International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 13; Issue: 4; April 2026
Website: www.ijrrjournal.com
Review Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Review on Deep Network Accelerators Towards Healthcare and Biomedical Applications

Pallavi Anil Talnikar[1], Hemangi Ravindra Zunjarrao[2]

[1]Lecturer in Automation & Robotics Department, Marathwada Mitra Mandal's Polytechnic, Thergaon, Pune, Maharashtra.
[2]Lecturer in, Automation & Robotics Department, Marathwada Mitra Mandal's Polytechnic, Thergaon, Pune, Maharashtra.

Corresponding Author: Pallavi Anil Talnikar

DOI: https://doi.org/10.52403/ijrr.20260436

ABSTRACT

Recent trends in deep learning (DL) imposed hardware accelerators as the most viable solution for several classes of high-performance computing (HPC) applications such as image classification, computer vision, and speech recognition. This survey summarizes and classifies the most recent advances in designing DL accelerators suitable to reach the performance requirements of HPC applications. In particular, it highlights the most advanced approaches to support deep learning accelerations including not only GPU and TPU-based accelerators but also design-specific hardware accelerators such as FPGA-based and ASIC-based accelerators, Neural Processing Units, open hardware RISC-V-based accelerators and co-processors. The survey also describes accelerators based on emerging memory technologies and computing paradigms, such as 3Dstacked Processor-In-Memory, non-volatile memories (mainly, Resistive RAM and Phase Change Memories) to implement in-memory computing, Neuromorphic Processing Units, and accelerators based on Multi-Chip Modules. The survey classifies the most influential architectures and technologies proposed in recent years, to offer the reader a comprehensive perspective in the rapidly evolving field of deep learning.

Index Terms: Deep network accelerators, deep learning, high-performance computing

INTRODUCTION

The large availability of biomedical data presents both tremendous opportunities and significant challenges for healthcare research. In particular, identifying associations among diverse data sources is a fundamental problem in developing reliable, data-driven medical tools based on machine learning approaches. To address this, previous studies have attempted to integrate multiple data sources and construct unified knowledge bases for predictive analysis and discovery. Although these models have shown promising results, machine learning-based predictive tools have not yet been widely adopted in clinical practice.

Healthcare is entering a new era in which abundant biomedical data are playing an increasingly important role. In this context, precision medicine aims to ensure that the right treatment is delivered to the right patient at the right time by considering various aspects of patient data, including molecular traits, environmental factors, electronic health records (EHRs), and lifestyle.

Several challenges hinder the effective utilization of biomedical data. These include high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity.

Additionally, the use of different medical ontologies, such as SNOMED-CT, UMLS, and ICD-9, introduces inconsistencies and conflicts. The same clinical condition may also be represented in multiple ways across datasets. For example, a patient with type 2 diabetes mellitus may be identified through laboratory values (e.g., hemoglobin A1C > 7.0), diagnostic codes (e.g., ICD-9 code 250.00), or textual descriptions in clinical notes. As a result, harmonizing these diverse representations into a unified semantic structure is a complex task.

Traditionally, biomedical research relies on domain experts to manually define relevant features or phenotypes. However, this supervised approach does not scale well and may overlook hidden patterns in the data. In contrast, representation learning techniques enable automatic extraction of meaningful features directly from raw data. Deep learning, a form of representation learning, uses multiple layers of nonlinear transformations to learn increasingly abstract representations of the input data.

Deep learning models have demonstrated remarkable success in fields such as computer vision, speech recognition, and natural language processing. Given their strong performance and continuous methodological advancements, deep learning offers promising opportunities for biomedical informatics. Several initiatives are already exploring its application in healthcare. For instance, Google DeepMind has announced efforts to apply deep learning techniques in healthcare, while Enlitic utilizes deep learning for detecting abnormalities in X-rays and CT scans.

Despite these advancements, deep learning has not yet been extensively evaluated across a wide range of medical problems. Nevertheless, it offers several advantages, including superior performance, end-to-end learning with integrated feature extraction, and the ability to handle complex, multi-modal data. To fully realize its potential in healthcare, further research is required to address challenges related to data characteristics such as sparsity, noise, heterogeneity, and temporal dependencies. Additionally, improved methods and tools are needed to integrate deep learning systems with healthcare workflows and clinical decision support systems.

LITERATURE REVIEW

Bio-signals play a significant role in health monitoring and disease diagnosis because they provide critical information about a person's physiological, pathophysiological, and emotional states. With the emergence of machine learning algorithms, intelligent bio-signal processing has become available, resulting in various sought-after biomedical applications. For example, Attia et al. proposed a rapid, inexpensive method based on a convolutional neural network (CNN) to detect the signature of atrial fibrillation using an electrocardiogram (ECG).[1]

Now a days a big part of data is being generated by IoT devices, used as "big data" and input feed by deep learning algorithms to yield meaningful information. Deep learning has numerous implications and is almost available in every aspect of our life. Medical uses and healthcare are among the most popular ones, and are still growing. EHRs, integrated administrative and medical data bases, digital images (radiography, mammography, and histology), data taken from mobile applications, medical devices IoT, genes data, and data coming from search engines are main sources of deep learning algorithms that can predict, diagnose, help decide clinically, etc. some others are used in biomedical and pharmaceutical fields such as molecular diagnostics, pharmacogenomics, identification of pathogenic variants, DNA Sequencing, gene splicing, personalized cancer care, and drug discovery [2].

Deep learning, by contrast, learns representations directly from raw data. It uses computational models made up of multiple processing layers, usually organized as neural networks, to automatically learn features at multiple levels of abstraction. The main differences between deep learning and traditional artificial neural networks (ANNs) lie in the depth of the architecture, the nature

of the connections between layers, and the ability to learn meaningful abstractions of the inputs. Traditional ANNs generally have only a few layers (often three) and are trained in a supervised manner to perform a specific task, with limited ability to generalize beyond it [3].

The Deep Neural Network delivers state-of-the-art performance in many applications. The complexity of the DNN models generally increases with application complexity, and the deployment of complex DNN models requires high computational power. General-purpose processors are unable to process complex DNNs within the required throughput, latency, and power budget. Therefore, domain-specific hardware accelerators are required to provide high computational resources with superior energy efficiency and throughput within a small chip area. In this paper, existing DNN hardware accelerators are reviewed and classified based on the optimization techniques used in their implementations. Each optimization technique generally improves one or more specific performance parameter(s). For example, the hardware optimized for sparse DNNs may provide poor performance for dense DNNs in terms of power and throughput. Therefore, understanding the tradeoff between different hardware accelerators helps to identify the best accelerator model for application deployment. We identify three major areas, ALU, dataflow, and scarcity, in hardware architectures having the potential to improve the overall performance of an accelerator. Existing hardware accelerators for inference are broadly classified into these three categories. As there is no standard model or performance metrics to evaluate the efficiency of the new DNN hardware in the literature, the classification model can help to identify appropriate performance parameters and benchmark accelerators [4].

Various types of data have been emerging in modern biomedical research, including electronic health records, imaging, sensor data, and text, which are complex, heterogeneous, poorly annotated, and generally unstructured. Traditional data mining and statistical learning approaches typically need to first perform feature engineering to obtain effective and more robust features from the data, and then build prediction or clustering models on top of them. There are many challenges at both steps in a scenario with complex data and insufficient domain knowledge. The latest advances in deep learning technologies provide new, effective paradigms to obtain end-to-end learning models from complex data. In this article, we review the recent literature on applying deep learning technologies to advance the healthcare domain. Based on the analyzed work, we suggest that deep learning approaches could be the vehicle for translating big biomedical data into improved human health. However, we also note limitations and needs for improved methods development and applications, especially in terms of ease-of-understanding for domain experts and citizen scientists. We discuss such challenges and suggest developing holistic and meaningful interpretable architectures to bridge deep learning models and human interpretability [5].

In deep learning, each layer transforms the input it receives from the layer below into increasingly abstract representations, often by optimizing local unsupervised objectives. A key advantage is that these features are not engineered by humans—they are learned automatically from data through a general-purpose learning procedure. Conceptually, deep neural networks process inputs layer by layer in a nonlinear fashion, using pre-training to initialize deeper layers and capture “deep structures” that generalize well. Finally, a supervised training phase fine-tunes the entire network via back propagation, optimizing it for the specific end-to-end task [6].

For many years, building a machine learning system required extensive engineering and domain expertise to manually convert raw data into meaningful internal representations. These representations allowed the learning component—often a classifier—to identify

patterns within the dataset. Traditional approaches typically relied on a single, often linear, transformation of the input space, which limited their ability to handle natural, unprocessed data effectively [7].

After reviewing the literature on these DL accelerators, we quantify the performance of various algorithms on different types of DL processors. The results allow us to draw a perspective on the potential future of spike-based neuromorphic processors in the biomedical signal processing domain. Based on our analysis and perspective, we conjecture that, for edge processing, neuro-morphic computing and Spiking Neural Networks (SNNs) [8] will likely complement DL inference engines, either through signaling anomalies in the data or acting as 'intelligent always-on watchdogs' which continuously monitor the data being recorded, but only activate further processing stages if and when necessary.

Deep learning methods are representation-learning algorithms with multiple levels of representation, obtained by composing simple but nonlinear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level [23]. Deep learning models demonstrated great performance and potential in computer vision, speech recognition and natural language processing tasks [24–27]

We expect this tutorial, review and perspective to provide guidance on the history and future of DL accelerators, and the potential they hold for advancing healthcare. Our contributions are summarized as follows:

- Our paper is the first to discuss the use of three different emerging and established hardware technologies for facilitating DL acceleration, with a focus on biomedical applications.
- We provide tutorial sections on how one may implement a typical biomedical task on FPGAs or simulate it for deployment on memristive crossbars.
- Our paper is the first to discuss how event-based neuromorphic processors can complement DL accelerators for biomedical signal processing.
- We provide open-source code and data to enable the reproduction of our results.

APPLICATIONS OF DEEP LEARNING IN HEALTHCARE

Deep learning approaches have already been successfully applied in several domains, particularly in computer vision and natural language processing. Existing literature clearly demonstrates the strong capabilities of deep learning in healthcare data analysis. The use of multi-layer neural networks for processing medical data has significantly enhanced predictive performance across various clinical applications and domains.

One of the key strengths of deep learning lies in its hierarchical learning structure, which enables the integration of diverse and heterogeneous data sources. Unlike traditional models that primarily focus on classification accuracy, deep learning emphasizes representation learning. This allows for better generalization and improved performance when dealing with complex healthcare datasets. Deep learning has the potential to drive the next generation of predictive healthcare systems. These systems can: Scale to handle millions or even billions of patient records.

Utilize a unified, distributed patient representation to support clinicians in their daily activities, rather than relying on multiple fragmented systems. An ideal healthcare framework would integrate various data sources such as Electronic Health Records (EHRs), genomics, environmental data, wearable devices, and social activity information. This would enable a comprehensive and holistic representation of an individual's health status. In such a system, deep learning models can be embedded into healthcare platforms (e.g., hospital EHR systems) and continuously updated to adapt to changes in patient populations. These deep

representations can support a wide range of clinical applications, including:
Disease risk prediction
Personalized treatment and prescriptions
Clinical decision support
Clinical trial recruitment
Medical research and data analysis
A notable example is the work by Wang et al., who won the Parkinson's Progression Markers Initiative data challenge. They used a temporal deep learning approach based on Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) to identify subtypes of Parkinson's disease. Since Parkinson's disease is progressive in nature, traditional vector- or matrix-based approaches are insufficient for capturing temporal disease patterns. In contrast, the LSTM model effectively identified three distinct subtypes, each representing specific disease progression trends.
This example highlights the significant potential of deep learning in addressing real-world healthcare challenges and developing more reliable and robust automated systems. More broadly, deep learning can serve as a guiding framework for both hypothesis-driven research and exploratory analysis in clinical domains. It supports tasks such as clustering, visualization of patient cohorts, and disease stratification. To fully realize this potential, it is essential to integrate statistical and medical considerations at every stage, including study design, experimental planning, model development, refinement, and data interpretation.
The descriptive statistics from Table 4.1 showed that the values were normally distributed about their mean and variance. This indicated that aggregate stock prices on the KSE and the macroeconomic factors, inflation rate, oil prices, exchange rate, and interest rate are all not too much sensitive to periodic changes and speculation. To interpret, this study found that an individual investor could not earn higher rate of profit from the KSE. Additionally, individual investors and corporations could not earn higher profits and interest rates from the economy and foreign companies could not earn considerably higher returns in terms of exchange rate. The investor could only earn a normal profit from KSE.

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: No conflicts of interest declared.

Ms. Pallavi Anil Talnikar completed her Master of Engineering in VLSI & Embedded System at the Savitribai Phule Pune University of Pune, Maharashtra, India. She is currently a senior lecturer in the Department of Automation and Robotics Engineering at Marathwada Mitra Mandal's Polytechnic, Pune

Ms. Hemangi Ravindra Zunjarrao completed her Master of Engineering in Electronic and Telecommunications at the Savitribai Phule Pune University of Pune, Maharashtra, India. She is currently a senior lecturer in the Department of Automation and Robotics Engineering at Marathwada Mitra Mandal's Polytechnic, Pune.

REFERENCES

1. S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. on Neural Networks*, vol. 4, pp. 570–578, July 1993.
2. J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436–44.
4. C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Trans. Image Process.*, vol. 10, no. 5, pp. 767–782, May 2001.
5. Cichocki and R. Unbehaven, *Neural Networks for Optimization and Signal*

Processing, 1st ed. Chichester, U.K.: Wiley, 1993, ch. 2, pp. 45–47.
6. W.-K. Chen, *Linear Networks and Systems*, Belmont, CA: Wadsworth, 1993, pp. 123–135.
7. H. Poor, *An Introduction to Signal Detection and Estimation*; New York: Springer-Verlag, 1985, ch. 4.
8. R. A. Scholtz, "The Spread Spectrum Concept," in *Multiple Access*, N. Abramson, Ed. Piscataway, NJ: IEEE Press, 1993, ch. 3, pp. 121–123.
9. G. O. Young, "Synthetic structure of industrial plastics," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
10. M. B. Kasmani, "A Socio-linguistic Study of Vowel Harmony in Persian (Different Age Groups Use of Vowel Harmony Perspective," *International Proceedings of Economics Development and Research*, ed. Chen Dan, pp. 359–366, vol. 26, 2011.
11. W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *Proc. 1987 INTERMAG Conf.*, 1987, pp. 2.2-1–2.2-6.
12. G. W. Juette and L. E. Zeffanella, "Radio noise currents n short sections on bundle conductors," presented at the IEEE Summer Power Meeting, Dallas, TX, June 22–27, 1990.
13. J. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
14. N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
15. J. P. Wilkinson, "Nonlinear resonant circuit devices," U.S. Patent 3 624 12, July 16, 1990.
16. *Letter Symbols for Quantities*, ANSI Standard Y10.5-1968.
17. *Transmission Systems for Communications,* 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44–60.
18. *Motorola Semiconductor Data Manual,* Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.
19. R. J. Vidmar. (August 1992). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21(3). pp.876–880.Available: http://www.halcyon.com/pub/journals/21ps03-vidmar
20. Gottlieb A, Stein GY, Ruppin E, et al. A method for inferring medical diagnoses from patient similarities. BMC Med 2013; 11:194.
21. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013; 35:1798–828.
22. Farhan W, Wang Z, Huang Y, et al. A predictive model for medical events based on contextual embedding of temporal sequences. J Med Internet Res 2016;4: e39.
23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436–44.
24. Abdel-Hamid O, Mohamed A-R, Jiang H, et al. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2014; 22:1533–45.
25. Deng L, Li X. Machine learning paradigms for speech recognition: an overview. IEEE Trans Audio Speech Lang Process 2013; 21:1060–89.
26. Cho K, Courville A, Bengio Y. Describing multimedia content using attention-based encoder–decoder networks. ArXiv 2015. http://arxiv.org/abs/1507.01053
27. Hannun A, Case C, Casper J, et al. Deep speech: scaling up end-to-end speech recognition. arXiv 2014. https://arxiv.org/abs/1412.5567
28. Google's DeepMind forms health unit to build medical software. https://www.bloomberg.com/news/articles/2016-02- 24/google-s-deepmind-forms-health-unit-to-build-medical-software (4 August 2016, date last accessed).
29. Enlitic uses deep learning to make doctors faster and more accurate. http://www.enlitic.com/index.html (29 November 2016, date last accessed).
30. Bengio Y, Lamblin P, Popovici D. Greedy layer-wise training of deep networks. Adv Neural Inf Process Syst 2007; 19:153–60

How to cite this article: Pallavi Anil Talnikar, Hemangi Ravindra Zunjarrao. Review on deep network accelerators towards healthcare and biomedical applications. *International Journal of Research and Review*. 2026; 13(4): 363-368. DOI: *https://doi.org/10.52403/ijrr.20260436*

******