International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Deep Learning for Big Data: A Survey of Architectures, Distributed Frameworks, and Emerging Challenges

Vijayakumar Soundrapandian

Department of Computer Science, Chhatrapati Shivaji Maharaj University

Abstract

Big data has become a defining feature of modern computing, and deep learning has emerged as the primary tool for extracting predictive value from such large-scale, high-velocity, heterogeneous data. This survey reviews the intersection of deep learning and big data along three axes: the neural architectures used to model large-scale data, the distributed computing frameworks (notably those built on Apache Spark) that make training such models tractable, and the persistent challenges of scalability, data quality, privacy, and interpretability that constrain real-world deployment. We further examine emerging responses to these challenges, including federated learning for privacy-preserving distributed training, explainable AI (XAI) for large-scale models, and edge/TinyML approaches for resource-constrained deployment. The survey closes with open research directions, including communication-efficient distributed training, robustness to non-IID data, and standardized benchmarking for big data deep learning systems.

Keywords: Big Data, Deep Learning, Apache Spark, Federated Learning, Explainable AI.

I. INTRODUCTION

Data volumes worldwide continue to grow at an exponential rate, with global data creation projected to reach into the hundreds of zettabytes by the mid-2020s [1]. This growth is commonly characterized along the dimensions of volume, variety, velocity, and veracity — the so-called "4 Vs" of big data [2], [3]. Extracting actionable insight from data at this scale exceeds the capacity of traditional statistical and shallow machine learning methods, which typically rely on hand-engineered features and do not scale efficiently to high-dimensional, high-volume inputs [4].

Deep learning, by contrast, learns hierarchical feature representations directly from raw data and has demonstrated strong performance across vision, language, and structured-data tasks. Its combination with big data platforms has therefore become a central research theme in both the machine learning and data engineering communities [5], [6]. This survey synthesizes recent literature on deep learning for big data, organized around architectures, distributed training frameworks, domain applications, and open challenges, with the goal of providing researchers and practitioners a structured entry point into the field.

II. BIG DATA CHARACTERISTICS

The 4 Vs framework — volume, variety, velocity, and veracity — captures the core difficulties of big data

analysis [2], [3]. Volume refers to the sheer scale of data generated by modern systems; variety reflects the heterogeneity of structured, semi-structured, and unstructured data sources; velocity concerns the rate at which data must be ingested and processed, often in near real time; and veracity addresses noise, incompleteness, and uncertainty in the data itself. Foundational work on big data in business and scientific contexts has emphasized that value is only realized when organizations can analyze data at this scale rather than merely store it [7], [8].

Deep neural networks are well suited to this setting because they can absorb large volumes of heterogeneous, weakly structured data and automatically learn useful representations, reducing dependence on manual feature engineering [4], [5]. However, realizing this advantage in practice requires computational infrastructure capable of training large models on datasets that do not fit in the memory of a single machine, which motivates the distributed frameworks discussed in Section IV.

III. DEEP LEARNING ARCHITECTURES

A range of neural architectures has been adapted for big data settings. Convolutional neural networks (CNNs) remain dominant for image and spatial data at scale, while recurrent architectures (RNNs, LSTMs) and, increasingly, transformer-based models are used for sequential and time-series data arising from sensor streams, logs, and text corpora [5], [6]. Surveys of the field note that the choice of architecture in big data pipelines is frequently driven as much by data engineering constraints — such as how data is partitioned and streamed — as by pure modeling accuracy [6]. This has led to growing interest in architectures explicitly designed for distributed or streaming ingestion, alongside standard model families adapted from smaller-scale settings.

IV. DISTRIBUTED TRAINING FRAMEWORKS

Because big data by definition exceeds single-machine capacity, training deep learning models on such data typically requires distributed computing frameworks. Apache Spark has become a common substrate for this integration, given its widespread adoption for large-scale data processing. Early systems such as DeepSpark distributed model replicas across Spark clusters and aggregated results using asynchronous elastic-averaging stochastic gradient descent [9]. Related frameworks — including CaffeOnSpark, DeepLearning4j, and SparkNet — similarly bridge Spark's data-parallel execution model with deep learning training loops, each offering different trade-offs between data locality, fault tolerance, and GPU utilization [10], [11].

More recent work has extended these ideas to mobile and remote-sensing big data pipelines, where deep learning models are trained directly over data resident in distributed file systems such as HDFS [12], [13]. Spark's own ecosystem has evolved to natively support this workload: as of Spark 3.4, built-in APIs such as TorchDistributor for PyTorch and the spark-tensorflow-distributor package allow distributed model training and inference to be expressed without third-party glue frameworks [14]. Broader surveys of the Spark ecosystem catalog these developments alongside Spark's core data-processing components, illustrating the convergence of big data engineering and deep learning tooling [15].

V. DOMAIN APPLICATIONS

Deep learning over big data has been applied across numerous domains. In healthcare, IoT-enabled smart health systems combine wearable and clinical sensor streams with machine learning and deep learning models to support diagnosis and monitoring at population scale [16]. In earth observation, distributed deep

learning on Spark has been used for geological remote-sensing interpretation, where imagery volumes routinely exceed single-node processing capacity [13]. These applications share a common pattern: raw data arrives continuously from many distributed sources and must be processed, modeled, and acted upon with an infrastructure that can scale horizontally, reinforcing the centrality of the distributed frameworks described in Section IV.

VI. KEY CHALLENGES

A. Scalability and Data Quality

Even with distributed training infrastructure, the quality of big data pipelines constrains model performance. Data-centric perspectives on deep learning emphasize that collection, labeling, and curation issues — rather than model architecture alone — are often the binding constraint on real-world performance at scale [17].

B. Privacy

Centralizing large volumes of sensitive data for training raises significant privacy and regulatory concerns. Federated learning has emerged as the primary architectural response, allowing model training across distributed data owners without centralizing raw data; recent surveys catalog both the opportunities this creates for big data applications and the accompanying challenges of communication overhead, system heterogeneity, and non-IID data distributions across clients [18], [19], [20], [21].

C. Interpretability

As deep models grow larger and are applied to higher-stakes big data applications, the ability to explain their predictions becomes increasingly important for trust, auditability, and regulatory compliance. Recent surveys distinguish inherent interpretability from post-hoc explainability techniques such as LIME and SHAP, and note the growing use of these methods in industry applications operating on large-scale data [22], [23], [24].

VII. EMERGING SOLUTIONS AND FUTURE DIRECTIONS

A complementary line of work addresses big data challenges by pushing computation toward the network edge rather than centralizing it. Surveys of deep learning for edge computing describe architectures for deploying models on resource-constrained devices to reduce latency and preserve data locality, extending naturally to IoT-generated big data streams [25], [26], [27]. TinyML research pushes this further, targeting extremely constrained microcontroller-class devices [28], while recent work explores combining edge deployment with large language models for intelligent, low-latency data analysis [29].

Taken together, federated learning, explainable AI, and edge/TinyML represent three complementary strategies for reconciling the scale of big data with practical constraints of privacy, trust, and resource availability — and each remains an active area of ongoing research.

VIII. OPEN RESEARCH DIRECTIONS

Several open problems remain. First, communication-efficient distributed and federated training algorithms are needed to reduce the overhead that currently limits scalability across heterogeneous clients [20]. Second, robustness to non-IID and low-veracity data remains an open modeling challenge, particularly in federated and streaming settings [19], [21]. Third, standardized benchmarks that jointly evaluate accuracy, training efficiency, privacy guarantees, and interpretability are largely absent from the

current literature, making cross-study comparison difficult. Finally, energy and cost efficiency of large-scale distributed training is an increasingly pressing concern as model and data sizes continue to grow.

IX. CONCLUSION

Deep learning and big data have become deeply intertwined: big data provides the scale that deep models need to realize their representational advantages, while deep learning provides a mechanism for extracting value from data whose scale exceeds traditional analytical methods. This survey has reviewed the architectures, distributed frameworks, applications, and challenges that define this intersection, and highlighted federated learning, explainable AI, and edge computing as the leading responses to the field's core constraints of privacy, trust, and resource availability. Continued progress will depend on advances in communication-efficient distributed training, robustness to real-world data conditions, and standardized evaluation practices.

REFERENCES

1. Data growth statistics reported in recent big data and edge computing literature project global data creation reaching into the hundreds of zettabytes range by the mid-2020s.
2. S. Sagiroglu and D. Sinanc, "Big data: A review," in Proc. 2013 Int. Conf. on Collaboration Technologies and Systems (CTS), 2013.
3. J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," National Science Review, vol. 1, no. 2, pp. 293–314, 2014.
4. X.-W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, vol. 2, pp. 514–525, 2014.
5. "A Survey on Deep Learning in Big Data," in Proc. IEEE Int. Conf., 2017 (IEEE Xplore Document 8005992).
6. "Exploring the Intersection of Machine Learning and Big Data: A Survey," Big Data and Cognitive Computing, vol. 7, no. 1, art. 13, 2023.
7. T. H. Davenport, P. Barth, and R. Bean, "How 'big data' is different," MIT Sloan Management Review, vol. 54, no. 1, pp. 43–46, 2012.
8. A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data," Harvard Business Review, vol. 90, no. 10, pp. 61–67, 2012.
9. "DeepSpark: A Spark-Based Distributed Deep Learning Framework for Commodity Clusters," arXiv:1602.08191, 2016.
10. "Deep Learning Frameworks on Apache Spark: A Review," ResearchGate preprint, 2018.
11. "Spark based distributed Deep Learning framework for Big Data applications," in Proc. IEEE Int. Conf. (IEEE Xplore Document 7777390), 2016.
12. "Mobile Big Data Analytics Using Deep Learning and Apache Spark," arXiv:1602.07031, 2016.
13. "Distributed Deep Learning for Big Remote Sensing Data Processing on Apache Spark: Geological Remote Sensing Interpretation as a Case Study," in Web and Big Data, Springer, 2024.
14. "Distributed Deep Learning Made Easy with Spark 3.4," NVIDIA Technical Blog, 2023.
15. "A Survey on Spark Ecosystem for Big Data Processing," arXiv:1811.08834, 2018.
16. "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System," PMC7786888, 2021.

17. "Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective," arXiv:2112.06409, 2021.
18. "A Comprehensive Survey of Federated Learning: Privacy, Security, and Applications," IEEE Xplore Document 11447669, 2025.
19. Q. Han et al., "Privacy preserving and secure robust federated learning: A survey," Concurrency and Computation: Practice and Experience, 2024.
20. "Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence," arXiv:2504.17703, 2025.
21. "Federated learning for big data: A survey on opportunities, applications, and future directions," Engineering Applications of Artificial Intelligence, ScienceDirect, 2025.
22. "Explaining explainability: A comprehensive survey on explainable artificial intelligence and relevant industry applications," ScienceDirect, 2026.
23. "Interpretability research of deep learning: A literature survey," Information Fusion, ScienceDirect, 2024.
24. "Explainable artificial intelligence (XAI): from inherent explainability to large language models," arXiv:2501.09967, 2025.
25. "Edge Deep Learning in Computer Vision and Medical Diagnostics: A Comprehensive Survey," arXiv:2605.06714, 2025.
26. "Deep Learning for Edge Computing Applications: A State-of-the-Art Survey," IEEE Xplore Document 9044329, 2020.
27. "Edge Computing with Artificial Intelligence: A Machine Learning Perspective," ACM Computing Surveys, 2022.
28. "Intelligence at the Extreme Edge: A Survey on Reformable TinyML," arXiv:2204.00827, 2022.
29. "Intelligent data analysis in edge computing with large language models: applications, challenges, and future directions," Frontiers in Computer Science, 2025.