WEP
Warwick
Evans
Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Recent Progress of Memristor-based Neuromorphic Computing

Jun Zhou *

Temple University, Philadelphia, U.S.A

* Corresponding Author Email: tun40202@temple.edu

Abstract. The evolution of memristors and their successful applications have positioned them as formidable candidates for the next generation of computer systems. With the rapid advancement of foundational ar- tificial intelligence applications, there is an increasing demand for computational power, energy efficiency, and stability. Memristors and the Neuromorphic Computing (NMC) systems they underpin hold signifi- can't potential to break through the von Neumann bottleneck. However, technical challenges remain in the application of NMC to computer systems. In this review, we focus on the performance of various structured memristors within Neuromorphic Computing and across different machine learning algorithms. We pro- vide an overview of the current challenges faced by NMC, including the structural limitations due to sneak paths and the inherent power consumption limitations, and offer a perspective on future developments and opportunities in the field.

Keywords: Warwick Evans; Publishing; Memristo; Neuromorphic Computing; Artificial Intelligence.

1. Introduction

Over the past several decades, CMOS (Complementary Metal-Oxide-Semiconductor) technology has become a cornerstone of the modern electronics industry. Itslow power consumption, high integration levels, and excep- tional cost-effectiveness have established CMOS as the preferred choice for the design of large-scale integrated circuits. In fields ranging from microprocessors and storage devices to image sensors, the achievements of CMOS technology are unparalleled. For instance, in the realm of storage technology, CMOS has enabled the rapid de- elopement of Dynamic Random Access Memory (DRAM) and non-volatile flash memory, now fundamental components of all computing devices. In the domain of image sensing, the integration of CMOS sensors has revolutionized digital imaging devices, making them more compact, energy-efficient, and cost-effective.

However, as semiconductor device sizes continue to shrink, they are approaching physical limits. In the post- Moore's Law era, the development bottleneck of von Neumann architecture may soon be inadequate to meet humanity's growing demands for computation and the processing of information, such as data and images.

The emergence of the memristor could address the aforementioned challenges. In 2008, a team at HP Labs discovered experimental evidence for the memristor predicted by Chua, based on an analysis of a thin film of titanium dioxide (TiO2), thereby linking the operation of RRAM devices to the memristor concept.

As a two-terminal device, the memristor is particularly advantageous for high-density integration. Capable of emulating the plasticity of biological synapses, it is ideally suited for executing machine learning algorithms, especially the process of weight adjustment in deep learning algorithms. [7] This makes memristor-based NMC systems potentially more efficient than traditional CMOS processors in handling complex pattern recognition and data processing tasks.

The development of memristors has facilitated mature applications of NMC in the fields of Artificial Neural Networks (ANN), Spiking Neural Networks (SNN), Convolutional Neural

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

Networks (CNN), and beyond—from the initial recognition of handwritten digits and images to more complex and computationally demanding video processing tasks.

In this review, we summarize the recent advancements in memristor-based Neuromorphic Computing (MR- based NMC), beginning with an evaluation of the performance and characteristics of various memristor struc- tures, including those made from materials such as TiOx, Hfox, TaOx, Ag, and Cu. We then explore their applications in different types of machine learning algorithms, primarily Artificial Neural Networks (ANN), Reservoir Computing, and Convolutional Neural Networks (CNN). The discussion culminates by addressing the challenges such as structural defects, sneak paths, and power consumption issues faced by MR-based NMC. We conclude with a prospective outlook on the potential for more efficient applications of memristors in future artificial intelligence models.

2. Sophisticated MR device for NMC

TiOx-based memristors: TiOx-based memristors, due to their excellent performance in resistive switching and synaptic emulation, stand at the forefront of NMC devices. The ease of fabrication and compatibility with CMOS technology makes them particularly appealing for integrated neuromorphic systems. Their dynamic conductance change, mim- icking synaptic plasticity, enables learning and memory functionalities essential for AI applications. Advances in layering, doping, and interface engineering have further enhanced their operational stability and scalability. kim2021, it shares a structural heritage with the earliest memristors discovered at HP Labs, have matured into a variety of sophisticated applications. In a significant Nature publication (1), a research team implemented a foundational artificial neural network within a dense crossbar array using transistor-free TiOx. The network, a single-layer perceptron, was capable of classifying simple 3x3 pixel binary images. The memristors in the study emulated biological synapses by executing vector-by-matrix multiplication, a crucial operation in neuromorphic computing. Despite the challenges of variability in the manufacturing process, by controlling and reducing this variability, the team demonstrated the system’s effective learning and classification capabilities. This develop- ment lays the groundwork for more complex neuromorphic systems, potentially surpassing the capabilities of traditional silicon-based circuitsin AI tasks.

Additionally, TiOx-based memristor technology has facilitated the real-time recognition of finger move- ments using skin-conformable electronic devices, especially in developing Human-Machine Interfaces (HMI) that integrate wearable sensors with neuromorphic computing. A system integrating ultra-thin titanium dioxide artificial synapse arrays with organic motion sensors was showcased, comfortably fitting on a finger and pre- cisely and promptly recognizing and learning finger movements. The integration of well-defined synaptic and photo-responsive electrical properties in the devices and sensors has successfully translated photonic signals into post-synaptic currents. This system could identify finger movement patterns for digits 0-9 with an accuracy rate up to 95 percent, maintaining high accuracy under various stress conditions and after up to 100 stress cycles.[21]

HfOx/TaOx-based memristors: This research article presents the key advantages and innovations of using HfOx memristor crossbars for analog signal processing and edge computing tasks. It highlights the 1T1R architecture, which allows precise control over the conductance of each memristor through a transistor integrated into each unit, significantly enhancing device responsiveness and configurability. The HfOx memristor arrays demonstrate high precision and energy efficiency in performing large-scale analog vector-matrix multiplications (VMM), effectively supporting appli- cations such as signal spectrum analysis, image compression, and convolutional filtering for IoT and edge computing. [2] Additionally, this research proposes a new method for handling negative values in discrete cosine transformations (DCT), further enhancing system error tolerance and stability, showcasing their potential in modern high-efficiency computing scenarios.[6]

Ag/Cu based memristors: The Ag/Cu alloyed memristors achieved more stable and controllable device operations compared to Ag-Ni, Ag-Ti, and Ag-Cr structures, demonstrating superior data retention capabilities. In the context of large-scale crossbar arrays, this configuration exhibited outstanding data retention capabilities. [15]

In image processing tasks, the Ag/Cu structure displayed numerous advantages. In this research, it shows the memristors' uniform switching behavior and stable data retention across a wide range of conductance levels are crucial for consistency in processing complex image data. This structure managed tasks such as image sharpening, blurring, and edge detection without the need for a transistor in each memristor unit, avoiding this limitation. [4] In the 32x32 transistor-less alloyed memristor crossbar array, a high device yield was maintained. The precision in programming analogue states suggests high reliability in practical applications.

Overall, TiOx-based memristors are primarily employed for emulating synaptic functions and are partic- ularly adept at facilitating learning and memory capabilities, making them highly attractive for constructing neuromorphic computing systems. The 1T1R architecture of HfOx/TaOx-based memristors offers enhanced precise control, suitable for large-scale analog vector-matrix multiplications, commonly used in edge computing and signal processing scenarios. Ag/Cu-based memristors stand out for their superior data retention capabili- ties and the ability to handle image processing tasks without a transistor at each unit, making them especially suitable for tasks such as image sharpening, blurring, and edge detection. These three distinct structures also offer possibilities through technical complementarity; for instance, while TiOx-based memristors excel in sim- ulating biological synapses, they may face challenges in data retention and device consistency, which can be compensated by the robust data retention and device stability of Ag/Cu-based memristors. In a composite neural network, an exploration into integrating these three types of memristors into a multifunctional device or system is feasible, with each type of memristor optimized for its area of strength. For example, utilizing TiOx- based memristors for rapid learning and adaptation, employing HfOx/TaOx-based memristors for efficient data processing and computation, and using Ag/Cu-based memristors for long-term data storage and complex image processing tasks.

3. ML algorithm achieved by memristor based NMC

The previous section outlined the basic applications of different memristor structures in neural networks. Due to their ability to emulate synaptic properties, alongside their capabilities for parallel processing, non-volatile storage, and compact integration, memristors can leverage these unique attributes for efficient, low-power, and fast-responsive computing within the realm of machine learning, particularly in simulating and optimizing neural network environments.

ANN: Artificial Neural Networks (ANNs) are computational models that mimic the structure and function of the human brain and are widely used in the fields of machine learning and deep learning. They excel particularly in tasks such as image recognition, language processing, and predictive analysis. The fundamental building units of ANNs are neurons, which are interconnected by synapses to form a complex network structure. Memristors, with their adjustable resistance properties, can emulate the functionality of biological synapses, making them well-suited as the neuronal units in ANNs. They can directly adjust weights between layers on the chip, reducing the need for data transmission between processors and memory, thereby enhancing computational speed and efficiency, and reducing power consumption. Additionally, the parallel processing capabilities of memristor arrays make them more efficient in handling complex multilayer neural networks, opening up new possibilities for the development of more advanced artificial intelligence applications.

Memristors have seen substantial mature research and successful applications in the field of Artificial Neural Networks (ANNs), particularly in areas such as image classification and video processing. As one of the earliest studies published in Nature, in 2014, a research group combined complementary metal-oxide-semiconductors (CMOS) with memristors and successfully built a network capable of recognizing 3x3 pixel black-and-white images using the delta rule algorithm.[23] The network was able to perfectly categorize these images into three classes (representing letters), marking an important step towards more complex neuromorphic networks with memristive technologies.

Another team in 2018 utilized multilayer memristor networks to build larger neural network arrays (1024x512), [10] achieving an accuracy of 97 percent on the MNIST dataset of handwritten digits. Compared to CMOS tech- nology, this structure offers potential advantages in speed and power consumption.

Reservoir: Reservoir Computing (RC) is a simplified neural network model primarily employed for tasks such as time series forecasting and pattern recognition. This model is based on a randomly generated dynamic system known as the "reservoir," which processes input data. The learning tasks are executed by the output layer. When imple- mented using memristors, Reservoir Computing leverages the unique physical and electrical characteristics of memristors, particularly their dynamic storage capabilities and adjustable resistance states. These attributes make memristors ideally suited for emulating the reservoir component of RC systems, which constitutes the non- linear and dynamic segments of the network. This synergy of efficient computation and low energy consumption renders the system highly impactful across various future domains.

RC utilizes the dynamic storage capabilities of a short-term memory reservoir to project features from tem- poral inputs into a high-dimensional space. This characteristic allows for the efficient computation of relatively complex tasks with lower training costs. For instance, one research team used a small system comprising only 88 memristors to accomplish the task of handwritten digit recognition, achieving an accuracy of 92.1 percent.[9] Another research team developed a memristor-based Recurrent Neural Network (RNN) for real-time and low- power signal recognition, achieving accuracies of 96.6 percent in real-time heart rate monitoring and 97.9 percent in real-time gesture recognition. [20]

CNN: Convolutional Neural Networks (CNN) have become one of the most important deep learning models, playing a crucial role in image processing tasks. However, the unique computation process of CNN, which involves extensive convolution operations, is also limited by the traditional von Neumann architecture. This limitation leads to significant energy consumption and data latency due to the data transfer between memory and processing units. Neuromorphic computing supported by memristors, which offers a non-von Neumann system, eliminates the cost of data transmission. Memristor arrays can independently perform parallel in-memory multiply-accumulate calculations (MAC), significantly enhancing computational efficiency and energy efficiency.

In this 2020 study, a five-layer CNN for recognizing the MNIST handwritten digits dataset was successfully constructed using 1024 one-transistor-one-memristor (1T1R) arrays. Achieving an accuracy of 96.19 percent, [5]

the use of three parallel memristive convolution machines reduced the error of the mCNN by threefold. The same team also implemented a three-dimensional circuit composed of eight layers of integrated memristor devices. After recognizing the handwritten digits database, the system processed pixels through a filter group to perform edge detection on moving objects in videos, demonstrating a successful application of memristors in the CNN domain. [20]

4. Challenge of NMC

Although memristors have been proven to be successful and stable in various neural network architectures, they still face several challenges and limitations in becoming the next-generation mainstream computing systems. One of the primary challenges is how to avoid losses and maximize the retention of their advantages when mem- ristors' unique circuit structures perform neural network weight computations. A mature alternative to current computing systems requires a stable structure and low power consumption to handle massive computations, yet memristors still have issues with power consumption. Moreover, due to their unique structural characteristics, the resistance values of memristors cannot be controlled by software or programming, which adds difficulty to further implementation in neural networks. **Sneak path:** To implement and simulate neural networks, each memristor in a memristor array is connected through rows and columns to form cross points. When attempting to modify the weight of anode, i. e., changing the state of a specific memristor, current may flow through other memristors in the array that have lower resistance. This phe- nomenon is known as the "sneak path" problem. Current mainstream solutions involve altering the structure of the memristor array and incorporating additional components to control the flow of current. Common structures include the 1-transistor-1-memristor (1T1R), 1-diode-1-memristor (1D1R), and 1-selector-1-memristor (1S1R) configurations. [3, 8] However, these solutions, while addressing the sneak path issue, also introduce higher power consumption and other deficiencies. For example, the 1D1R structure compromises the bidirectionality advan- tage of memristors; the 1T1R structure transforms the system into a three-terminal structure, losing the original two-dimensional configuration; and the 1S1R structure is challenging to utilize in practical applications. These methods, while providing a resolution for the sneak path problem, also bring new challenges and technical obstacles.

Power issue and precious address: Memristors inherently have low energy consumption when storing and processing data, yet the overall power consumption of the system does not solely depend on the memristor units themselves. Due to previously dis- cussed issues such as the "sneak path" problem and non-ideal current pathways in memristor arrays, peripheral circuits must precisely control the voltage and current at each memristor node to ensure correct read and write operations. This requirement increases the complexity of the circuit design and the overall power consumption. For example, in systems with a 1T1R structure, each memristor requires a transistor, which adds additional energy consumption. The cumulative energy used infrequent switching states also impacts the system's overall efficiency.

Another challenge for memristor arrays is precise addressing, which ensures that each read and write op- eration affects only the targeted memristor. Due to the physical and electrical characteristics of memristors, their resistance cannot be precisely controlled, leading to computational errors. Although structural changes such as the 1D1R and 1T1R configurations have been proposed to resolve current deficiencies in memristors, these modifications introduce other shortcomings. Finding solutions that address existing issues without intro- ducing new problems, while maintaining the low power characteristics of memristors, is a significant challenge for memristor technology.

5. Conclusion

To overcome the limitations of traditional computing systems based on the von Neumann architecture, significant progress has been made in neuromorphic computing (NMC) utilizing memristors, yet many challenges remain. As technology advances and our understanding of memristor properties deepens, we anticipate that developing memristors as independent computing units will be key to achieving efficient, scalable, and low-power computing systems. To achieve this goal, potential solutions include optimizing the materials and design of memristors, integrating independent memristors, and developing computing architectures and algorithms specifically for memristors. As we advance our understanding of memristor properties, it is crucial to explore their potential in low-power computing models. The findings of Fuetal. [1] offer promising insights into the use of memristors under bio-voltage conditions, which could revolutionize energy efficiency in computing systems. Additionally, continuing to improve the combination of memristors and transistors to reduce power consumption and enhance operability will also be an effective way to address these challenges. [13, 16].

The article does not mention the application of memristors in other machine learning models; however, this does not imply that they lack potential in such models. Should the current challenges and limitations facing memristors be resolved, they hold promise for demonstrating unique advantages in complex architectures such as Transformers and Knowledge Attention Networks (KAN). For instance, in Transformer models, the parallel processing capabilities and low power consumption of memristors could significantly enhance processing efficiency and energy savings, which is crucial for handling the extensive matrix operations required by attention mechanisms. Similarly, in KAN, the non-volatile storage capability and compact integration of memristors could effectively support long-term memory and rapid access to external knowledge bases, optimizing the model's learning and response speeds. Nevertheless, to deploy memristors in these advanced applications, overcoming the technical barriers associated with their precise control and large-scale integration remains necessary. This will require ongoing research efforts to ensure the practical viability and effectiveness of these technologies.

Year	Device	Algorithm	Architecture	Application	Performance/Accuracy Rate
2015	Metal-oxide memristors	Single-layer perceptron (ANN)	1T1R	Image processing	Perfect classification of 3x3-pixel images
2018	Hafnium Oxide-based Memristors	In situ learning in multi-layer neural network	1T1R (One Transistor One Memristor) Array	Image Processing	91.71% accuracy on MNIST dataset
2017	Dynamic Memristors	Reservoir Computing	Memristor Array	Handwritten Digit Recognition, Temporal Information Processing	88.1% accuracy on handwritten digit recognition
2022	Pd/TaOx/Ta Memristors	Self-Organizing Map (SOM)	128x64 1T1R Memristor Array	Data Clustering, Image Processing, Solving Optimization Problems	High energy efficiency and competitive accuracy in tasks
2020	TiN/TaOx/HfOx/TiN Memristors	Hybrid Training for Memristor CNN	1T1R	Handwritten Digit Recognition, Temporal Information Processing	95.63% accuracy post-weight transfer; 97.99% after training
2018	Diffusive Memristors	Unsupervised Learning	Fully Integrated 1T1R Memristive NN	Pattern Classification	High accuracy in pattern classification
2022	Dynamic Memristors Arrays	Reservoir Computing	Fully Analogue Reservoir Computing System	Arrhythmia Detection, Dynamic Gesture Recognition	96.6% for arrhythmia detection, 97.9% for dynamic recognition
2023	Ultrathin Titanium Oxide-based Artificial Synapse Array	Spike-Timing Dependent Plasticity (STDP)	Spike-Timing Dependent Plasticity (STDP)	Real-Time Finger Motion Recognition	Up to 95% accuracy in recognizing finger-writing patterns
2020	One-dimensional (1D) transistors on Ag wire	Stochastic Gradient Descent (SGD)	Fiber-based Artificial Multi-Synapses	Wearable Neuromorphic Applications	Approximately 90% for MNIST and 70% for ECG patterns
2019	HfAlyOx/TaOx Memristors	In-memory computing	128x64 1T1R Memristor Array	Image Processing, Deep Learning Models	High efficiency in processing with significant energy and speed
2023	CMOS-integrated Memristors	High-Precision Programming with Denoising Process	256x256 Memristor Array	Machine Learning Acceleration, Edge Computing	Achieved high precision with 2048 conductance levels in memristors
2020	h-BN/TaOx/NbOx Memristors	Analog Neuromorphic Computing	High-Density Crossbar Arrays	Image Recognition	High device yield with over 98% accuracy in mimicking NC tasks

Figure 1. Summary of some memristor research and applications over the years. The information in the table comes from [23, 10, 9, 12, 5, 18, 20, 21, 22, 14, 17, 19], with the citations listed in the same order as the table.

References

[1] Fu, T., Liu, X., Gao, H. et al. Bioinspired bio-voltage memristors. Nat Commun, 11, 1861 (2020).

[2] Jiang, H., Han, L., Lin, P. et al. Sub-10 nm Ta channel responsible for superior performance of a HfO2 memristor. Sci Rep, 6, 28525 (2016).

[3] Midya, R., Wang, Z., Zhang, J., Savel'ev, S.E., Li, C., Rao, M., Jang, M.H., Joshi, S., Jiang, H., Lin, P., Norris, K., Ge, N., Wu, Q., Barnell, M., Li, Z., Xin, H.L., Williams, R.S., Xia, Q., Yang, J.J. Anatomy of Ag/Hafnia-based selectors with 1010 nonlinearity. Adv. Mater., 29, 1604457 (2017).

[4] You, B.K., Kim, J.M., Joe, D.J., Yang, K., Shin, Y., Jung, Y.S., Lee, K.J. Reliable memristive switching memory devices enabled by densely packed silver nanocone arrays as electric-field concentrators. ACS Nano, 10, 9478 - 9488 (2016).

[5] Yao, P., Wu, H., Gao, B. et al. Fully hardware-implemented memristor convolutional neural network. Nature, 577, 641 – 646 (2020).

[6] Li, C., Hu, M., Li, Y. et al. Analogue signal and image processing with large memristor crossbars. Nat Electron, 1, 52 – 59 (2018).

[7] Wang, Z., Joshi, S., Savel'ev, S. et al. Memristors with diffusive dynamics as synaptic emulators for neuro- morphic computing. Nat Mater, 16, 101 – 108 (2017).

[8] Choi, S., Jang, S., Moon, JH. et al. A self-rectifying TaOy/nanoporous TaOx memristor synaptic array for learning and energy-efficient neuromorphic systems. NPG Asia Mater, 10, 1097 – 1106 (2018).

[9] Du, C., Cai, F., Zidan, M.A. et al. Reservoir computing using dynamic memristors for temporal information processing. Nat Commun, 8, 2204 (2017).

[10] Li, C., Belkin, D., Li, Y. et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. Nat Commun, 9, 2385 (2018).

[11] Kim, H., Mahmoodi, M.R., Nili, H. et al. 4K-memristor analog-grade passive crossbar circuit. Nat Commun, 12, 5198 (2021).

[12] Wang, R., Shi, T., Zhang, X. et al. Implementing in-situ self-organizing maps with memristor crossbar arrays for data mining and optimization. Nat Commun, 13, 2289 (2022).

[13] Yan, M., Huang, C., Bienstman, P. et al. Emerging opportunities and challenges for the future of reservoir computing. Nat Commun, 15, 2056 (2024).

[14] Xia,Q., Yang, J.J. Memristive crossbar arrays for brain-inspired computing. Nat Mater, 18, 309–323 (2019).

[15] Yeon, H., Lin, P., Choi, C. et al. Alloying conducting channels for reliable neuromorphic computing. Nat Nanotechnol, 15, 574 – 579 (2020).

[16] Wang, Z., Wu, H., Burr, G.W. et al. Resistive switching materials for information processing. Nat Rev Mater, 5, 173 – 195 (2020).

[17] Rao, M., Tang, H., Wu, J. et al. Thousands of conductance levels in memristors integrated on CMOS. Nature, 615, 823 – 829 (2023).

[18] Wang, Z., Joshi, S., Savel'ev, S. et al. Fully memristive neural networks for pattern classification with unsupervised learning. Nat Electron, 1, 137 – 145 (2018).

[19] Chen, S., Mahmoodi, M.R., Shi, Y. et al. Wafer-scale integration of two-dimensional materials in high- density memristive crossbar arrays for artificial neural networks. Nat Electron, 3, 638 – 645 (2020).

[20] Zhong, Y., Tang, J., Li, X. et al. A memristor-based analogue reservoir computing system for real-time and power-efficient signal processing. Nat Electron, 5, 672 – 681 (2022).

[21] Cho, H., Lee, I., Jang, J. et al. Real-time finger motion recognition using skin-conformable electronics. Nat Electron, 6, 619 – 629 (2023).

[22] Ham, S. et al. One-dimensional organic artificial multi-synapses enabling electronic textile neural network for wearable neuromorphic applications. Sci Adv, 6, eaba1178 (2020).

[23] Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., Adam, G. C., Likharev, K. K., & Strukov, D. B. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature, 521, 61 - 64 (2015).