iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XII Month of publication: December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76719

A Comprehensive Review of Machine Learning Algorithms for Big Data Analytics

Gyanendra Kumar Gautam[1], Vishesh Saxena[2], Vineet Mishra[3], Vinod Kumar[4]
[1, 2,4]Assistant Professor, [3]Associate Professor, Department of Computer Science, Agra Public College of Technology & Management, Agra

Abstract: This study examines the role of machine learning algorithms in big data analysis, highlighting how their integration enables intelligent data processing and decision-making across diverse application domains. The continuous growth in data volume, velocity, and variety has made traditional data analysis techniques inadequate, thereby necessitating advanced analytical tools for efficient big data processing, analysis, and storage. Machine learning algorithms address these complexities by enabling automated pattern discovery, predictive modeling, and intelligent decision-making in large-scale and heterogeneous big data environments. This paper presents a structured review of machine learning algorithms for big data analysis, covering the background and characteristics of big data (5Vs) and examining supervised, unsupervised, and reinforcement learning techniques used in large-scale data analytics. The practical applications of these algorithms are discussed across multiple sectors, including healthcare, banking, retail, manufacturing, and telecommunications. The study further reviews emerging trends influencing the convergence of big data and machine learning, with particular emphasis on ethical artificial intelligence, data privacy, security concerns, and scalability issues. Key challenges such as model interpretability, scalability, data quality, and data diversity in big data analysis are critically discussed. Finally, the paper outlines future research directions, including real-time big data analytics, edge computing, automated machine learning, and ethical artificial intelligence frameworks, and presents a comparative tabular summary of machine learning algorithms applied in big data analysis to support structured understanding and future research.

Keywords: Machine Learning, Data Analytics, Big Data Analysis, Unsupervised Learning, Supervised Learning, Reinforcement Learning, Predictive Analytics, Scalability, Privacy, Edge Computing, Data Privacy and Security.

I. INTRODUCTION

A growing number of informed individuals in the IT and analytics communities are becoming watchful while using the term "big data." The size and complexity of big data make it hard for conventional systems and data-warehousing technologies to handle and process. Big data is produced by both humans and technology, as well as by the natural world. Large amounts of data, which might be structured, semi-structured, or unstructured from many sources are created as a result of the development of services and technologies[1].

1) *Types of Big Data:* Big Data has been divided into following categories-

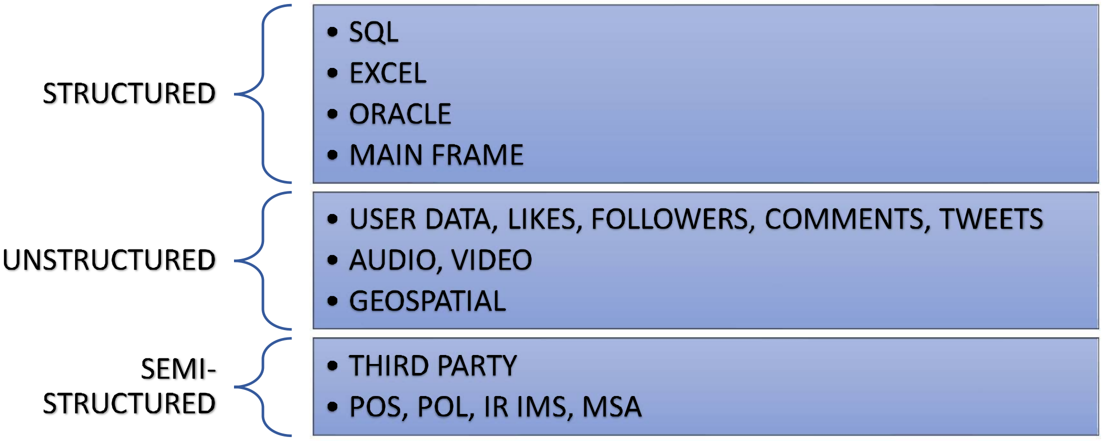

Fig. 1. Types of Big Data

2) *Big Data Characteristics (5Vs):* There are 5Vs of Big Data that are:

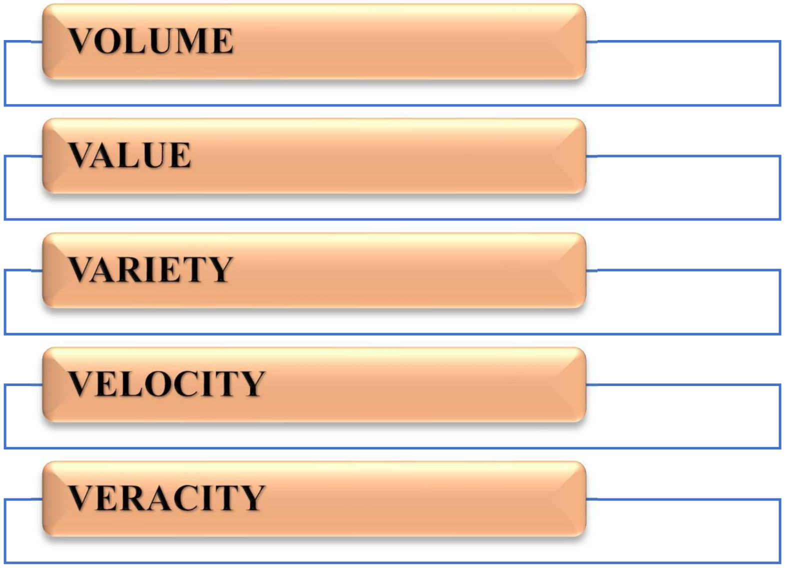

Fig. 2. Characteristics of Big Data

3) *Origin of Big Data:* Big Data can take place form different sources -

- Data Explosion: With the advent of digitalization around the end of the 20th century, more and more data began accumulating at a rapid pace. This caused the generation of rapid and unprecedented growth of data volumes that the rather rudimentary conventional data handling techniques could not cope including data from transactions, sensors, social media, and other sources.
- Internet Growth: In the 1990s, the internet spiral began, which also accelerated the processes of data creation and data dissemination. The world wide web transformed into macromedia of both unstructured (text, multimedia) and structured (databases), which raised new challenges and opportunities for data management and analysis.
- Technological Advancements: Organizations gained capacity to collect, save, and work on more and more large amount of data sets due to processing power and storage capacity advancements. This also created challenges that led to the publishing of Big data with its volume, variety and velocity, distributed computing framework and scalable databases.

4) *Big Data Technologies:* Highlighted big Data problems also called for solutions such as Apache Hadoop, Apache Spark, and all.

5) *Applications of Big Data:* Some application areas of big data are as follows-

Table 1: The key applications of Big Data

Sector	Applications
Healthcare	• Personalized Medicine • Medical Imaging Analysis • Drug Discovery • Healthcare Management
Finance	• Fraud Detection • Algorithmic Trading • Risk Management • Market Analysis.
Retail	• Customer Segmentation • Inventory Optimization • Dynamic Pricing

Sector	Applications
	• Recommendation Engines
Manufacturing	• Predictive Maintenance • Quality Control • Supply Chain Optimization • Energy Efficiency
Telecommunications	• Network Optimization • Service Personalization • Real-time Analytics • IoT Data Management

6) *Challenges of Big Data:* Big data volume, velocity, variety, and authenticity provide a number of issues. The following are some major obstacles related to big data shows with block diagram.

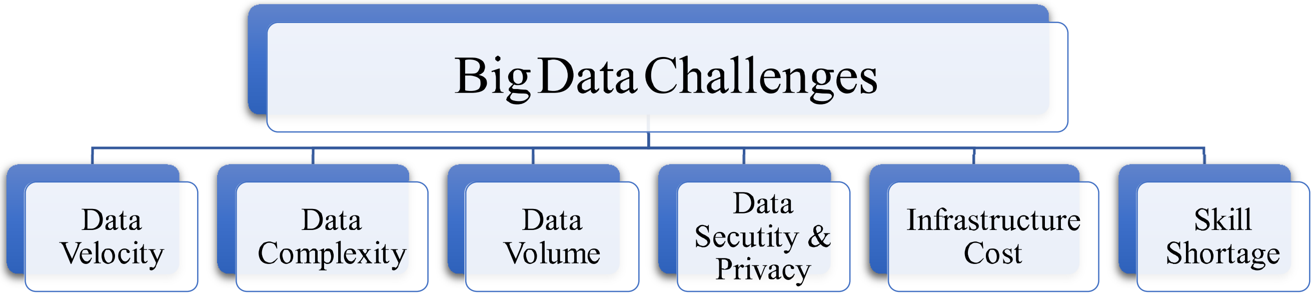

Fig. 3. Big Data Challenges Block Diagram

7) *Role of Big Data in Future:* Big Data will continue to shape the future in many different fields. The following crucial roles are anticipated of it:

- Artificial Intelligence and Machine Learning: Big Data will keep advancing the development of AI and ML. As the volume and variety of datasets increases, AI/ML algorithms will grow consequently and allow for better analysis, forecasting and systemization of the processes in the industries such as healthcare, finance, manufacturing and others.
- Data Privacy and Security: The more data there is the more the need to secure it and its contents. Advances in BD technologies will assist in increasing the encryption levels, addressing possible security threats, and ensuring that legal requirements (like GDPR, CCPA) are adhered to in order to safely extract valuable information.
- Smart Cities and Urban Planning: BD will support smart cities' strategies by providing data integration techniques from other sources (including but not limited to sensors, social media, traffic flow). BD can provide a range of services such as resources management, traffic management, safety services, and urban planning.
- Healthcare and Precision Medicine: BD will enhance healthcare systems as a result of development in genomics and personalized health care, and management of health at the population level. Through evaluating large scale medical databases, forecasted and resultant illnesses will also be treated, profiled and epidemiological tendencies will be determined at levels above customary measures.

II. TECHNIQUES OF MACHINE LEARNING

Machine learning is a core discipline of artificial intelligence that enables computer systems to learn patterns from data and improve performance without being explicitly programmed. This capability is especially important in big data environments, where large volumes of complex and heterogeneous data require automated and adaptive analytical techniques.

1) *Supervised Learning:* It is the approach that provides a clear methodology for training algorithms on how to recognize patterns and forecast outcomes based on labelled datasets. These algorithms include the following:
- Linear Regression: It is a classical supervised approach used to establish the relationship that may exist between Two or more variables using Regression equation Models.
- Support Vector Machines (SVM): SVM is a common technique applied in machine learning which is defined as multi-class classifier for remote sensing image Classification and enhancing the classification accuracy of multitemporal satellite images. For a given sample, it is capable of regression and classification to determine the fitting hyperplane which best cuts the data space. The SVM has been demonstrated to effectively classify both the uniform and the diverse features in morphology in within remote [2].
- Neural Networks: Neural Networking is an effective computing modelling artificial brain, which applies neural interconnections for stimuli processing. Thanks to improved hardware and back propagation technologies, neural networks are becoming widely utilized for big data applications such as image and voice recognition[3]

2) Unsupervised Learning: This learning approach employs a computer to learn on its own without human involvement. The purposes of unsupervised learning are the restructuring of the input data into new attributes or a set of objects with the recognizable patterns. Typical algorithms are:
- K-Means Clustering: It teaches the algorithm to be able to perform autonomously on unlabelled, unclassified data and to permit the computer to do so. That being said, the task of making a machine in this case is to categorize disordered information based on similarities, structures and discrepancies only.
- Hierarchical Clustering: In the field of data mining, cluster analysis is a set of techniques for the construction of nested orthogonal partitions. During the process every datum point is initialized as a single cluster and clusters which are in close proximity get merged stepwise until a satisfaction criterion is achieved.
- Principal Component Analysis: This approach was first developed in 1901 by the mathematician Karl Pearson. It doesn't allow too much circumstance that although the data from space of high axon number has been projected to lower space the lower space data must have its variances to be as high as possible.

3) *Reinforcement Learning:* Reinforcement learning is a machine learning paradigm in which an autonomous agent learns by interacting with its environment. Instead of relying on labeled data, the agent receives feedback in the form of rewards or penalties based on the actions it performs. Through continuous interaction, the agent aims to learn an optimal policy that maximizes the cumulative reward signal over time, making reinforcement learning suitable for sequential decision-making and dynamic big data environments [4**].** Typical algorithms used in reinforcement learning include:
- Q-learning: A model-free algorithm that learns optimal action–value functions through iterative reward-based updates.
- Deep Q-Networks (DQN): An extension of Q-learning that uses deep neural networks to handle high-dimensional and large-scale data.
- Actor–Critic Methods: A hybrid approach that combines value-based and policy-based learning to improve learning stability and performance in complex environments.

III. APPLICATIONS OF MACHINE LEARNING INTO BIG DATA

1) Healthcare: Making accountable predictions on the possible outcome of the patient, predicting illness and having customized therapy plans for the patients.
2) Finance: Segmentation and targeting of customers, trading using algorithms, management of risks, and fraud detection.
3) Retail: Control of the stock, analysis of customers' behaviour, and suggestion of systems.
4) Transportation: It is applied in the prediction of traffic jam, the most efficient routes and also self-driven automobiles.

IV. CHALLENGES IN MACHINE LEARNING BASED BIG DATA ANALYSIS

1) Scalability: Big data initiatives can grow fast. Cloud computing is the result's scalability of Big Data problem. It presents some drawback, such as managing and carrying out different tasks to meet each object of workload objective in an economical manner.

2) Data Quality and Variety: The huge amount of data generated can also have an effect on the correctness of data and quality. Making sure the data is appropriate, current, and full can be challenging given the volume of data being generated. This might lead to assumptions or fallacy in the data analysis, generating inaccurate or insufficient discovery.

3) Interpretability: Explanation of the big data interpretability: To communicate its conclusions properly, one has to understand complex models that are used. It is quite challenging to remain transparent without compromising privacy and proprietary information. Some of the significant challenges include a balance between accuracy and interpretability. Ensuring to retain interpretability while obtaining significant insights still remains an essential objective in making successful use of large data.

4) Privacy and Security: Big data gives privacy and security challenges that including protect sensitive data, complying with laws such as GDPR, and reducing the likelihood of illegal access or data breaches. These tasks call for strong encryption, anonymization, and access control mechanisms.

V. FUTURE DIRECTIONS

The growth of machine learning and artificial intelligence for deeper insights, edge computing for real-time analytics, and ethical concerns about data governance and privacy will probably be the main aim of big data in the coming times.

1) Integration of Domain Knowledge: with a view to enhance model performance and interpretability, domain-specific knowledge will be progressively integrated into future machine learning systems.

2) Automated Machine Learning: The aim of the promptly growing field of automated machine learning research is to make machine learning obtainable to non-experts in the field by automating the processes required to build high-performance machine-learning pipelines for specific use cases[5].

3) Ethical AI and Fairness: with a view to ensure fairness in machine learning models and address biases in big data, further study and development are necessary.

4) Advancements in Deep Learning: As deep learning architectures and techniques continue to evolve, more complicated big data modelling will be possible, increasing scalability and accuracy.

5) Edge Computing: A promising approach is to directly install machine learning models on edge (Internet of Things) devices to process data locally instead of depending on centralized servers.

6) Human-Machine Collaboration: Improving human-in-the-loop systems, in which machine learning supports human judgment calls rather than taking over completely.

Table 2: Summary of selected studies addressing big data problems using machine learning techniques

Author(s)	Year	Title	ML Techniques Used	Application Area	Key Findings/Contributions
Meena et al. [6]	2020	Traffic Prediction for Intelligent Transportation System using Machine Learning	Deep Learning, Genetic Algorithms	Autonomous vehicles and traffic management	Developed a traffic prediction tool to support autonomous vehicle integration and improve traffic flow accuracy.
Shingate et al.[7]	2020	Adaptive Traffic Control System using Reinforcement Learning	Reinforcement Learning	Traffic light management	Optimized real-time traffic signal timing using deep neural networks and reinforcement learning.
Singh et al.[8]	2021	Machine learning based distributed big data analysis framework for next generation web in IOT	Extreme Learning Machines (ELM), K-means, PCA	Internet of Things (IoT)	Achieved high accuracy in IoT data classification using scalable ML techniques.
Nassar & Kamal [9]	2021	ML & Big Data Analytics in Cybersecurity Threat Detection	Big Data Analytics, Machine Learning	Cybersecurity	Demonstrated real-time threat detection using Big Data and ML, and addressed privacy and ethical issues.
Punia et al. [10]	2021	Performance Analysis of ML Algorithms for Big Data Classification	K-NN, Naïve Bayes, Decision Trees	Social media analytics	Evaluated different ML classification algorithms on social media data.

Author(s)	Year	Title	ML Techniques Used	Application Area	Key Findings/Contributions
Rahul et al. [11]	2021	Machine Learning Algorithms for Big Data Analytics	Supervised Learning, Reinforcement Learning	Industrial Applications	Addressed the effectiveness of ML in sectors such as banking, healthcare, and manufacturing.
Ang & Seng[12]	2021	ML and Big Data With Hyperspectral Information in Agriculture	ML, Deep Learning, Parallel Discriminant Analysis	Agriculture	Improved agricultural productivity by applying ML to hyperspectral data for crop management and soil monitoring.
Kumar et al. [13]	2022	Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research	ML algorithms, Blockchain	Sustainable Finance	Proposed new research avenues for applying ML to green finance and carbon financing.
Manley et al.[14]	2022	A review of machine learning and big data applications in addressing ecosystem service research gaps	ML for uncertainty reduction	Cybersecurity	Addressed the importance of combining ML and Big Data for mapping ecosystem services and cyber security applications.
Deekshetha et al. [15]	2022	Traffic Prediction Using ML	Regression Models, TensorFlow	Intelligent Transportation Systems	Developed real-time traffic prediction system using ML libraries, improving traffic flow predictions.
Khoshaba et al. [16]	2022	Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics	Apache Spark, Apache Mahout	Big Data processing	Compared Apache Spark and Mahout for ML in Big Data environments, highlighting their impact on large dynamic datasets.
Zamani et al. [17]	2024	Multi-disease Prediction in Health Informatics	SSA-EL, DOA-EL, FPA-EL	Healthcare Informatics	Improved multi-disease prediction accuracy using heuristic ML models.
Tayseer et al. [18]	2024	IoT Integration for Machine Learning System using Big Data Processing	ML, Big Data Analytics	E-learning	Explored how Big Data and ML can enhance personalized e-learning experiences and security.
Adewusi et al. [19]	2024	Business Intelligence in the Era of Big Data	Predictive Analytics, ML	Business Intelligence	Highlighted the role of ML in improving business decision-making and gaining competitive advantages through data analytics.
Judijanto et al. [20]	2024	Big Data Technology for Predicting Disease Spread	Integrated Data Sources, Big Data	Healthcare, Infectious Disease Control	Demonstrated improved predictions of disease spread using integrated Big Data platforms for quick decision-making in health crises.

VI. CONCLUSION

Machine learning has emerged as a key enabler for extracting value from big data by supporting automated analysis, predictive modeling, and intelligent decision-making over large and diverse datasets. This paper reviewed the fundamental concepts of big data, discussed supervised, unsupervised, and reinforcement learning techniques, and examined how these methods are applied across domains such as healthcare, finance, retail, transportation, and smart systems. The review highlights the growing reliance on machine learning–driven analytics to handle data complexity and support data-driven operations in modern organizations.

The main contribution of this study lies in providing a structured and consolidated overview of machine learning algorithms for big data analysis, supported by a comparative literature summary that links techniques with application areas. By identifying key research trends and practical challenges, this paper offers useful insights for researchers and practitioners seeking to design effective big data analytics solutions. Future work in this area is expected to emphasize real-time analytics, automated machine learning, edge-based processing, and responsible AI practices, further strengthening the role of machine learning in large-scale data analytics.

REFERENCES

[1] Ishwarappa and J. Anuradha, "A brief introduction on big data 5Vs characteristics and hadoop technology," Procedia Comput Sci, vol. 48, no. C, pp. 319–324, 2015, doi: 10.1016/j.procs.2015.04.188.

[2] M. S. Chowdhury, "Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting," Environmental Challenges, vol. 14, no. October 2023, p. 100800, 2024, doi: 10.1016/j.envc.2023.100800.

[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[4] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to reinforcement learning," Deep Reinforcement Learning: Fundamentals, Research and Applications, pp. 47–123, 2020, doi: 10.1007/978-981-15-4095-0_2.

[5] M. Baratchi et al., Automated machine learning: past, present and future, vol. 57, no. 5. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10726-1.

[6] G. Meena, D. Sharma, and M. Mahrishi, "Traffic Prediction for Intelligent Transportation System using Machine Learning," Proceedings of 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, ICETCE 2020, no. February 2020, pp. 145–148, 2020, doi: 10.1109/ICETCE48199.2020.9091758.

[7] Kranti Shingate, Komal Jagdale, and Yohann Dias, "Adaptive Traffic Control System using Reinforcement Learning," International Journal of Engineering Research and, vol. V9, no. 02, pp. 443–447, 2020, doi: 10.17577/ijertv9is020159.

[8] S. K. Singh, J. Cha, T. W. Kim, and J. H. Park, "Machine learning based distributed big data analysis framework for next generation web in iot," Computer Science and Information Systems, vol. 18, no. 2, pp. 597–618, 2021, doi: 10.2298/CSIS200330012S.

[9] A. Nassar and M. Kamal, "Machine learning and big data analytics for cybersecurity applications: A systematic review," *Journal of Big Data*, vol. 8, no. 1, pp. 1–24, 2021, doi: 10.1186/s40537-021-00441-1.

[10] S. K. Punia, M. Kumar, T. Stephan, G. G. Deverajan, and R. Patan, "Performance analysis of machine learning algorithms for big data classification: Ml and ai-based algorithms for big data analysis," International Journal of E-Health and Medical Communications, vol. 12, no. 4, pp. 60–75, 2021, doi: 10.4018/IJEHMC.20210701.oa4.

[11] K. Rahul, R. K. Banyal, P. Goswami, and V. Kumar, Machine learning algorithms for big data analytics, vol. 1227, no. January. Springer Singapore, 2021. doi: 10.1007/978-981-15-6876-3_27.

[12] K. L. M. Ang and J. K. P. Seng, "Big data and machine learning with hyperspectral information in agriculture," IEEE Access, vol. 9, pp. 36699–36718, 2021, doi: 10.1109/ACCESS.2021.3051196.

[13] S. Kumar, D. Sharma, S. Rao, W. M. Lim, and S. K. Mangla, "Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research," Ann Oper Res, 2022, doi: 10.1007/s10479-021-04410-8.

[14] K. Manley, C. Nyelele, and B. N. Egoh, "A review of machine learning and big data applications in addressing ecosystem service research gaps," Ecosyst Serv, vol. 57, no. September, p. 101478, 2022, doi: 10.1016/j.ecoser.2022.101478.

[15] H. R. Deekshetha, A. V. Shreyas Madhav, and A. K. Tyagi, "Traffic Prediction Using Machine Learning," Lecture Notes on Data Engineering and Communications Technologies, vol. 116, pp. 969–983, 2022, doi: 10.1007/978-981-16-9605-3_68.

[16] F. Khoshaba, S. Kareem, H. Awla, and C. Mohammed, "Machine learning algorithms in Bigdata Analysis and its applications: A review," HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, no. July, pp. 1–8, 2022, doi: 10.1109/HORA55278.2022.9799848.

[17] A. S. Zamani, A. H. A. Hashim, A. S. A. Shatat, M. M. Akhtar, M. Rizwanullah, and S. S. I. Mohamed, "Implementation of machine learning techniques with big data and IoT to create effective prediction models for health informatics," Biomed Signal Process Control, vol. 94, no. April, 2024, doi: 10.1016/j.bspc.2024.106247.

[18] F. Tayseer et al., "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING IoT Integration for Machine Learning System using Big Data Processing," Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE, vol. 2024, no. 14s, pp. 591–599, 2024, [Online]. Available: www.ijisae.org

[19] Adebunmi Okechukwu Adewusi, Ugochukwu Ikechukwu Okoli, Ejuma Adaga, Temidayo Olorunsogo, Onyeka Franca Asuzu, and Donald Obinna Daraojimba, "Business Intelligence in the Era of Big Data: a Review of Analytical Tools and Competitive Advantage," Computer Science & IT Research Journal, vol. 5, no. 2, pp. 415–431, 2024, doi: 10.51594/csitrj.v5i2.791

[20] Loso Judijanto, Hermansyah, K. P. Ningsih, D. Anurogo, and M. Firdaus, "The Role of Big Data Technology in Predicting and Managing the Spread of Infectious Diseases," J. of World Future Medicine, Health and Nursing, vol. 2, no. 2, pp. 219–230, 2024