Journal of Neuromorphic Intelligence

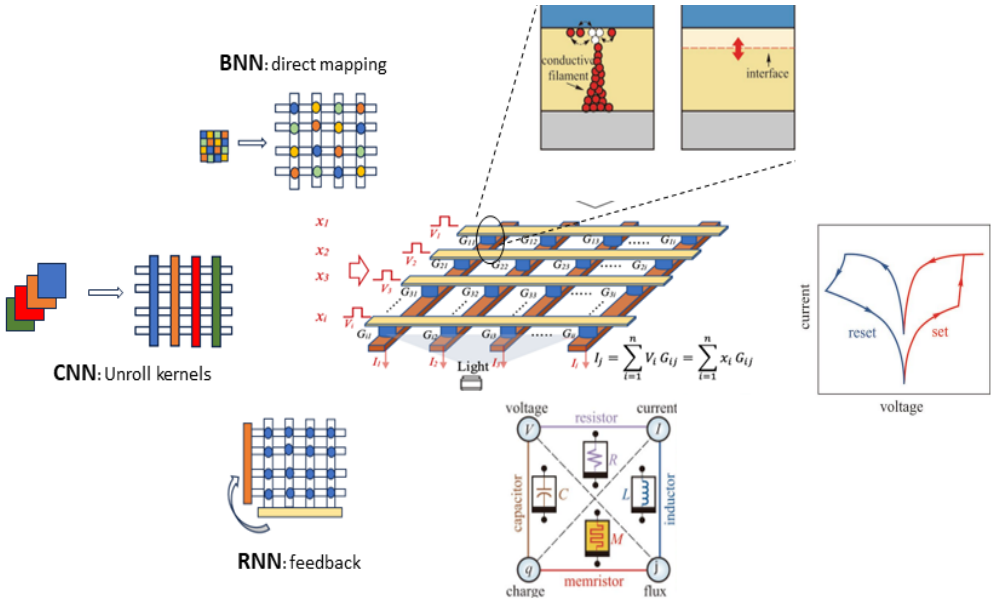
REVIEW ARTICLE

VOLUME 01 | JULY 2024

Bridging Synapses: A Comparative Review of Machine Learning Algorithms in Memristor Technology

Prashant Dhakal[1], Jae Gwang Kim[2], Aolin Hou[2], Xiaofei Wu[1], Shiren Wang[1,2,*]

1 Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA
2 Department of Material Science and Engineering, Texas A&M University, College Station, TX, USA
** Corresponding author: s.wang@tamu.edu*



Received: 10 May 2024

Accepted: 5 June 2024

Published: 10 July 2024

Check for updates

ABSTRACT

The integration of memristor-based nanodevices into machine learning systems has garnered significant attention due to their rapid, low-energy, and non-volatile switching capabilities. Memristors, with their unique ability to retain an internal resistance state based on voltage and current history, excel in executing core operations like vector matrix multiplication, addressing the von Neumann bottleneck. This review explores the synergy between memristors and machine learning, highlighting their potential to enhance neuromorphic computing through energy-efficient and highly parallel architectures. The paper examines current research, practical implementations, and emerging challenges, providing a comprehensive analysis of the future directions in this interdisciplinary field.

Keywords: ***Neuromorphic computing, Neuromorphic engineering, Bio-sensor, Wearable devices***

1. INTRODUCTION

In the pursuit of advancing computing technologies, the spotlight has turned towards memristor-based nanodevices, heralded for their scalability down to sub-10 nm dimensions and their prowess in delivering rapid, low-energy, and non-volatile electrical switching capabilities. Memristors, quintessentially described as 'memory resistors', have the unique ability to preserve an internal resistance state reflective of the voltage and current history applied to them. This characteristic sets them apart, as their functionality transcends the capabilities of traditional resistors, capacitors, and inductors[1]. The essence of memristors extends beyond their simple circuit representation, offering a glimpse into their microscopically altered states via an external two-terminal resistance. Initially conceptualized as devices linking charge and magnetic flux, memristors have evolved in definition to be characterized by a pinched-hysteresis loop, with its size being frequency-dependent[2]. The integration of memristors in the realm of machine learning presents an exciting frontier. Machine learning, the field dedicated to enabling systems to learn from data and improve from experience, has seen its algorithms, especially artificial neural networks (ANNs), become indispensable in processing complex, real-world environments. These networks, which demand substantial computational and memory resources, have found applications across a spectrum of areas including speech recognition and natural language processing [3]. Memristors excel in executing core operations like vector matrix multiplication (VMM) within dense crossbar arrays, leveraging analog computation laws for efficient processing. This approach epitomizes the "computing by physics" paradigm, offering a significant boost in performance for matrix-intensive tasks and addressing the von Neumann bottleneck—a major constraint in system performance [4,5]. Neuromorphic computing, gaining momentum as an alternative to von Neumann architectures, benefits markedly from memristors. Coined by Carver Mead in 1990, neuromorphic computing today embodies systems inspired by biological neural networks, distinguished by their highly parallel and low-power nature. Memristors, in this context, shine as the most ubiquitous component, drawing parallels to biological synapses in their ability to exhibit STDP-like behavior. The diversity of materials from which memristors can be fabricated further underscores their adaptability and potential for energy-efficient circuitry, aligning with the fault-tolerant characteristics of neural network models to counteract device variation effects. Despite the burgeoning interest and preliminary successes in weaving together the narratives of memristors and machine learning, there exists a discernible void in comprehensive analyses that synergize these interdisciplinary pursuits. This review is poised to bridge this gap, embarking on an exhaustive exploration of the confluence between memristors and machine learning algorithms. It scrutinizes an array of algorithms optimized for memristor-based systems, evaluates their applicability across various domains, and delineates the attendant benefits and challenges of these groundbreaking computing paradigms. Our discourse is bifurcated into a critique of the prevailing research landscape involving memristors within machine learning frameworks, and an exploration of nascent challenges and promising avenues for research, particularly emphasizing the co-design of algorithms and memristive hardware to maximize the efficacy of artificial intelligence applications. The rest of the paper is organized as follows: Section 2 talks about Machine Learning algorithms with theoretical underpinnings and practical implementations. Section 3 broadens the discussion to encompass the multifaceted applications of ML integration with memristor. The final section offers reflections and perspectives on future research directions, charting a course for further inquiry into this compelling intersection of disciplines.

2. MACHINE LEARNING ALGORITHMS

As the digital age advances, the quest for more efficient, faster, and smaller computing technologies becomes increasingly critical. Within this context, memristors, heralded for their unique electrical properties, emerge as a pivotal innovation. These devices, characterized by their ability to remember previous states of electrical resistance, have sparked a revolution in computing paradigms, particularly in the realm of machine learning. The intrinsic ability of memristors to perform computations directly within memory architectures not only promises to dramatically enhance computational speed and reduce energy consumption but also opens up new avenues for the development of sophisticated machine learning algorithms. This section delves into the transformative role of memristors in machine learning, exploring their integration across various algorithmic frameworks—from neural networks to unsupervised learning and beyond. By facilitating direct, in-memory computation, memristors offer a compelling solution to the traditional bottlenecks encountered in machine learning operations, such as the von Neumann bottleneck, thereby heralding a new era of efficiency and performance in artificial intelligence applications

2.1 ANN BASED ML ALGORITHMS

Artificial neural networks (ANNs) are designed based on how human brains work and can adjust to different tasks because of how they are connected. They have been a game-changer for machine learning, offering new ways to handle tasks like understanding language, recognizing images, and spotting objects[6][7]. But, as ANNs get better, they also need more data storage and processing power[8], which can be a problem, especially for devices with limited computing resources like smartphones and IoT devices. To solve this, researchers are looking into using memristor arrays to make ANNs run faster. This has become a very interesting area of study. Memristor-based neural networks mainly focus on two things: making decisions (inference) and learning (training). We look at some of the key studies on how to build these networks, especially for making decisions.

The simplest kind of ANN is the Multilayer Perceptron (MLP), which is just layers connected together that process inputs and can be easily used with memristor arrays. MLPs are popular for early experiments with memristor networks[9-11]. For example, in 2018, Hu and colleagues showed how a single-layer network could recognize handwritten digits using a special kind of memristor setup[12] consisting of 128 X 64 1T1R

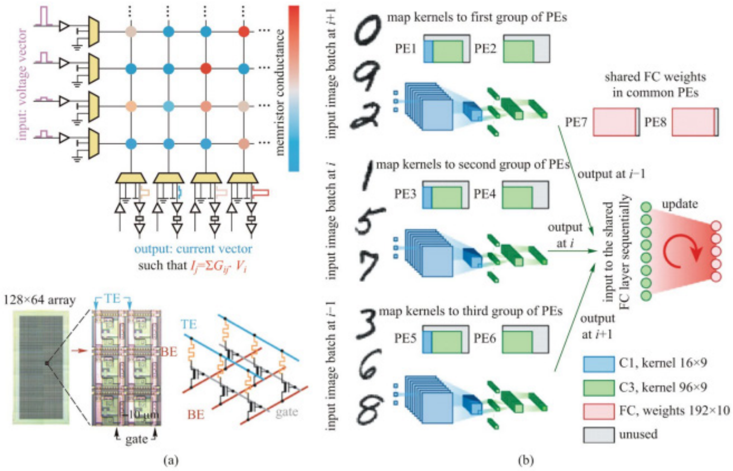


Figure 1. Schematics of the hardware-implemented memristive neural network. a MVM scheme with memristor arrays and the periphery circuits using the fabricated 128 × 64 1T1R array. Adapted from Ref. [21]. b Sketch of the hardware system operation flow with the duplicated convolutional weights and the hybrid training method. Adapted from Ref. [41]. The Authors, published by Intelligent Computing, under a Creative Commons Attribution License 4.0 (CC BY 4.0). Reprinted with permission.

array, achieving an accuracy of 89.9% on a standard test (fig 2a). This setup converted between analog and digital signals, combining the best of both worlds: the efficiency of memristors and the flexibility of digital computing. More recently, Kim and team [13]demonstrated an MLP on a simpler, passive memristor array 64 X 64. They found ways to deal with the challenges of programming these devices accurately thorough statistic of the target conductance with respect to the initial state and the programming pulses over the fabricated 4096 devices.

Another approach uses Convolutional Neural Networks (CNNs), which are great for tasks like image recognition. In 2020, Yao and team [14]made a CNN using memristors 128 X 16 1T1R chips(fig 2b) that was much more energy-efficient than traditional computer chips. Another study by Lin and colleagues[15] proposed a new 3D memristor array design with a customized eight layer 3D memristor array for the experimental demonstration of hardware-implemented CNN that could process data more effectively, showing great potential for future applications.
Besides MLPs and CNNs, researchers are exploring other types of memristor-based networks for different uses, like processing video or translating languages. These studies show how versatile memristor networks can be, opening up new possibilities for computing technology.

2.1 NON ANN BASED ML ALGORITHMS

Even though artificial neural networks (ANNs) have gotten really good at speeding up certain calculations, making big strides in learning and decision-making processes, there's still a big challenge: as data gets more complex, it gets harder to handle. This problem, known as the "curse of dimensionality," makes it tough to work with data that's spread out and hard to compare because it's so high-dimensional. Solutions to the curse of dimensionality could be dimension ignorable computation or efcient dimension reduction methods. Using in-memory computing, we can process this complex data much faster because it doesn't slow down, no matter how complicated the data gets. However, for machine learning tasks that need special ways of organizing or estimating data, memristor arrays come into play, proving to be really helpful in making these tasks faster.

Table 1. Summary of the representative experimentally demonstrated memristor arrays

Array type	Memristor Material	Array Size	ML Algorithm	References
1T1R	Ti/HfOx/TiN	128 x 64	BNN	[15]
1T1R	TiN/TaOx/HfOx/TiN	512 x 1024	BNN	[16]
1T1R	W/TiN/TiON	-	ANN	[17]
1T1R	TiN/TaOx/HfOx/TiN	-	CNN	[18]
1T1R	TiN/TaOx/HfAlO/TiN	-	MLP	[19]
1T1R	Ta/HfOx/Pt	128 x 64	MLP,LSTM, RL	[20-22]

When it comes to analyzing and searching data, finding similarities is crucial but also really affected by the curse of dimensionality. Techniques like measuring Euclidean distances(ED) or cosine similarity between data points are used in various algorithms, from grouping data to finding similar items or organizing maps. The most widely studied experiments with memristors for calculating these similarities focused on the most common method, Euclidean distance, showing that memristors can speed up these calculations. Yu et al [23] used dot products from a memristor array to estimate Euclidean distance, applying it to classify orientations in a network, a simple approach but effective for certain tasks. This introduced the idea of measuring similarities with memristor arrays for the first time. And, Jeong et al[24], realized a fully hardware-based ED calculation method to achieve the K-means data clustering algorithm, which is widely utilized in various applications(fig 2a-b).
Another strategy to manage complex data is to pick out the important features and simplify the data. Principal Component Analysis (PCA) is a common method that reorganizes the original data into new, unrelated groups, focusing on the most important parts. Choi et al. [25] showed that memristor arrays can speed up this detailed feature selection process in PCA, using an unsupervised learning approach, highlighting how memristors can not only handle straightforward calculations quickly but also more complicated data organization tasks (fig 2c-d).

3.APPLICATIONS OF MACHINE LEARNING WITH MEMRISTORS

Combining memristor technology with machine learning has completely changed how we think about computing and has opened up new possibilities for exciting uses in different areas. Thanks to special features of memristors, like their brain-like abilities and the way they can process data right where it's stored, we're looking at a future with faster processing, better energy use, and improved data storage. We're on the edge of a big change in technology, with memristor-based machine learning showing up in everything from small devices that put AI in our hands to advanced systems that make us rethink how computers work. This part talks about the main areas where memristors and machine learning come together, showing how they're changing fields like manufacturing, health care. By looking into these uses, we see how memristors can improve current machine learning methods and even lead to new kinds of algorithms and models we haven't thought of yet.

3.1 EDGE COMPUTING

Artificial intelligence (AI) is being used more and more in embedded applications like monitoring patients and ensuring safety in buildings and industries. These applications prefer to process data locally (at the "edge" of the network) for better security and to use less energy for sending data back and forth. However, putting AI into environments with very limited energy sources is tough because AI usually needs a lot of power, which often means AI tasks have to be done in the cloud or closer to the user ("the fog") instead of on the device itself. A promising way to solve this is by using systems based on memristors, which can greatly cut down the amount of energy AI needs to run. This could lead to edge AI systems that don't need batteries because they can get energy from their surroundings. Memristors are special

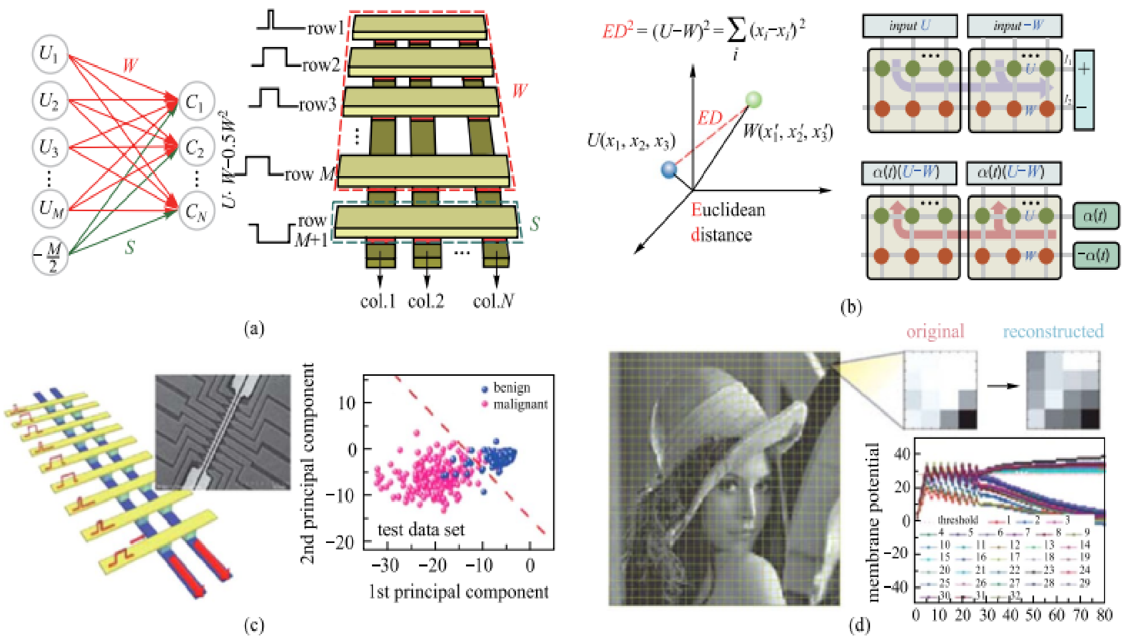


Figure 2. Machine learning algorithms on memristor arrays. a. K-means data clustering by calculating −2U W_n + S_n for data comparison. Adapted from Ref. [23]. b Mapping method and operation steps of the Euclidean distance engine for competitive learning. Adapted from Ref. [24]. c Implementation of PCA to the memristor arrays and the experimental verifcation with the Breast Cancer data set. Adapted from Ref. [25]. d Demonstration of the sparse coding. Adapted from Ref. [25].

because they can keep data even when there's no power, which is a big plus.

There's a growing trend to move AI tasks from big, power-hungry data centers to smaller devices like smartphones and smartwatches, even though these devices have limited battery life and processing power. Memristors, which let us do computations right where data is stored, could make these devices much more energy-efficient. With more and more smart devices needing to process lots of data efficiently, edge computing becomes key. Edge computing means moving data processing from powerful cloud servers to the devices at the very edge of the internet, like sensors that gather data from the real world. Since a lot of this real-world data is unorganized (like images or sounds), it's crucial to use neural network techniques, like deep learning, right on these devices to make sense of it all. Relying only on cloud servers for this could lead to unsustainable levels of power use. So, it's expected that in 2 to 5 years, edge computing will become very common as deep learning, the Internet of Things, and smart sensors improve and help edge computing grow.

Memristor-based neural networks fit well into tiny edge-computing systems and are inspired by how the human brain works. These networks could one day replace the traditional computing system we use now. Memristors offer many benefits for creating new kinds of computing systems that don't follow the old rules: they can be made very small, use very little energy, keep data without power, and can be stacked in 3D, making them very efficient for future technologies.

In a recent study by Wang etal [26], a novel noise distribution normalization (NDN) method is proposed to enhance privacy in edge AI applications, particularly for IoT systems. This method uniquely employs the inherent cycle-to-cycle variations of memristors as a natural source of Gaussian distributed noise, bypassing the need for additional software or hardware overhead. By integrating this noise directly during the weight-update process, the study showcases a memristor-based solution that achieves differential privacy without extra circuitry. A case study implementing this method in differentially private stochastic gradient descent (DP-SGD) demonstrates a 3.5% to 15.5% improvement in recognition accuracy over traditional approaches. This advancement not only capitalizes on memristors' variability for privacy protection but also proposes a practical and efficient hardware implementation for enhancing data privacy in edge AI systems.

Jean-Michel Portal's[27] work introduces a binarized neural network powered by solar energy, featuring 32,768 memristors optimized for edge applications. This innovative circuit, designed for digital near-memory computing, operates effectively in varied illumination, shifting seamlessly to approximate computing under low light without needing external calibration. Demonstrated through image classification tasks, the network maintains functionality and accuracy even in low-power scenarios, showcasing its potential for self-powered AI in fields like health and environmental monitoring. The approach marks a significant step towards sustainable, efficient edge computing.

3.1 NEUROMORPHIC COMPUTING

In recent years, the development of Compute-in-Memory (CIM) technology has advanced alongside

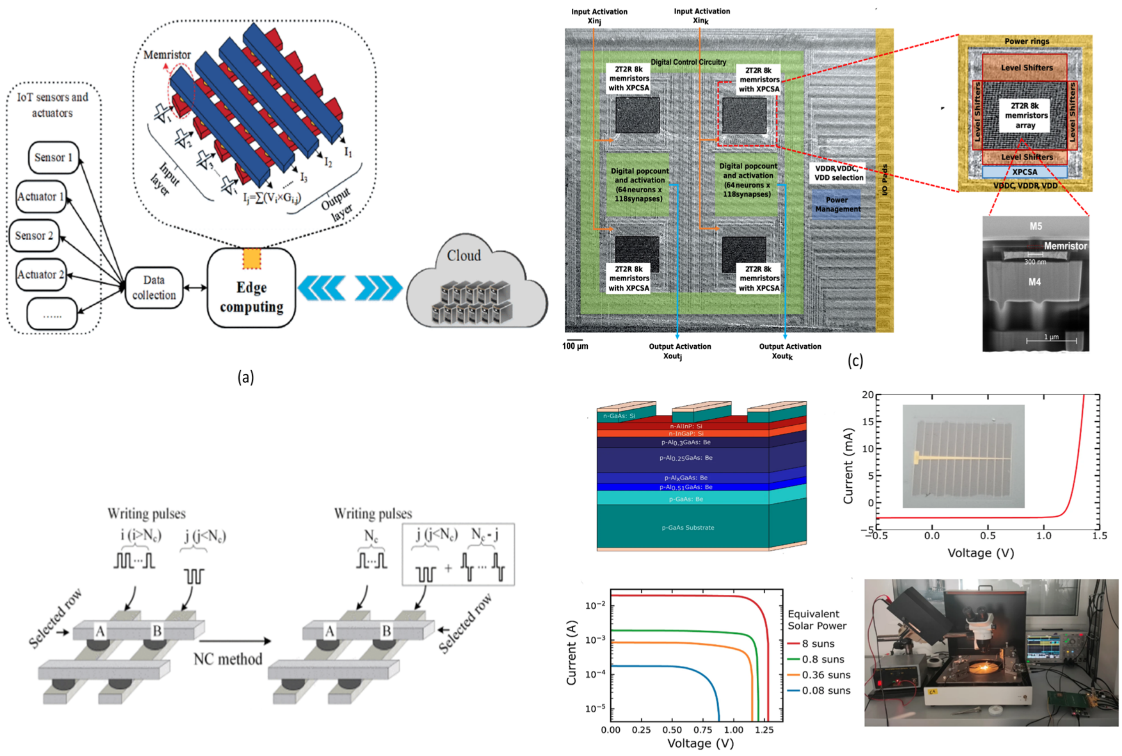


Figure 3. a. Hardware implementation of neural networks using memristor crossbar Adapted from Ref. [26]. b Workflow of the NDN method, where n represents the number of pulses that are used to update a weight and m represents the number of positive and negative PNs.Adapted from Ref. [26]. c. Overview of the fabricated memristor-based binarized neural network.d. Measurements of the binarized neural network powered by a miniature solar cell. Adapted from Ref. [27]

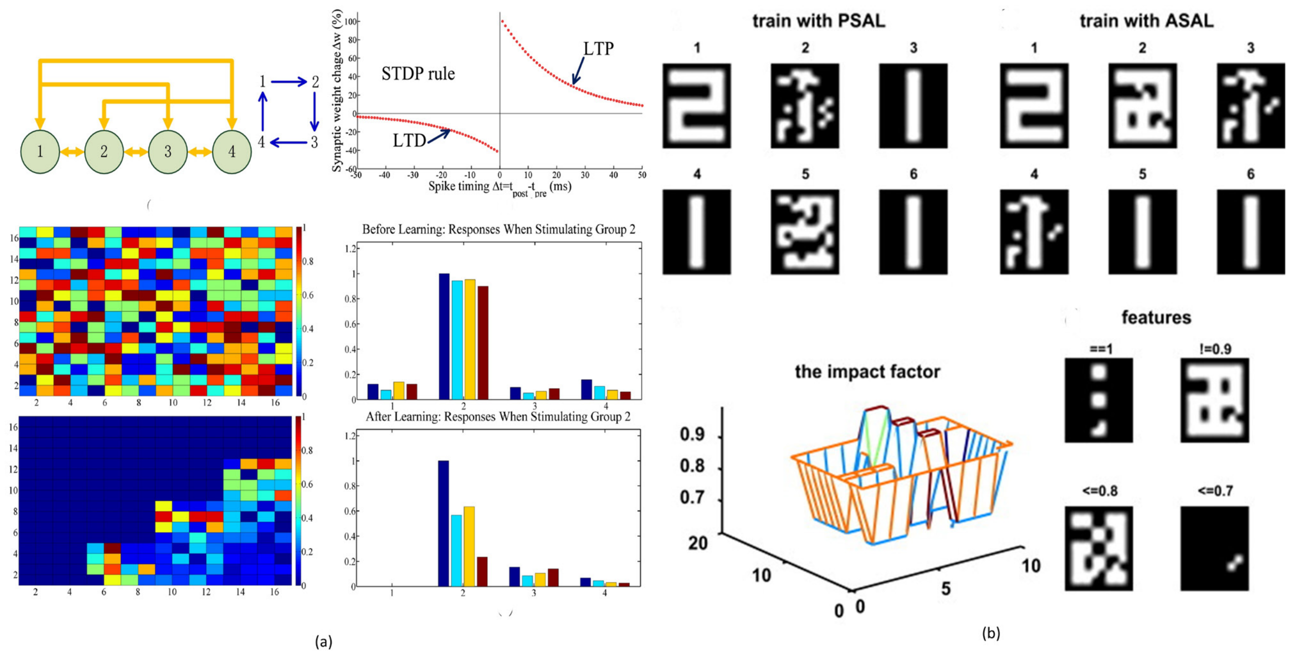


Figure 4. a. Energy comparison of different modulation schemes in neural network built by single memristor-based synapses for sequence learning.Adapted with permission[28]. Copyright 2020, Elsevier. **b.** The simulation of the binary number image. Adapted from Ref. [29]

new nonvolatile devices such as resistive random access memories (RRAMs), phase change random access memories (PCRAMs), and ferroelectric random access memories (FeRAMs). CIM integrates memory and processing functions within the same module, effectively overcoming the challenges that traditional AI chips face, leading to enhanced efficiency and addressing the limitations of conventional architectures. This innovative approach, combined with algorithmic advancements inspired by the biological brain, has enabled neural networks to extract abstract features from vast amounts of data through deep hierarchical nonlinear processing. Neural networks are particularly effective for AI tasks requiring substantial computational power, such as self-driving vehicles and robotics, as they help reduce energy consumption. Neuromorphic chips, which integrate both hardware and algorithm design, offer a remarkable combination of low energy requirements and exceptional parallel processing capabilities, making them ideal for neural computing and intelligent learning. These chips emulate human brain functionalities, executing multiple operations simultaneously and processing complex tasks efficiently. As AI progresses and Moore's Law faces constraints, there is a growing need for alternative technologies that can provide enhanced capabilities beyond traditional CMOS technology.

Memristors are becoming key players in brain-inspired computing for a couple of big reasons. First, they work a lot like the connections between neurons in our brains, showing behaviors similar to synaptic plasticity, which is how neurons strengthen or weaken their connections based on activity. Some even think this brain-like behavior can help explain how learning happens in nature[28-29]. Second, memristors can be used to make energy-saving circuits, and many studies have looked into how they can reduce power use in brain-like systems[28]. Memristors are often used to mimic the connections between neurons, supporting learning processes similar to those in our brains, including both general learning and more specific kinds like spike-timing-dependent plasticity[29].

4. CONCLUSION AND OUTLOOK

This paper provides a comprehensive review of historical developments in neuromorphic computing. Over time, the motivations behind the creation of neuromorphic computers have evolved. The main reasons for building these computers are to have low power consumption, handle many tasks at once, work in real time, and potentially learn and adapt. The paper explores various machine learning models utilized within memristors, acknowledging that a single perfect model might not exist due to each model's unique advantages and disadvantages. Therefore, the future of neuromorphic computing might encompass various models, from simple to complex ones mimicking biological brains. Training algorithms for these computers are also discussed, highlighting the need for new approaches designed specifically for neuromorphic systems rather than adapting existing ones. This area holds great promise for innovation. The aim of our paper was to offer readers an overview of research within neuromorphic computing particularly with memristors, hoping to inspire further innovative contributions to the field and encourage the adoption of neuromorphic computers in various applications. 7]. Recently, advancements in materials science and nanotechnology have facilitated the development of novel sensing technologies, such as Multi-Walled Carbon Nanotubes (MWNTs) and gold nanoparticles. These sensing front-ends exhibit improved sensitivity characteristics, making them suitable for integration with spike-based processing[25, 27]. The increasing application of carbon nanotubes in electronic sensing is primarily due to their superior electrical conductivity compared to carbon black. Research indicates that

REFRENCES

1. Luo, T., Liu, S., Li, L., Wang, Y., Zhang, S., Chen, T., ... & Chen, Y. (2016). DaDianNao: A neural network supercomputer. IEEE Transactions on Computers, 66(1), 73-88.
2. Reagen, B., Whatmough, P., Adolf, R., Rama, S., Lee, H., Lee, S. K., ... & Brooks, D. (2016). Minerva: Enabling low-power, highly-accurate deep neural network accelerators. ACM SIGARCH Computer Architecture News, 44(3), 267-278.
3. Itoh, M., & Chua, L. O. (2011). Memristor hamiltonian circuits. International Journal of Bifurcation and Chaos, 21(09), 2395-2425.
4. Shafiee, A., Nag, A., Muralimanohar, N., Balasubramonian, R., Strachan, J. P., Hu, M., ... & Srikumar, V. (2016). ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. ACM SIGARCH Computer Architecture News, 44(3), 14-26.
5. Backus, J. (1978). Can programming be liberated from the von Neumann style? A functional style and its algebra of programs. Communications of the ACM, 21(8), 613-641.
6. Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.R., Dahl, G., Ramabhadran, B.: Deep convolutional neural networks for large-scale speech tasks. Neural Netw. 64, 39–48 (2015)
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.
8. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.
9. Yao, P., Wu, H., Gao, B., Eryilmaz, S. B., Huang, X., Zhang, W., ... & Qian, H. (2017). Face classification using electronic synapses. Nature communications, 8(1), 15199.
10. Xia, Q., & Yang, J. J. (2019). Memristive crossbar arrays for brain-inspired computing. Nature materials, 18(4), 309-323.
11. Zhang, T., Yang, K., Xu, X., Cai, Y., Yang, Y., & Huang, R. (2019). Memristive devices and networks for brain inspired computing. physica status solidi (RRL)–Rapid Research Letters, 13(8), 1900029.
12. Hu, M., Graves, C. E., Li, C., Li, Y., Ge, N., Montgomery, E., ... & Strachan, J. P. (2018). Memristor based analog computation and neural network classification with a dot product engine. Advanced Materials, 30(9), 1705914.
13. Kim, H., Mahmoodi, M. R., Nili, H., & Strukov, D. B. (2021). 4K-memristor analog-grade passive crossbar circuit. Nature communications, 12(1), 5198.
14. Yao, P., Wu, H., Gao, B., Tang, J., Zhang, Q., Zhang, W., ... & Qian, H. (2020). Fully hardware-implemented memristor convolutional neural network. Nature, 577(7792), 641-646.
15. Yin, S., Sun, X., Yu, S., & Seo, J. S. (2020). High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90-nm CMOS. IEEE Transactions on Electron Devices, 67(10), 4185-4192.
16. Yu, S., Li, Z., Chen, P. Y., Wu, H., Gao, B., Wang, D., ... & Qian, H. (2016, December). Binary neural network with 16 Mb RRAM macro chip for classification and online training. In 2016 IEEE International Electron Devices Meeting (IEDM) (pp. 16-2). IEEE.
17. Xue, C. X., Chiu, Y. C., Liu, T. W., Huang, T. Y., Liu, J. S., Chang, T. W., ... & Chang, M. F. (2021). A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices. Nature Electronics, 4(1), 81-90.
18. Yao, P., Wu, H., Gao, B., Tang, J., Zhang, Q., Zhang, W., ... & Qian, H. (2020). Fully hardware-implemented memristor convolutional neural network. Nature, 577(7792), 641-646.
19. Yao, P., Wu, H., Gao, B., Eryilmaz, S. B., Huang, X., Zhang, W., ... & Qian, H. (2017). Face classification using electronic synapses. Nature communications, 8(1), 15199.
20. Wang, Z., Li, C., Song, W., Rao, M., Belkin, D., Li, Y., ... & Yang, J. J. (2019). Reinforcement learning with analogue memristor arrays. Nature electronics, 2(3), 115-124.
21. Li, C., Wang, Z., Rao, M., Belkin, D., Song, W., Jiang, H., ... & Xia, Q. (2019). Long short-term memory networks in memristor crossbar arrays. Nature Machine Intelligence, 1(1), 49-57.
22. Li, C., Belkin, D., Li, Y., Yan, P., Hu, M., Ge, N., ... & Xia, Q. (2018). Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. Nature communications, 9(1), 2385.
23. Yu, S. (2014, June). Orientation classification by a winner-take-all network with oxide RRAM based synaptic devices. In 2014 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1058-1061). IEEE.
24. Jiang, Y., Kang, J., & Wang, X. (2017). RRAM-based parallel computing architecture using k-nearest neighbor classification for pattern recognition. Scientific reports, 7(1), 45233.
25. Choi, S., Shin, J. H., Lee, J., Sheridan, P., & Lu, W. D. (2017). Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. Nano letters, 17(5), 3113-3118.
26. Fu, J., Liao, Z., Liu, J., Smith, S. C., & Wang, J. (2020). Memristor-based variation-enabled differentially private learning systems for edge computing in iot. IEEE Internet of Things Journal, 8(12), 9672-9682.
27. Jebali, F., Majumdar, A., Turck, C., Harabi, K. E., Faye, M. C., Muhr, E., ... & Portal, J. M. (2024). Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell. Nature Communications, 15(1), 741.
28. Deng, L., Wang, D., Zhang, Z., Tang, P., Li, G., & Pei, J. (2016). Energy consumption analysis for various memristive networks under different learning strategies. Physics Letters A, 380(7-8), 903-909.
29. Chen, L., Li, C., Huang, T., & Wang, X. (2014). Quick noise-tolerant learning in a multi-layer memristive neural network. Neurocomputing, 129, 122-126.

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENTS

The authors are grateful for the funding support from Texas A&M University