Global Trends in Science and Technology

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

Machine Learning for Big Data Analytics: A Comprehensive Review

Tsendayush Erdenetsogt[1*], Mehtab Jamal[2]

[1]University of the Potomac, USA

[2]Gomal University, Pakistan

[1]Tsendayush.Erdenetsogt@student.potomac.edu, [2]Mehtabbinjamal@gmail.com

open access

Corresponding Author

Tsendayush Erdenetsogt

Tsendayush.Erdenetsogt@student.potomac.edu

Article History:
Submitted: 10-03-2026
Accepted: 20-04-2026
Published: 25-04-2026

Keywords

Machine Learning, Big Data Analytics, Deep Learning, Data Mining, Scalable Systems, Artificial Intelligence, Real-Time Processing.

ABSTRACT

This review examines the convergence of machine learning and big data analytics, showcasing how machine learning has contributed to big data analytics and the extraction of valuable insights from large-scale data. It explores the properties of big data, principles of machine learning, scalable algorithms and data pipelines that are backed by distributed and cloud computing platforms. The research explores use cases in healthcare, finance, retail, social media, and smart cities, showing the broad reach of data science. Challenges including scalability, data quality, privacy, and interpretability are discussed, as are emerging areas such as federated learning, explainable AI, and edge computing. It concludes that this integration is critical to future "smart, fast and fair" analytics.

INTRODUCTION

The rapid proliferation of digital technologies in recent years has resulted in a massive increase in the generation of data in various areas, including health care, finance, social media, transportation, and research. This trend, known as big data, is not only massive in volume but also in velocity, variety and complexity. Conventional data processing and analysis techniques are no longer sufficient to gain valuable insights from these massive and diverse data sets [1]. This creates a demand for more sophisticated computational techniques to effectively process, analyses and learn from big data in real

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

time. Machine learning (ML), a branch of artificial intelligence, has become a promising approach to tackle these issues [2].

Machine learning offers a way to automatically learn patterns, relationships, and representations from data and build intelligent systems, which can inform data-driven decision-making. By combining machine learning with big data analytics, there are new opportunities for discovery, prediction and automation in a range of sectors [3]. This enables businesses to leverage data to gain insights, which in turn can be used to increase efficiency, enhance customer experience, and develop new business opportunities. Machine learning and big data complement each other. The presence of large data sets improves the learning capability and generalizability of machine learning models, especially in challenging applications such as image classification, language processing, and recommender systems [4].

Machine learning approaches facilitate efficient processing of big data through distributed computing and parallel processing technologies. Other technologies, including cloud computing and supercomputing, also enable this integration by offering the infrastructure to process and train big data models [5]. While the integration of machine learning in big data analytics has great promise, it also presents challenges.

Data quality, scalability, computational efficiency, and privacy are key challenges. Moreover, the importance of interpretability and fairness in machine learning algorithms has become more prominent, particularly in critical domains like healthcare and finance [6]. To overcome these challenges, it is necessary to design effective algorithms, effective data management solutions and ethical guidelines to ensure the responsible deployment of data-driven technologies [7].

This article seeks to give an overview of machine learning and big data analytics. It examines the underlying principles, techniques, technologies and applications that characterize this dynamic area. It also identifies existing challenges and opportunities for future research, providing a glimpse into future trends and prospects. This article aims to provide a comprehensive overview of this field by synthesizing current literature, and to be a useful resource for researchers, practitioners and policymakers alike looking to learn and advance machine learning-based big data analytics.

BIG DATA LANDSCAPE REVISITED

The term big data has undergone a major transformation in the last ten years from its original meaning of simply referring to volume to a more sophisticated notion of volume, context and velocity. Nowadays, big data are a living, evolving landscape of structured, semi-structured and unstructured data produced from various sources such as social media, sensors, mobile devices, enterprise data systems, and scientific instruments. This shift has significantly impacted data collection, storage,

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

processing, and analysis, requiring more sophisticated and flexible analytical methods [8].

Historically, big data has been characterized by the "5Vs": volume, velocity, variety, veracity, and value. Although these attributes still apply, in today's data world we have to consider variability (variable data flows) and visualization (the need to visually communicate complex data insights) [9]. The variety of data formats (from text to images, and streaming sensor data) adds another layer of complexity, so data storage and processing strategies must adapt. With the ever-increasing growth in data, managing the data lifecycle is becoming increasingly important for businesses [10].

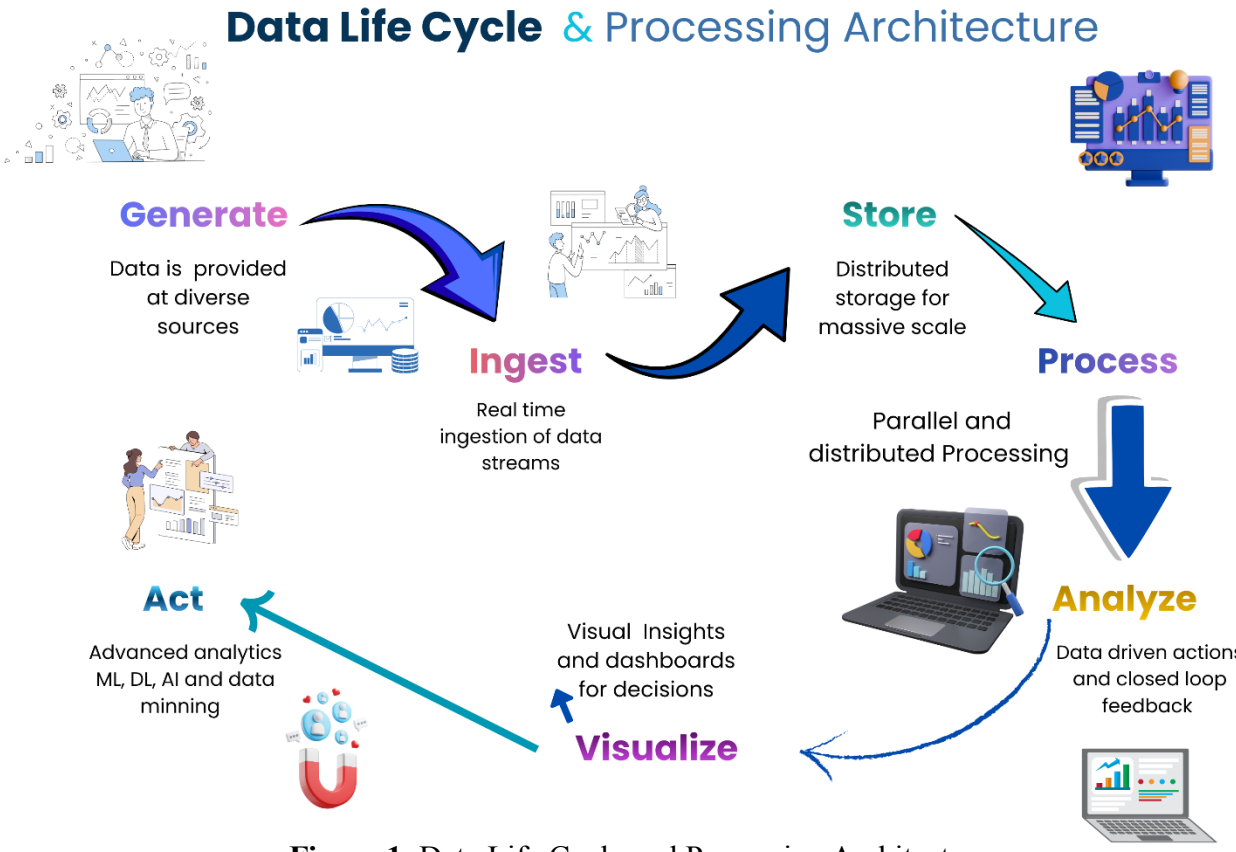


Figure 1. Data Life Cycle and Processing Architecture

The data generation and processing environment is another key component of the big data ecosystem. Data is often generated in a distributed manner, and needs to be ingested and processed in real time. With the emergence of technologies like Internet of Things (IoT) and cloud computing applications, there is a constant flow of data that needs to be ingested and processed in near real-time [11]. This has resulted in the emergence of distributed data architectures that combine distributed data stores and parallel processing platforms to support the processing of large data workloads [12].

Despite improvements in technology and tools, there are a number of analytical challenges in big data processing. Data quality is a key challenge, as large-scale data sets can be noisy, redundant, have missing values and inconsistencies, which can affect the quality of analyses. Moreover, data integration and standardisation is complicated by the diversity of sources. Scalability poses another

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

challenge, as existing algorithms may not be efficient enough to work with large data volumes, resulting in higher computational resources and time [13].

Data security and privacy issues also play a significant role in big data. With the collection and storage of large volumes of sensitive data, data security and regulatory compliance issues are critical. Furthermore, ethical issues such as data use and ownership are increasingly important, especially in the context of personal or behavioral data. The contemporary big data environment is complex, varied, and ever-evolving. [14] This knowledge is crucial to develop analytical approaches and harness the power of data-driven technologies.

MACHINE LEARNING FOUNDATIONS IN A DATA-INTENSIVE ERA

Machine learning (ML) is a crucial component of data analytics, especially in scenarios where data is big, complex and ever-changing. Machine learning, at its essence, is about designing algorithms that learn from data to make predictions or decisions with limited human intervention. In the era of big data, machine learning fundamentals are being revisited and refined to cope with issues of scale, dimensionality, and processing speed [15].

A key characteristic of machine learning is the classification of learning into supervised, unsupervised, and reinforcement learning. In the big data world, these paradigms need to adapt to situations where data is not only large but also dynamic and sometimes unstructured. Supervised learning, which requires labeled data, is hampered by the costs and time required for large-scale data labeling [16]. Unsupervised learning, which is important for uncovering patterns in unlabeled data, is essential in big data applications. Reinforcement learning, while traditionally used in well-defined settings, is also being investigated for big data, real-time decision-making scenarios [17].

Another key aspect is the need to modify algorithms to work with high-dimensional and diverse data. Big data is typically high dimensional with many features that can be redundant and not relevant to the task. This requires the application of dimensionality reduction, feature selection and representation learning techniques to enhance model performance and efficiency [18]. Further, contemporary machine learning increasingly relies on automatic feature learning, particularly with deep learning models, to automatically learn complex features directly from the data without significant human effort.

Assessment of machine learning models in large-scale settings is also challenging. Classic performance metrics like accuracy, precision, recall, and F1-score continue to play a significant role, but need to be interpreted with caution in the presence of data imbalance or data streams [19]. In addition, efficiency and scalability take on significant importance, with models needing to handle large data volumes efficiently in terms of time and computational resources. Methods like cross-

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

validation and batch learning are modified or replaced by more scalable techniques such as online evaluation and distributed cross-validation [20].

In addition, machine learning in big data applications demands the incorporation of sophisticated technologies such as distributed computing platforms and hardware accelerators like GPUs and TPUs. These allow for parallelism and accelerate training, making it possible to use complex algorithms on large data. Through overcoming challenges of scale, complexity and efficiency, machine learning is pushing the boundaries of what is possible in big data analytics [21].

DATA TO KNOWLEDGE: BIG DATA PIPELINES POWERED BY ML

The journey from data to insights is a multi-step process that underpins analytics. For big data, this process is realised via scalable and machine learning (ML)-optimised pipelines that combine data ingestion, preprocessing, model building and deployment. Such pipelines are crafted to manage the diversity, volume and speed of big data, while delivering insights quickly and accurately [23[. Data ingestion is the initial step in an ML-driven big data pipeline. Data is sourced from a variety of often-distributed sources including sensors, user actions, transactional systems and external databases. Because of the velocity and continuous nature of data, both batch and real-time data ingestion is required [24]. Real-time data processing and distributed data stores allow for the ingestion of data streams and efficient storage of data at scale. Scalable and high-speed ingestion is essential to ensure the pipeline's responsiveness.

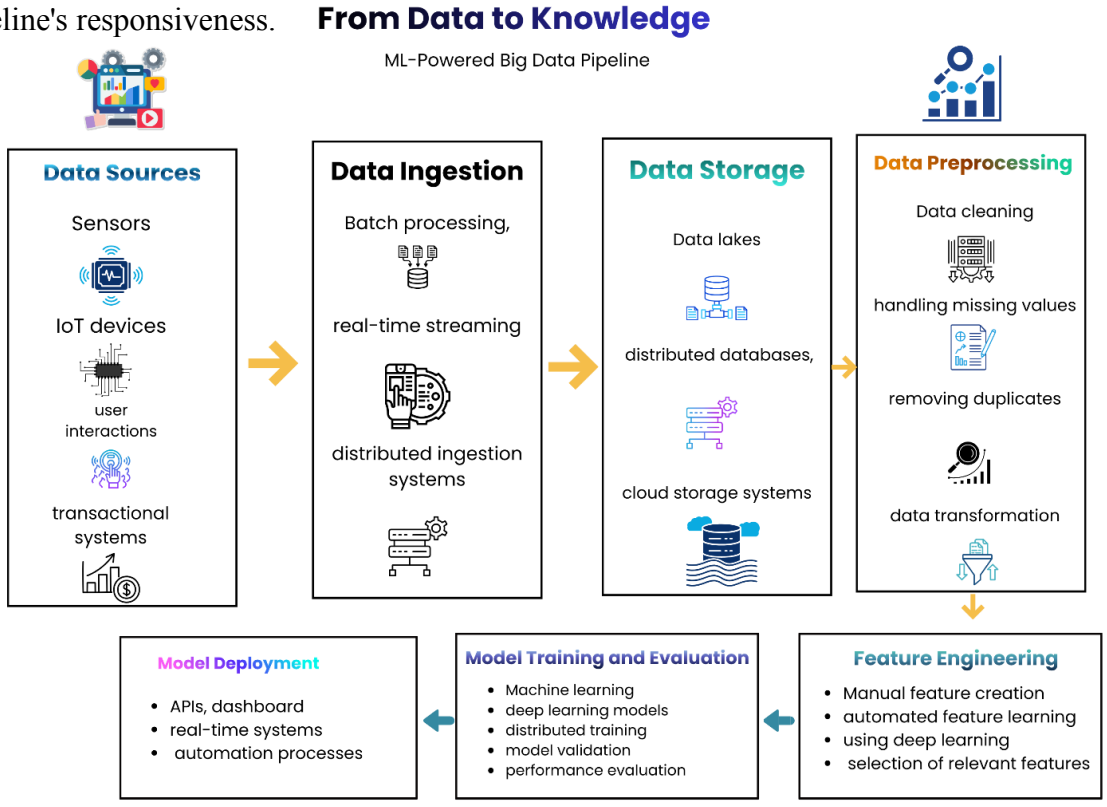


Figure: 2 Machine learning big data pipeline

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

Once data is ingested, data preprocessing is essential for cleaning and transforming data. Big data is frequently dirty, inconsistent, and missing data points, which must be cleaned. This includes techniques like imputing missing data, eliminating duplicates, standardizing data formats, and merging data from various sources [25]. Also, feature engineering or feature learning is done to create relevant variables that enhance model performance. While older methods require the manual creation of features, newer data pipelines increasingly involve feature learning, especially using deep learning models that can automatically learn features from raw data [26].

Models are then deployed for use, where they can make predictions or assist decision-making. This can be through real-time systems, dashboards, or automated processes. Models need to be monitored to maintain their accuracy and effectiveness, particularly as underlying data patterns shift. This brings us to the notion of automated pipelines and model lifecycles, whereby models are continually trained as new data emerges [27]. Machine learning-powered big data pipelines offer a systematic and scalable way to transform large volumes of data into valuable insights. By connecting various stages of the process, these pipelines allow companies to fully leverage machine learning in big data settings [28].

SCALABLE MACHINE LEARNING TECHNIQUES

With the ever increasing size of data, conventional machine learning techniques can lose their efficiency, accuracy and responsiveness. This has given rise to scalable machine learning techniques that are designed to work with big data. These methods aim to modify the algorithms, models and training approaches to manage enormous amounts of data, large numbers of features, and evolving data streams without sacrificing efficiency, speed or accuracy [29]. Parallel and distributed machine learning is a major technique to achieve scalability. Here, datasets are distributed across multiple nodes, enabling parallel execution of machine learning tasks. Distributed systems allow algorithms like decision trees, clustering models, and gradient-based optimization algorithms to parallelise the processing of data to speed up execution [30].

Incremental and online learning is another key approach, particularly in dynamic and streaming data scenarios. Incremental learning is different from batch learning, which trains models on a static dataset, by allowing the model to be updated with new data over time. It avoids the need to retrain the model and allows it to adapt in real time to the changing environment. Online learning is commonly applied in scenarios like recommendation engines, fraud prevention and real-time analytics, where frequent updates are essential to keep the model relevant and accurate [31].

Deep learning has also contributed to scalable machine learning. Deep neural networks have the ability to learn intricate representations from large and unstructured data such as images, text, and

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

audio. For scalability, deep learning models are typically trained on dedicated hardware (such as GPUs and TPUs) and using parallel training approaches such as data parallelism and model parallelism [32]. Training strategies such as mini-batch gradient descent and adaptive learning rates also contribute to the effectiveness of large-scale models.

Aside from single models, hybrid and ensemble techniques have emerged as a means to enhance predictive accuracy in big data scenarios. Ensemble methods aggregate multiple models to enhance prediction accuracy and performance, while hybrid methods blend different learning algorithms to harness their combined power. These approaches can be parallelized in a distributed environment for scalability in big data settings [33]. Scalability comes with its own set of challenges, such as greater system complexity, communication bottlenecks between distributed processors and resource allocation problems. Trade-offs between speed and accuracy are important to consider, especially with constrained computational resources scalable machine learning approaches are vital for enabling big data analytics [34]. These methods allow for efficient data processing and learning from large datasets, ensuring that machine learning can be effectively and practically applied in a world of growing data.

INFRASTRUCTURE AND ECOSYSTEM SUPPORT

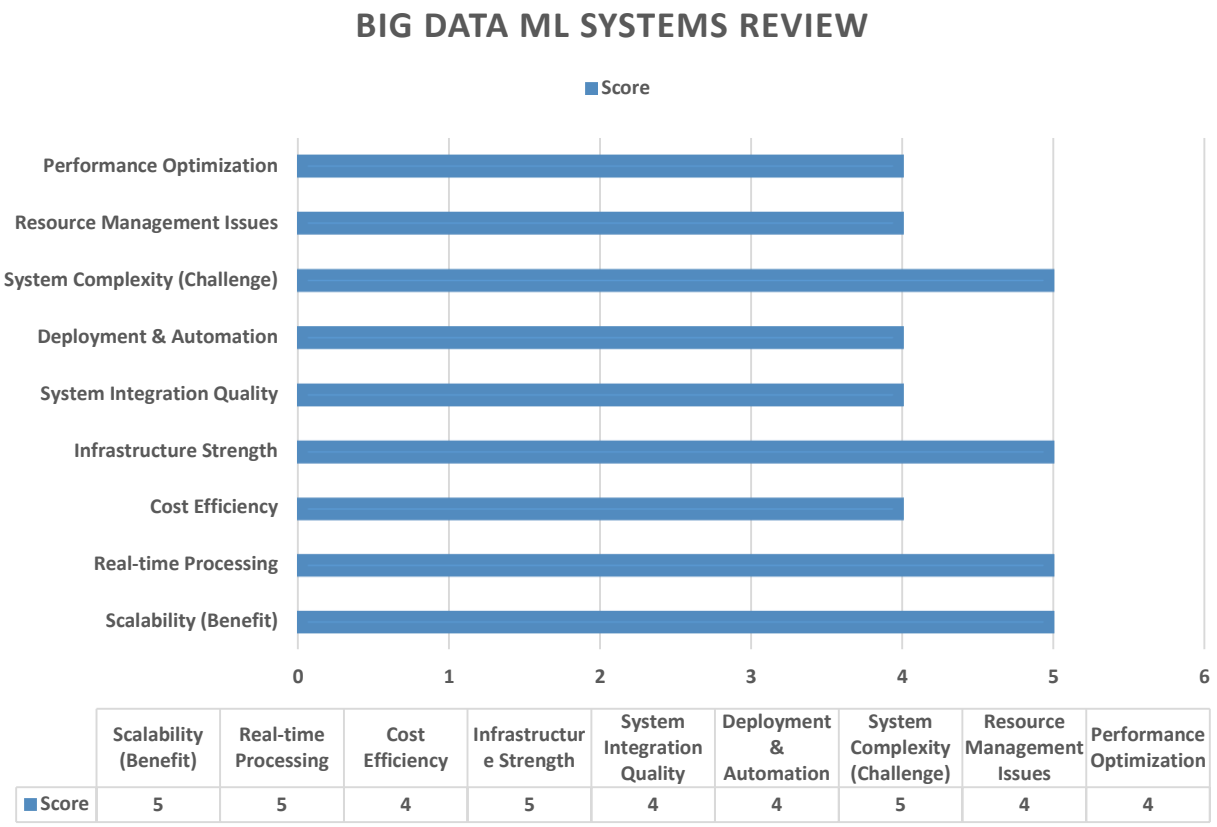
To effectively implement machine learning (ML) for big data analytics, a powerful infrastructure is essential, complemented by a comprehensive ecosystem of tools and technologies. With rapidly increasing data sizes and increasingly sophisticated analytical tasks, the old computing paradigms are not adequate. Big data analytics requires scalable, agile and efficient infrastructure for the management of storage, processing and real-time analytics [35].

Central to this infrastructure are distributed storage and computing systems. Big data systems decentralize data storage to allow simultaneous access and processing, rather than using a centralised approach. This approach not only facilitates scalability but also redundancy, since data is stored in multiple copies. Distributed file systems and cluster computing frameworks enable large data to be divided into smaller pieces and processed in parallel, thereby speeding up the process. These systems are the foundation of large-scale machine learning systems, enabling the training of models on large data sets [36].

Another crucial element in the ecosystem is stream processing systems that process data streams. In many use cases, such as financial transactions, social media analysis and sensor networks, data is arriving in continuous streams and needs to be processed in real-time. Stream processing systems allow data to be ingested, processed and analyzed in real-time, enabling timely detection of patterns and events [37]. This is crucial for real-time applications such as fraud, anomaly and predictive

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

maintenance. The advent of cloud computing has also transformed infrastructure support for big data and machine learning. With the cloud, computing resources, storage, and other services can be provisioned as needed without the initial cost of purchasing hardware. Resources can be dynamically scaled to meet demand, providing cost-effective scalability [38].



	Scalability (Benefit)	Real-time Processing	Cost Efficiency	Infrastructure Strength	System Integration Quality	Deployment & Automation	System Complexity (Challenge)	Resource Management Issues	Performance Optimization
Score	5	5	4	5	4	4	5	4	4

Figure 3. Big Data ML Systems Review

Cloud-based machine learning platforms provide integrated workflows for data preprocessing, training, deployment and monitoring of models. These environments may include features for automation, versioning and collaboration, streamlining the development process. A rich ecosystem of other technologies, including containerization and orchestration technologies, also exist to improve portability and scalability [37]. Containers enable machine learning models and their associated dependencies to be encapsulated in isolated environments, providing a consistent environment from development to production. Orchestration tools assist in managing the containers, automating aspects such as resource management, load balancing and system resilience [38].

However, there are still issues with integration, complexity and resource allocation. It can be challenging to connect different tools and platforms into a unified system, requiring careful planning and expertise. Additionally, ensuring performance and reliability in distributed systems requires

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

ongoing tuning and optimization [39]. Ecosystem support and infrastructure are essential for supporting scalable and efficient machine learning for big data analytics. Through the use of distributed processing, real-time analytics and cloud computing, robust analytical frameworks can be developed to support the needs of contemporary data-intensive applications [40].

CROSS-DOMAIN APPLICATIONS AND CASE STUDIES

The combination of machine learning (ML) and big data analytics has led to game-changing applications in various fields. Through the use of big data and sophisticated learning algorithms, businesses can now discover patterns, make predictions, and automate decision-making. These applications not only showcase the potential of machine learning-supported big data analytics across different domains but also its contributions towards enhancing efficiency, innovation and customer experience [41].

Big data analytics and machine learning have transformed medical diagnosis, treatment, and patient care in the healthcare industry. Big data in healthcare encompasses electronic health records, medical imagery, and genetic data, which can be used to detect patterns and predict outcomes. Machine learning models help healthcare professionals to detect diseases like cancer and heart disease early, and tailoring treatment plans based on individual patient data [42]. Further, continuous monitoring systems leverage real-time data from wearable sensors to offer real-time health monitoring and alerts. Big data analytics powered by machine learning is used in the financial sector for credit risk, fraud detection and trading. Banks and financial institutions handle large volumes of transactional and market data to identify anomalies that could flag fraudulent transactions. Machine learning algorithms can be employed to predict credit risk, personalize investment portfolios and segment customers. Real-time data processing allows for rapid and accurate decision-making in fast-moving markets [43]. Big data and machine learning has also revolutionized the retail and e-commerce sector. Data about customer preferences, purchase patterns, and online interactions can be used to provide product recommendations and targeted promotions. This not only improves customer experience but also boosts revenue and loyalty. Predictive analytics also help with inventory and demand planning, cutting costs and optimizing the supply chain [44]. Social media networks are another major application domain, where a large amount of user-generated data are harnessed to gain insights about trends, sentiments and preferences. Machine learning is applied for sentiment analysis, content recommendation, and content moderation to filter out inappropriate content [45]. This information can be used by companies, governments and academics to gain insight into social trends and sentiments.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

In smart cities and the Internet of Things (IoT), big data analytics using machine learning techniques are used for urban planning, traffic management, energy efficiency and environmental monitoring. Real-time data from sensors and IoT devices are used to enhance public services and urban infrastructure. For instance, smart traffic management systems alleviate congestion, and smart grids save energy [46]. Interdisciplinary applications demonstrate the broad application and benefits of ML-based big data analytics. These examples highlight how big data can be translated into valuable insights, leading to informed decision-making and improved system efficiency across various sectors [47].

KEY CHALLENGES IN ML-BASED BIG DATA ANALYTICS

While significant progress has been made in machine learning (ML) and big data technologies, a number of key challenges remain. These stem from the complexities, volumes, and privacy concerns of data, and the computational resources required for modern analytics platforms. Overcoming these challenges is critical in developing robust, fast and ethical big data analytics systems. A key issue is the balance between scalability and accuracy [48]. As the volume of data increases, machine learning algorithms need to be scaled for efficient processing. But increasing scalability often comes at the cost of model accuracy, as simpler models or approximations are used. On the other hand, predictive accuracy is achieved by complex models like deep neural networks, which are resource-intensive and hard to use for real-world applications. Strike a balance between these two properties is a key challenge in big data analytics [49].

Another critical challenge is data quality, such as bias, noise and missing data. Big data is frequently sourced from various diverse sources, which can lead to inconsistencies, missing data and errors. Data quality issues can severely impact the accuracy and effectiveness of machine learning algorithms, which can result in incorrect predictions and insights [50]. Biased data can lead to biased models, which can perpetuate certain inequalities or lead to unjust outcomes. Cleaning, validation and preprocessing of big data is crucial and difficult.

Data security and privacy is also a critical consideration in ML big data systems. Big data often includes personal, financial, or medical data, which is a prime target for attacks. Encryption, secure storage, and access control are essential for safeguarding this information. Meanwhile, there are efforts to create privacy-preserving technologies such as anonymization, differential privacy, and federated learning to allow data analysis while protecting individual privacy. But these solutions are challenging to implement at large scale [51].

Computational complexity is another limiting factor. Machine learning models trained on large datasets require substantial computational resources. Despite advances in parallel and distributed

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

computing and cloud computing, there are significant challenges in balancing speed, efficiency, cost and energy consumption. Often, companies have to decide which model to use based on a trade-off between complexity and practicality [52].

Moreover, there is a growing concern for the interpretability of machine learning models. Complex models, such as deep learning models, can be treated as "black boxes" in which the decision-making process is not completely understood. This can create barriers to trust and acceptance, particularly in sensitive areas like healthcare, finance, and criminal justice [53]. There are several intertwined issues associated with scalability, data quality, privacy, computational complexity and interpretability in big data analytics using ML. Addressing these challenges is crucial for developing big data technologies that are not only efficient and effective but also safe, ethical, and ready for real-world use [54].

NEW PARADIGMS AND FRONTIERS

Machine learning (ML) for big data analytics is an ever-evolving field, with technological innovations and the need for smarter, faster, and more reliable systems leading to new developments. As existing approaches struggle to cope with the scale, complexity, and timeliness demands, new paradigms are being developed to address the limitations and reshape the landscape of data processing and usage. A key emerging trend is explainable artificial intelligence (XAI) [55]. With increasingly sophisticated machine learning models, especially deep learning, decisions are often made without clear explanations. XAI seeks to create techniques to explain model predictions in a human-interpretable way, while maintaining accuracy [56]. This is particularly crucial in critical applications like healthcare, finance and legal advice, where explanations and trust are necessary. Methods such as feature importance, model simplification and interpretable models are currently being developed to enhance explain ability and trustworthiness [57].

Also, there is an increasing interest in federated learning, which allows machine learning models to be trained across multiple devices or servers without the need to transfer data to a central server. This is especially important in applications where privacy is paramount, like mobile devices, health systems, and banks. Rather than sharing data, model updates are shared, which enhances privacy while allowing for collective learning [58]. Federated learning is a significant advancement towards privacy-preserving artificial intelligence and smart systems.

Big data analytics also includes critical areas of research in edge and fog computing. This approach, as opposed to cloud computing, involves processing data near the source, such as IoT devices and sensors. This helps minimise delays, data transfer, and reliance on centralised systems, and is particularly suited to real-time applications such as autonomous driving, smart cities and manufacturing. Fog computing takes this a step further by introducing a layer between edge devices

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

and cloud computing, facilitating better data management and processing [59].

Moreover, foundation models and large pre-trained models are revolutionizing the field of machine learning. They are pre-trained on large and diverse datasets and can be easily fine-tuned for specific tasks [60]. These models' capacity for cross-domain adaptation has boosted the state-of-the-art in natural language processing, computer vision and multimodal learning. But they also present challenges in terms of computational, bias and environmental sustainability, which are ongoing research topics [61].

Other current trends include automated machine learning (AutoML), which will automate the model selection and hyperparameter tuning process, and green AI, which will reduce the carbon cost of running large-scale computations. The next frontier of machine learning for big data analytics is driven by developments that place a stronger emphasis on explainability, decentralization, speed and scalability [62]. These new paradigms are likely to help overcome existing challenges and to open up exciting opportunities for data intelligence systems.

SYNTHESIS AND FUTURE OUTLOOK

The combination of machine learning and big data analytics is a major technological development of our time. The previous sections clearly demonstrate that this integration has revolutionised the methods of data collection, analysis and understanding, allowing businesses to transition from descriptive to predictive and prescriptive analytics. The integration of all these advances reflects a trend: greater dependence on data-driven systems for decision-making in every aspect of our lives [63].

At an abstract level, machine learning offers the intelligence to draw useful insights from large-scale and complex data sources, while big data systems offer the data scale and variety needed to build effective models. They operate in a symbiotic relationship, with better algorithms making data more useable, and better data making algorithms better [64]. This in turn has led to successful applications in medical diagnosis, financial prediction, and recommendation services, as well as smart urban planning, showcasing the real-world benefits of ML-enabled big data platforms [65].

But the state-of-the-art synthesis also uncovers structural challenges. Scalability, data integration, privacy and interpretability remain major barriers to adoption. Although distributed computing platforms and cloud computing have alleviated some of these challenges, they have also raised new challenges in terms of system integration, cost, and energy efficiency. Additionally, issues related to bias, fairness, and transparency are increasingly important in the ethical use of machine learning systems [66].

The future of the field is likely to see a number of developments. A key direction is the development

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

of more autonomous and adaptive systems able to learn from data streams in a continuous manner without the need for regular retraining. This will be enabled by technologies such as online learning, reinforcement learning and self-supervised learning [67]. Another major trend is a growing focus on trustworthy AI systems, in which explain ability, fairness and accountability will become essential requirements.

In addition, we will see increasing integration of edge computing and big data analytics to bring intelligence to the data source. This will be crucial for low-latency and highly reliable applications, like autonomous vehicles and industrial robots. At the same time, the design of energy-efficient machine learning models will be important to mitigate the environmental footprint of big data analytics [68]. The integration of machine learning and big data analytics is a rapidly evolving area with tremendous potential, as well as enormous risks. Looking ahead, we will see a shift towards more empowered, distributed and ethical data systems. As research progresses, the emphasis will increasingly be on moving beyond handling large amounts of data to meaningful, ethical and useful intelligence for industry and society [69].

CONCLUSION

The review of machine learning for big data analytics showcases an emerging interdisciplinary field that has revolutionized data processing, analysis and use in various areas. The integration of machine learning methods with big data platforms has allowed companies and researchers to overcome conventional data processing constraints and has opened up the possibility of extracting valuable insights from large, complex and diverse data. This shift has not only improved analytical insights but also decision-making in fields including healthcare, finance, retail, transportation and smart cities.

A central message of this review is that big data and machine learning go hand in hand. Big data offers the volume, variety, and complexity of data needed to inform sophisticated machine learning algorithms, and machine learning provides the computational smarts needed to make sense of and extract value from this data. They create a symbiotic relationship that enhances predictive analytics, automation and decision-making. These technologies are being used in a wide range of applications, including fraud prevention in banking, personalised product recommendations in e-commerce, and early diagnosis in medicine.

Yet, while these developments have been made, there remain a number of challenges. Challenges like scalability, data integrity, privacy management, computational overhead, and interpretability still need to be addressed for widespread deployment. The growing dependence on distributed processing and cloud computing platforms has helped to overcome some of these issues, but it has also raised new problems in terms of system complexity, cost-effectiveness and energy efficiency. In addition,

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

ethical concerns regarding algorithmic bias, fairness and transparency are playing an increasingly crucial role, especially in critical applications that impact human lives.

The rise of emerging paradigms like federated learning, explainable AI, edge computing and foundation models suggests a clear trend towards more distributed, interpretable and efficient approaches. These developments seek to overcome conventional challenges and enable the use of machine learning in real-time and privacy-preserving scenarios. Meanwhile, the rise of automated machine learning and sustainable AI highlights the importance of making machine learning more accessible and environmentally friendly.

The next frontier of machine learning in big data analytics is likely to revolve around creating more responsive, smart and ethical systems. Online learning from real-time data streams, enhanced model transparency, and closer integration with edge computing will be key features of future analytics tools. Moreover, the focus on ethical AI will ensure that technological development adheres to societal norms and regulatory standards.

Machine learning for big data analytics is an integral component of data science, contributing to innovation and the development of intelligent systems across various domains. Despite these obstacles, continued research and development efforts are driving innovation in this field. The future of the field is bright and it has the potential to have an even bigger impact as it moves towards more scalable, explainable and smart data-driven technologies.

REFERENCES

[1]. Gupta P, Sharma A, Jindal R. Scalable machine-learning algorithms for big data analytics: a comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2016 Nov;6(6):194-214.'

[2]. Rane NL, Paramesha M, Choudhary SP, Rane J. Machine learning and deep learning for big data analytics: A review of methods and applications. Partners Universal International Innovation Journal. 2024 Jun 25;2(3):172-97.

[3]. Sen S, Agarwal S, Chakraborty P, Singh KP. Astronomical big data processing using machine learning: A comprehensive review. Experimental Astronomy. 2022 Feb;53(1):1-43.

[4]. Kirola M, Memoria M, Dumka A, Joshi K. A comprehensive review study on: optimized data mining, machine learning and deep learning techniques for breast cancer prediction in big data context. Biomedical and Pharmacology Journal. 2022 Mar 31;15(1):13-25.

[5]. Li W, Chai Y, Khan F, Jan SR, Verma S, Menon VG, Kavita F, Li X. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. Mobile networks and applications. 2021 Feb;26(1):234-52.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[6]. Devaraj J, Madurai Elavarasan R, Shafiullah GM, Jamal T, Khan I. A holistic review on energy forecasting using big data and deep learning models. International journal of energy research. 2021 Jul;45(9):13489-530.

[7]. Olaniyi OO, Okunleye OJ, Olabanji SO. Advancing data-driven decision-making in smart cities through big data analytics: A comprehensive review of existing literature. Current Journal of Applied Science and Technology. 2023 Aug 18;42(25):10-8.

[8]. Sarker S, Arefin MS, Kowsher M, Bhuiyan T, Dhar PK, Kwon OJ. A comprehensive review on big data for industries: challenges and opportunities. Ieee Access. 2022 Dec 26;11:744-69.

[9]. El-Sayed A, Abougabal M, Lazem S. Practical big data techniques for end-to-end machine learning deployment: a comprehensive review. Discover Data. 2025 Apr 15;3(1):11.

[10]. Salkuti SR. A survey of big data and machine learning. International Journal of Electrical and Computer Engineering (IJECE). 2020 Feb 15;10(1):575-80.

[11]. Nti IK, Quarcoo JA, Aning J, Fosu GK. A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics. 2022 Jan 25;5(2):81-97.

[12]. Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K. Efficient machine learning for big data: A review. Big Data Research. 2015 Sep 1;2(3):87-93.

[13]. Singh N, Singh DP, Pant B. A comprehensive study of big data machine learning approaches and challenges. In2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS) 2017 Dec 11 (pp. 80-85). IEEE.

[14]. Ameen DD, Kareem SW, Hasan SB. A Big Data, Bigger Impact: A Comprehensive Review of Machine Learning Advancements. In2024 International Conference on Electrical Engineering and Computer Science (ICECOS) 2024 Sep 25 (pp. 1-6). IEEE.

[15]. Naeem S, Ali A, Anam S, Ahmed MM. An unsupervised machine learning algorithms: Comprehensive review. International Journal of Computing and Digital Systems. 2023 Mar 2.

[16]. Ponnusamy VK, Kasinathan P, Madurai Elavarasan R, Ramanathan V, Anandan RK, Subramaniam U, Ghosh A, Hossain E. A comprehensive review on sustainable aspects of big data analytics for the smart grid. Sustainability. 2021 Dec 1;13(23):13322.

[17]. Zhang W, Gu X, Tang L, Yin Y, Liu D, Zhang Y. Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge. Gondwana Research. 2022 Sep 1;109:1-7.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[18]. Szymańska E. Modern data science for analytical chemical data–A comprehensive review. Analytica chimica acta. 2018 Oct 22;1028:1-0.

[19]. Ahmed A, Xi R, Hou M, Shah SA, Hameed S. Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. IEEE Access. 2023 Oct 10;11:112891-928.

[20]. Ashqar RI, Ramos CM. Machine-learning holistic review in tourism and hospitality. InThe International Conference on Global Economic Revolutions 2023 Feb 27 (pp. 78-84). Cham: Springer Nature Switzerland.

[21]. Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: A comprehensive review. IEEE access. 2020 Dec 31;9:4843-73.

[22]. Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M. Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials. 2018 Jun 6;20(4):2923-60.

[23]. Zhang Q, Yang LT, Chen Z, Li P. A survey on deep learning for big data. Information Fusion. 2018 Jul 1;42:146-57.

[24]. Wang J, Xu C, Zhang J, Zhong R. Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems. 2022 Jan 1;62:738-52.

[25]. Jha K, Doshi A, Patel P, Shah M. A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture. 2019 Jun 1;2:1-2.

[26]. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors. 2018 Aug 14;18(8):2674.

[27]. Kakani AB, Nandiraju SK, Chundru SK, Vangala SR, Polam RM, Kamarthapu B. Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. International Journal of Emerging Trends in Computer Science and Information Technology. 2021 Jun 30;2(2):26-34.

[28]. Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, Agushaka JO, Eke CI, Akinyelu AA. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering applications of artificial intelligence. 2022 Apr 1;110:104743.

[29]. Himeur Y, Elnour M, Fadli F, Meskin N, Petri I, Rezgui Y, Bensaali F, Amira A. AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial intelligence review. 2022 Oct 15;56(6):4929.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[30]. Sharifani K, Amini M. Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal. 2023;10(07):3897-904.

[31]. Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS journal of photogrammetry and remote sensing. 2019 Jun 1;152:166-77.

[32]. Mienye ID, Swart TG, Obaido G. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. Information. 2024 Aug 25;15(9):517.

[33]. Kibria MG, Nguyen K, Villardi GP, Zhao O, Ishizu K, Kojima F. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. IEEE access. 2018 May 17;6:32328-38.

[34]. Botín-Sanabria DM, Mihaita AS, Peimbert-García RE, Ramírez-Moreno MA, Ramírez-Mendoza RA, Lozoya-Santos JD. Digital twin technology challenges and applications: A comprehensive review. Remote Sensing. 2022 Mar 9;14(6):1335.

[35]. Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Machine learning and knowledge extraction. 2023 Nov 20;5(4):1680-716.

[36]. Sarker IH, Kayes AS, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. Journal of Big data. 2020 Jul 1;7(1):41.

[37]. Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An introductory review of deep learning for prediction models with big data. Frontiers in artificial intelligence. 2020 Feb 28;3:4.

[38]. Rai R, Tiwari MK, Ivanov D, Dolgui A. Machine learning in manufacturing and industry 4.0 applications. International Journal of Production Research. 2021 Aug 18;59(16):4773-8.

[39]. Wang J, Lu S, Wang SH, Zhang YD. A review on extreme learning machine. Multimedia Tools and Applications. 2022 Dec;81(29):41611-60.

[40]. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-Dabbagh BS, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, Duan Y. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data. 2023 Apr 14;10(1):46.

[41]. Bahroun Z, Anane C, Ahmed V, Zacca A. Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability. 2023 Aug 29;15(17):12983.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[42]. Nath AG, Udmale SS, Singh SK. Role of artificial intelligence in rotor fault diagnosis: a comprehensive review. Artificial Intelligence Review. 2021 Apr 1;54(4).

[43]. Baryannis G, Validi S, Dani S, Antoniou G. Supply chain risk management and artificial intelligence: state of the art and future research directions. International journal of production research. 2019 Apr 3;57(7):2179-202.

[44]. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends. 2020 May 15;1(1):56-70.

[45]. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics. 2018 Jun 1;114:57-65.

[46]. Sircar A, Yadav K, Rayavarapu K, Bist N, Oza H. Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research. 2021 Dec 1;6(4):379-91.

[47]. Wang S, Cao J, Philip SY. Deep learning for spatio-temporal data mining: A survey. IEEE transactions on knowledge and data engineering. 2020 Sep 22;34(8):3681-700.

[48]. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar SS. A survey on deep learning: Algorithms, techniques, and applications. ACM computing surveys (CSUR). 2018 Sep 18;51(5):1-36.

[49]. Roh Y, Heo G, Whang SE. A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering. 2019 Oct 8;33(4):1328-47.

[50]. Wang Y, Chen Q, Hong T, Kang C. Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Transactions on smart Grid. 2018 Mar 22;10(3):3125-48.

[51]. Da Costa KA, Papa JP, Lisboa CO, Munoz R, de Albuquerque VH. Internet of Things: A survey on machine learning-based intrusion detection approaches. Computer Networks. 2019 Mar 14;151:147-57.

[52]. Çınar ZM, Abdussalam Nuhu A, Zeeshan Q, Korhan O, Asmael M, Safaei B. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Sustainability. 2020 Oct 5;12(19):8211.

[53]. Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J, Zhang Y. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water research. 2020 Mar 15;171:115454.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[54]. Çınar ZM, Abdussalam Nuhu A, Zeeshan Q, Korhan O, Asmael M, Safaei B. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Sustainability. 2020 Oct 5;12(19):8211.

[55]. Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y, Livingood W. A review of machine learning in building load prediction. Applied Energy. 2021 Mar 1;285:116452.

[56]. Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems. 2022 Mar 7;34(11):8135-53.

[57]. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN computer science. 2021 Nov;2(6):1-20.

[58]. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. Journal of medical systems. 2018 Nov;42(11):226.

[59]. Nguyen G, Dlugolinsky S, Bobák M, Tran V, Lopez Garcia A, Heredia I, Malík P, Hluchý L. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review. 2019 Jun;52(1):77-124.

[60]. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence. 2021 Feb 17;44(7):3523-42.

[61]. Dargan S, Kumar M, Ayyagari MR, Kumar G. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning: S. Dargan et al. Archives of computational methods in engineering. 2020 Sep;27(4):1071-92.

[62]. Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep learning approach for intelligent intrusion detection system. IEEE access. 2019 Apr 3;7:41525-50.

[63]. Misra NN, Dixit Y, Al-Mallahi A, Bhullar MS, Upadhyay R, Martynenko A. IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet of things Journal. 2020 May 29;9(9):6305-24.

[64]. Ahmad Z, Shahid Khan A, Wai Shiang C, Abdullah J, Ahmad F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies. 2021 Jan;32(1):e4150.

ISSN : 3079-5761
Volume 2, Number 2, 2026
https://doi.org/10.70445/gtst.2.2.2026.199-218

[65]. Ma X, Wu J, Xue S, Yang J, Zhou C, Sheng QZ, Xiong H, Akoglu L. A comprehensive survey on graph anomaly detection with deep learning. IEEE transactions on knowledge and data engineering. 2021 Oct 8;35(12):12012-38.

[66]. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics. 2023 Aug 25;13(17):2760.

[67]. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data. 2021 Mar 31;8(1):53.

[68]. Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, Park CW, Choudhary A, Agrawal A, Billinge SJ, Holm E. Recent advances and applications of deep learning methods in materials science. npj Computational Materials. 2022 Apr 5;8(1):59.

[69]. Zhang D, Yin J, Zhu X, Zhang C. Network representation learning: A survey. IEEE transactions on Big Data. 2018 Jun 25;6(1):3-28.