decisions with reliability and precision because of the presence of these uncertainties in data [1].

The existence of some problems like real-time processing requirements, scalability and computational complexity are due to integration of uncertainty into BDA which makes it very difficult for further analysis. This article gives us an extensive overview through a number of case studies which highlight the challenges and achievable remedies to cogitate the impact of uncertainty on big data analytics. The current study aims to explain existing techniques that can manage uncertainty in big data environments and to select prospective areas with additional investigations for ensuring trustworthy and accurate analytics [2]. The digital age surprisingly generates huge amount of data. Internet processes an unbelievable 1826 petabytes of data everyday which is asserted by the National Security Agency [3]. To put this into perspective, 2.5 quintillion bytes of daily data generation by the globe take place. The last two years on its own produced 90% of the global data and this contradicts the previous prediction of an accelerating increase in growth of data output. Facebook members upload photographs of 300 million, 510,000 numbered comments, and total count of 293,000 for update of status daily and Google solely process an overwhelming 3.5 billion queries every day. There is no doubt about the abundance of data in the world which offers opportunities as well as challenges [4].

Innovative approaches for analyzing data are crucial to making sense of this data deluge. By transforming big data into knowledgeable data, these necessary techniques give businesses and organizations more capacity to make decisions and offering actionable insights [5]. For instance, by seeing patterns in patient histories, healthcare professionals may provide more efficient and cost-effective care by analyzing massive databases, such as Electronic Health Records. Big data has unique properties called the five V's: large volume, poor veracity, high velocity, wide variety, and high value. These qualities make traditional data analytics difficult to use. Big data also presents additional challenges related to validity, viability, variability, and viscosity. Techniques for Artificial intelligence (AI) that includes machine learning, natural language processing, computational intelligence, and data mining, deliver faster, more precise, and more accurate answers for large data sets. These cutting-edge analytical methods aim to find important information, unidentified relationships, and hidden patterns in enormous datasets [7]. For example, past patient data can be thoroughly analyzed to enhance treatment strategies and aid to detect illness early. Additionally, organizations can make better-informed decisions by using predictive models and simulations. Though BDA powered by AI has a lot of capabilities, it also faces challenges. Uncertainty can sometimes arise, often because of the characteristics that make up big data itself. Unstructured, partial, or noisy data are examples of the sources of uncertainty caused by the V characteristics. Biased training data can cause machine learning algorithms to produce less-than-ideal results, and dealing with imprecise and missing information presents substantial hurdles. [8]. When we expand these problems to the significant data level, any mistakes or flaws in the analytics process are amplified. Because uncertainty significantly impacts result accuracy, BDA must handle it (Refer to Fig. 1 for BDA Process).

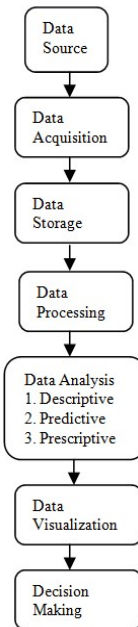


Figure 1: Process of Big Data Analytics

Remarkably, a few studies were performed to fully understand the influence of uncertain characteristics in big data analytics, even though it is becoming increasingly important. Computational intelligence (CI), natural language processing (NLP) and machine learning (ML) are some of the state-of-the-art AI methods in BDA through which the article fills this void space by summarizing them from the standpoint of uncertainty problems [9]. It also suggest possible fields of inquiry for related future studies. This study contributes by analyzing uncertainty difficulties in each of the five V's, reviewing procedures that take uncertainty into account, and discussing approaches to deal with these challenges. This is the first thorough examination of uncertainty in respect of big data analytics, as far as we are aware. The remaining portion of this study is categorized into sections that pertain to background data, the uncertainty viewpoint on various AI approaches, mitigation solutions in brief, and future research prospects in this fascinating and developing topic.

II. BIG DATA

Over time, big data has undergone a revolutionary metamorphosis. It was hailed as the next big thing in 2011, with increased unheard-of productivity, creativity, and competitiveness. By 2018, there were an astounding 3.7 billion Internet users, underscoring the size of the digital world. Data in this digital world expanded exponentially, from one zettabyte in 2010 to seven zettabytes by 2014. The three V's (Volume, Velocity, and Variety) were developed in 2001 in response to this data explosion, and they define the core difficulties in handling 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] (Refer to Fig. 2 for Layers of Big Data).

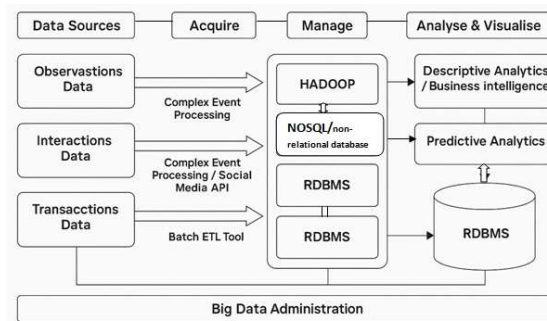


Figure 2: Different Layers of Big Data



Figure3. 5 V's of Big Data

When discussing big data, the term 'volume' refers to the continuous production of the huge quantity of data. It is unfeasible to define a uniform criterion for what qualifies as a "big dataset," as this relies on several variables, including the nature and duration of the data. Although exabyte (EB) and zettabyte (ZB) files are typically regarded as extensive data, difficulties can still arise with smaller datasets. For example, an astounding 2.5 petabytes (PB) of data collected by Walmart over a million customers per hour raises concerns about scalability and unpredictability [12]. The vast and varied nature of big data makes it difficult for current data analysis techniques to scale. "Variety" refers to data formats that pose a challenge to conventional algorithms built for consistent inputs, such as semi-structured (like NoSQL), unstructured (like text, multimedia), and structured (like relational databases). On multi-modal, noisy, or incomplete data, these approaches frequently perform poorly [13]. In order to control noise and inconsistency, data preparation procedures such as integration, transformation, and cleaning are essential. By eliminating incorrectly labeled samples, machine learning classification accuracy can be increased [14]. "Velocity" refers to how quickly data is created and needs to be processed, such as when IoT devices generate continuous sensor data. Delays in pacemaker data transmission in the medical field can be lethal; similarly, real-time processing is necessary for cyber-physical systems to prevent failure [15]. Analytics dependability is impacted by "veracity," which deals with data quality issues like imprecision, ambiguity, and inconsistency. For example, sentiment analysis or illness detection may be distorted by subpar healthcare or social media data. Poor data quality is thought to cost the American economy billions of dollars every year [16]. Finally, "Value" refers to the capacity of data to guide choices. Businesses such as Amazon (recommendation engines), Google (location data for maps), and Facebook (user

behavior for ads) are prime examples of how wise business decisions can be made and services can be improved through the utilization of big data (Refer to Fig. 3 for 5 V's of Big Data).

III. UNCERTAINTIES

Every step of pipeline for data is rife with uncertain behaviour of big data analytics, from collection (caused by sampling problems and environmental variability) to processing of data that is inherently complex and multimodal, such as medical records (text, numbers, and images). Missing values and noise are serious issues: While 90% of attributes may be missing from medical diagnosis transcripts, 80–90% of social network data lacks relational links [17]. By 2015, IBM projected that 80% of the world's data would be obscured by uncertainty [18], and that uncertainty significantly influences the accuracy of analytical models, particularly due to biased or incomplete training data. The degree of uncertainty increases as the dataset becomes larger and more complex [19]. As a result, methods like fuzzy sets are being utilized more and more to deal with imprecise or noisy data. Several methods exist to deal with uncertainty: Bayesian theory assesses likelihood based on historical data; belief function theory combines evidence when it is insufficient; classification entropy measures prediction uncertainty; and probability theory deals with randomness. Fuzzy logic is essential for handling linguistic and perceptual ambiguities, while Shannon's entropy measures missing or ambiguous data. Additionally, rough set theory facilitates the process of easy reasoning with unclear or insufficient data [20]. Despite being complementary in nature, these tools combine well with each other with regard to the modeling and management of uncertain nature of big data.

IV. METHODOLOGICAL ASPECTS

Hidden patterns, correlations, and user trends in huge amounts of data are the revelation of BDA which remains undetected by conventional methods [11] [14] [20]. Traditional analytical tools must be rethought in light of the five V's of big data because of processing time and storage limitations. In addition to raising global income, the expanding use of big data technologies has improved techniques for analysis such as CI, data mining, ML and NLP. Methods such as granular computing, sampling, incremental learning, parallelization, divide-and-conquer, instance selection and feature selection have all shown promise in addressing big data challenges. Parallelization reduces computational load and increases speed by dividing work among multiple processors. The divide-and-conquer strategy is especially helpful when controlling large databases because it divides big problems into smaller ones. Incremental learning is used to continuously adapt models to streaming data, ensuring continuous improvement. Sampling enables the discovery of patterns in large datasets by using representative subsets. Granular computing divides data into diminutive portions known as granules and thus it reduces complexity, ultimately enables assisting in managing uncertainty. Feature selection improves data representation by selecting pertinent attributes, which is essential for scalable analytics. Instance selection minimizes the dataset's size for efficient training and classification. Even though these techniques improve decision-making, processing speed, and cost-effectiveness, uncertainty remains a significant issue in BDA and needs further research to ensure reliable, scalable solutions.

A. Uncertainty addressal techniques of big data analytics

The influence of uncertainty on three important artificial intelligence (AI) techniques used in big data analytics—ML, NLP, and CI—is examined in this section. While there are availability of numerous analytical methods, we focus on these three and examine the inherent uncertainties they contain as well as strategies for reducing them.

Machine Learning and its role in Big Data

In data analytics, Machine Learning (ML) is an essential technology that helps create predictive models and supports data-driven decision-making. However, the characteristics of big data—large volumes, high velocity, diverse data types, low-value density, and incomplete data—pose challenges for traditional machine learning methods. Uncertainties brought on by issues like skewed training data and unexpected data types exacerbate these difficulties [11]

Examples of some contemporary machine learning approaches concentrating on big data analysis are transfer learning, feature learning, deep learning, active learning and distributed learning. Computers are able to perform automatic revelation of meaningful delineation from unprocessed data through feature learning method. Deep learning method, though suffers from computational comprehensiveness, can successfully analyze many significant perceptual structures from extensive datasets with various sources. In this respect, the performances

of machine learning models are determined by significant domination of data format. Computations amongst various systems are distributed through scalable performance. The performance of the model is improved through transfer learning on application of insightful conclusions from related fields. Active learning minimizes label expenses and selects the most informative examples and in this way, it advances the process. [14]. Deep learning, active learning and fuzzy logic methods, with increased adaptability, control uncertainty in machine learning with logical relevance [13] [15].

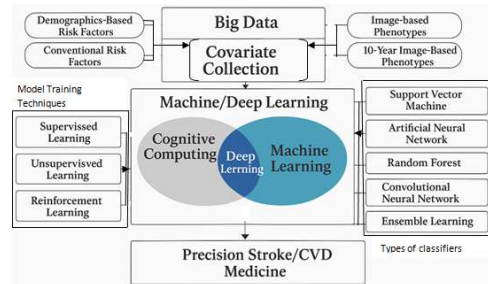


Figure 4. Machine learning and deep learning system framework using big data

Natural Language Processing in Big Data context

One of the application areas of machine learning is Natural language processing (NLP) which expedites the machines to be capable of generation, comprehension and assessment of human language in proper way. One crucial feature of BDA allows it to analyze vast amounts of real-time effectuated data. Classification, question-answering, clustering, opinion mining, topic modeling, information extraction and text summarization are some applications in which NLP methods viz. word sense disambiguation, part-of-speech tagging and lexical acquisition are induced [20]. In this respect, NLP assists in detection of fraud by carefully examining bank records and identifying criminal connections [10]. Nonetheless, biological terminology, unclear phrases structures and uncertain keyword with irrelevancy are some of the difficulties encountered by NLP in automation of POS tagging. Large datasets remain challenging to control regardless of the support of applications such as IBM Content Analytics (ICA).Improvement could result from the combination of probabilistic and fuzzy uncertainty modeling purposes ,however further study is needed [6][18].

Computational Intelligence Technique with Big Data Analytics

Computational intelligence (CI) is based on naturalistic approaches viz. artificial neural networks, evolutionary algorithms and fuzzy logic and is essential for managing large dimension, complex and uncertain behavior which are intrinsically complex to handle for traditional methods.CI techniques are especially helpful to deal with ambiguous cases like comprehension of imprecise data regarding emotions for analyzing sentiment through sentiment analysis(SA) technique[10][19].However, big datasets often face utility and authenticity problems. With the goal of effective and dynamic adjustment to the changes in processing of big data, development of new CI techniques is required. Techniques like Machine learning, AI and swarm intelligence make data analysis even better for things like statistical inference, collaborative filtering and predictive modeling. Fuzzy logic is excellent for simulation of qualitative uncertainty when it uses comprehensible fuzzy rules and quantifiers that are linguistic in nature [5] [10].Evolutionary algorithms (EAs) provide helpful approaches to big data issues by resembling natural selection. To properly handle these complexities, more research is necessary because problems like noise, motion and unforeseen factors in the environment can still have a detrimental effect on CI algorithms [4][16].

V. SUGGESTED STRUCTURAL ARRANGEMENT

Our proposal lies in developing an analytical framework consisting of multiple layers with six key components to address the uncertain nature in big data that affects core leveled education. At first, educational big data is collected and preprocessed which may include cleanliness and reduction of noise in terms of student feedback and performance logs. A subsequent division is made of uncertainty into two types such as epistemic and aleatoric categories where probabilistic inference is used. Thereafter, dimensionality reduction and wrapper-based feature selection techniques are used for elimination of redundant attributes, thus resulting in enhanced model generalizability. Fuzzy logic uses membership functions to explain implicit cases, thus managing borderline instances. Predictive modeling uses techniques of supervised learning like decision trees and neural

networks for processing the optimized dataset. Despite being mentioned lastly, a feedback loop supposes to affect tailored strategies and curriculum design which, eventually has an influence on choosing about educational policy. Altogether, allowance of ambiguity for systematic removal is possible when educational analytics is reliably improved.

VI. DETAILS OF IMPLEMENTATION

The public dataset student-por.csv is used in the implementation phase containing real life educational data like behavior characteristics, academic grades (G1-G3) and demographics of Portuguese secondary school students. A framework with multiple layers is demonstrated in the code. Initially preprocessing process controls label uncertainty ,then student data is encoded before assigning the labels with fuzzy logic and performance bands (Low, Medium, High)relying upon the final grade (G3). Random Forests and wrapper-based feature selection are applied to reduce dimensionality, leaving only the top 12 predictors. Academic performance is modeled with 85% accuracy and high F1-scores after a neural network is trained using the chosen features, particularly for the "High" performance class. Some important goals of the proposed framework are to reduce unknowns, improve predictive reliability, and support data-driven decision-making in educational systems and these are indeed reflected through this practical implementation.

VII. RESULT ANALYSIS

The result summary of the suggested implementation is depicted in following table:

TABLE I: KEY FINDINGS OF THE RESULT

Metric / Class Label	Precision	Recall	F1-Score	Support
Low (0)	0.88	0.81	0.84	26
Medium (1)	0.72	0.65	0.68	20
High (2)	0.86	0.90	0.88	84
Overall Accuracy	-	-	0.85	130
Macro Avg	0.82	0.79	0.80	130
Weighted Avg	0.84	0.85	0.84	130

A. Brief Analysis

With the highest F1-score (0.88) for the High performance class and an overall accuracy of 85%, the model demonstrated dependable predictions for students who perform well. Due to overlap or ambiguity in borderline grades, the Medium group's performance was marginally worse (F1-score: 0.68), confirming the applicability of fuzzy labeling. The wrapper-based selection method is validated by the top features (G2, G1, failures, absences) matching behavioral and academic indicators. Overall, the findings show dependable performance classification and efficient uncertainty handling, as intended by the suggested educational framework (Refer to Table I).

B. Visual Interpretation: Confusion Matrix

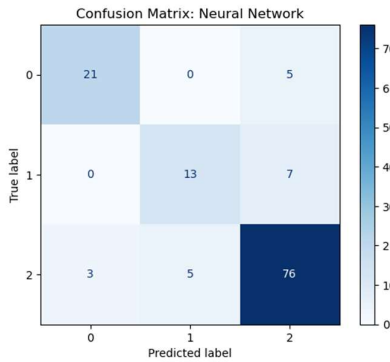


Figure 5: Confusion Matrix for the outcome

The neural network's classification performance in the study is demonstrated graphically by the uploaded confusion matrix. It shows that 76 out of 84 "High" performance cases, 13 out of 20 "Medium" performance cases, and 21 out of 26 "Low" performance cases were accurately identified by the model. The "Medium"

category, in particular, shows the misclassifications, thus emphasizing the difficulties faced in imprecise and overlapping educational data. The suggested hybrid framework necessitates its significance through the effective combination of fuzzy logic for handling label ambiguity, wrapper-based feature selection and deep learning. In addition to that, though the academic data is unclear or noisy, it manifests that it is reliable in ameliorated prophetic manner (Fig. 5).

VIII. DISCUSSION

The reliability of the model is often compromised by the intrinsic uncertain nature of big data arising from imprecise, incomplete or inconsistent data although it may transform educational decision-making. The proposed combined method with hybrid structure, achieved through the integration of machine learning, fuzzy logic and probabilistic modeling, permits to analyze uncertain sensitivity towards the changes of student behaviour. The Python implementation successfully substantiates this framework using the student-por.csv dataset. Four components viz. neural network-based predictive modeling, wrapper-based feature selection using Random Forests, fuzzy performance labeling to manage ambiguity and data preprocessing are incorporated by this framework. Enhanced accuracy and interpretability with facilitation of the creation of feedback-driven educational viewpoints along with practical feasibility for policymakers and institutions are achieved by this proposed multi-layered methodology.

IX. CONCLUSIONS

In summary, with the objective to thoroughly examine the influence of uncertainty in big data analytics, this study assessed a range of modern methodologies which include machine ML, NLP and computational intelligence CI, particularly in the educational sector. Different kinds of uncertainty, such as those resulting from ambiguous patterns, insufficient training data or inconsistent labeling, can have a detrimental impact on analytical outcomes, which is clearly highlighted by the article. Suggested mitigation strategies, like active learning improve reliability of the model by reinforced importance of choosing important data subsets. The proposed hybrid framework, which combines fuzzy logic for uncertainty-aware labeling, wrapper-based feature selection for dimensionality reduction, and neural networks for predictive modeling, has been effectively utilized to analyze student performance data. A confusion matrix and accurate classification metrics supported the resulting model's 85% accuracy, pointing out its efficacy in transforming noisy, complex educational datasets into accurate observations. In spite of being taken into consideration important big data attributes (volume, variety, velocity, and veracity), the study perceives gaps in addressing the "value" dimension pertaining to decision-making. Future directions focus on the development of novel NLP and ML algorithms so that decisions can be made in real time, inspection can be done on how scalable the current analytics tools are so that can be experimentally evaluated on large datasets, and observations can be performed for the refinement of CI metaheuristics designed to manage uncertain characteristics in big data scenario. All of these findings necessitates the development of multi-layered, uncertainty-aware approaches so that trustworthy educational analytics and policy suggestions can be generated.

REFERENCES

- [1] C. K.-S. Leung, 'Big data analysis and mining', in Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global Scientific Publishing, 2018, pp. 338–348.
- [2] R. Ak and R. Bhinge, 'Data analytics and uncertainty quantification for energy prediction in manufacturing', in 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 2782–2784.
- [3] K. U. Jaseena, J. M. David, and Others, 'Issues, challenges, and solutions: big data mining', CS & IT-CSCP, vol. 4, no. 13, pp. 131–140, 2014.
- [4] S. Fosso Wamba, A. Gunasekaran, R. Dubey, and E. W. T. Ngai, 'Big data analytics in operations and supply chain management', Annals of Operations Research, vol. 270, no. 1, pp. 1–4, 2018.
- [5] J. Pei, 'Some new progress in analyzing and mining uncertain and probabilistic data for big data analytics', in International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft 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, 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', *Journal of Evaluation in Clinical Practice*, vol. 24, no. 3, pp. 638–645, 2018.
- [10] A. Karami, 'A framework for uncertainty-aware visual analytics in big data', in *CEUR Workshop Proceedings*, 2015, vol. 1510, pp. 146–155.
- [11] M. M. N. H. K. Kholaiif, 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', *Sustainable Energy Technologies and Assessments*, vol. 53, p. 102622, 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 in Engineering Systems, Part A: Civil Engineering*, vol. 4, no. 1, p. 04018003, 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 *International Conference on 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', *Engineering*, vol. 5, no. 6, pp. 1010–1016, 2019.
- [15] C. K.-S. Leung and Y. Hayduk, 'Mining frequent patterns from uncertain data with MapReduce for big data analytics', in *International Conference on 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', *Annals of the American Association of Geographers*, vol. 108, no. 6, pp. 1513–1520, 2018.
- [17] M. Fahmideh and G. Beydoun, 'Big data analytics architecture design—An application in manufacturing systems', *Computers & Industrial Engineering*, vol. 128, pp. 948–963, 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 Wireless communications*, vol. 23, no. 5, pp. 68–74, 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, 2020.