Development of Crosspoint Memory Arrays for Neuromorphic Computing

Saverio Ricci, Piergiulio Mannocci, Matteo Farronato, Alessandro Milozzi, and Daniele Ielmini

Abstract Memristor-based hardware accelerators play a crucial role in achieving energy-efficient big data processing and artificial intelligence, overcoming the limitations of traditional von Neumann architectures. Resistive-switching memories (RRAMs) combine a simple two-terminal structure with the possibility of tuning the device conductance. This Chapter revolves around the topic of emerging memristor-related technologies, starting from their fabrication, through the characterization of single devices up to the development of proof-of-concept experiments in the field of in-memory computing, hardware accelerators, and brain-inspired architecture. Non-volatile devices are optimized for large-size crossbars where the devices' conductance encodes mathematical coefficients of matrices. By exploiting Kirchhoff's and Ohm's law the matrix–vector-multiplication between the conductance matrix and a voltage vector is computed in one step. Eigenvalues/eigenvectors are experimentally calculated according to the power-iteration algorithm, with a fast convergence within about 10 iterations to the correct solution and Principal Component Analysis of the Wine and Iris datasets, showing up to 98% accuracy comparable to a floating-point implementation. Volatile memories instead present a spontaneous change of device conductance with a unique similarity to biological neuron behavior. This characteristic is exploited to demonstrate a simple fully-memristive architecture of five volatile RRAMs able to learn, store, and distinguish up to 10 different items with a memory capability of a few seconds. The architecture is thus tested in terms of robustness under many experimental conditions and it is compared with the real brain, disclosing interesting mechanisms which resemble the biological brain.

S. Ricci (✉) · P. Mannocci · M. Farronato · A. Milozzi · D. Ielmini
Dipartimento Di Elettronica, Informazione E Bioingegneria (DEIB), Politecnico Di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy
e-mail: saverio.ricci@polimi.it

© The Author(s) 2024

F. Amigoni (ed.), *Special Topics in Information Technology*,
PoliMI SpringerBriefs, https://doi.org/10.1007/978-3-031-51500-2_6

1 Introduction

With the advent of the Internet-Of-Things and with the ever-growing number of people gaining the possibility to purchase smartphones and tablets capable to store a large amount of photo, video, music and applications in a single portable device, the global amount of data has increased exponentially, which raises strong requirements in terms of energy efficiency and processing speed for data analysis [1–3]. To satisfy these requirements, the computing performance of modern computers has increased steadily in the past few decades thanks to the scaling down of the transistor dimensions and the consequent higher density of information being stored in the same area, as predicted by Moore's law. The downscaling is now approaching its natural end mainly due to the increasing leakage of the complementary metal–oxide–semiconductor (CMOS) transistors due to their extreme miniaturization [2]. The operating frequency of each transistor has already reached an upper limit set by the maximum acceptable power dissipation, preventing further speed improvement at the device level to avoid an excessive temperature increase of the chip.

If on one side we have reached a limit on data transport speed due to the transistors, on the other side we have to consider that there is an additional limit imposed by the fact that conventional computing systems are based on the von Neumann architecture [4, 5], where memory and processing units are physically separated, which leads to an additional inevitable bottleneck due to the necessary data movement between the two separated units, which causes significant latency and energy consumption. This latency becomes significant when operation must be repeated thousands or millions of times, as it happens to tensor products and matrix multiplications, where the operation between the elements of the matrices cannot be done in parallel but only one operation after the other, finally collecting all the results.

Alternative in-memory computing approaches are becoming increasingly attractive to develop novel logics and neuromorphic computations to overcome Von Neumann bottleneck issues [4–6]. Indeed, typical operations like image learning, pattern recognition and decision exhibit high computational cost for boolean CMOS processors, while, for human brain, they represent elementary processes. In this scenario, the development of new devices designed specifically for neuromorphic computing could enable high density and low power networks to properly operate learning and recognition tasks. Among the various emerging memories, also known as memristors, resistive switching memories appear as one of the most promising technologies for in-memory computing, thanks to the CMOS-compatible fabrication process, the small area and the analog programming.

Differently from conventional memories based on transistors, which are able to store binary values only, specifically "1" (transistor in pass mode) and "0" (transistor switched off), memristors can store information in their electrical properties, like the resistance (or conductance) for example, in an analog way. Moreover, by organizing these memories in a matrix configuration, also known as crosspoint architecture, the matrix–vector multiplication is performed in one step only, carrying out all the single elements multiplications simultaneously exploiting the Kirchhoff's law [5, 6, 8].

Because of the novelty of this technology, problems of reliability and integration with existing technologies affect the emerging memories and further studies are required to overcome the limits by optimizing the materials and their responses and developing architecture designs and algorithms to exploit the innovative features and the strong parallelism of the physical multiplication [4, 5]. In this scenario, this Chapter focuses on the topic of RRAMs for high-density crosspoint arrays, starting from their fabrication, through the characterization of single devices up to the development of proof-of-concept experiments in the field of in-memory computing, hardware accelerators and brain-inspired architecture.

2 Non-volatile RRAMs

A resistive switching memory is a two-terminal device where the conductance can be manipulated by externally applied voltage pulses. The main structure is composed by an oxide layer sandwiched between two metals, in the so-called Metal–Insulator-Metal.

(MIM) structure. The RRAM switching mechanism refers to the possibility of creating and disrupting a conductive path across the oxide, creating a conductive bridge between the metals, by locally changing the oxygen vacancy concentration and for this reason they are also known as RedOx RRAMs (ReRAM). By applying a positive voltage to the top electrode (TE), the oxygen vacancies can migrate and reallocate inside the oxide layer with a consequent change of the electrical properties, where the formed oxygen vacancy-based conductive channel dictates a low resistance state (LRS), as depicted in Fig. 1a. The application of a negative voltage to the TE, instead, induces a vacancy dispersion into the oxide, the conductive path is dissolved and the resistivity rises-up, bringing the device in a high resistive state (HRS).

The typical electrical response of a RRAM is reported in Fig. 1b, where the hysteresis of the I-V curves changes according to the maximum current [1, 3, 6, 7], called compliance current (I_C). The dependence of the conductance as a function of the I_C is clearly visible in Fig. 1c, with a linear dependence linked to the possibility of enlarging the conductive channel diameter by increasing the current [3, 7]. Inversely, with the increase of the reset amplitude the conductive state is brought back to the HRS and the larger the voltage, the less conductive the device is, as seen in Fig. 1c. The exponential behavior is explained as the presence of an activation energy required to move the vacancies and the defects, resulting in an Arrhenius-like process.

The tunability of the conductance is the key point of the RRAM technology and the advantage is clear when the devices are organized in a matrix configuration, with the TEs and the BEs placed orthogonally. By exploiting Kirchhoff's and Ohm's law the matrix–vector-multiplication between the conductance matrix and a voltage vector is computed in one step only [7–9]. Each element in the matrix must be programmed properly to a desired value, by using multiple set and reset operations, as seen in Fig. 2a, where a device is programmed passing from 0 μS to 82 μS using set operation and then reset till the target of 73 μS. Figure 2b and c report the before

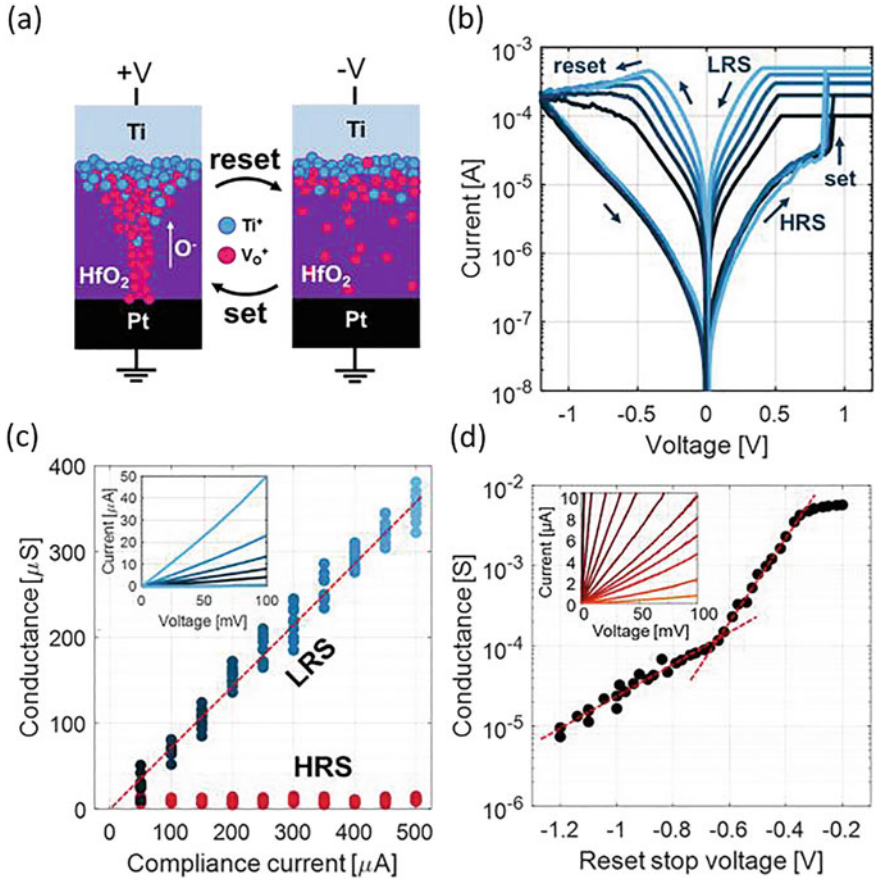

Fig. 1 Physical mechanism and quasi static characterization of Pt/HfO2/Ti non-volatile RRAMs. **a** Sketch of the switching mechanism with the formation and dissolution of the conductive channel. **b** I-V curves at different compliance currents in logarithmic scale. The device passes from the HRS to different LRS states through set transitions and then reset [7]. **c** Conductance levels as a function of I_C. **d** Conductance levels associated with the reset amplitudes. The values spread in a range between 5 mS and 10 μS

and after programming of an 8 × 8 crossbar (visible in Fig. 2d). The final matrix encodes the coefficients of the covariance matrix of the Wine dataset [9], with an acceptable maximum error of ±3 μS.

The power iteration is an algorithm able to extract the eigenvector components of a matrix by computing vector matrix multiplication between the matrix and the vector obtained in the previous step [8, 10]. After some iterations the values converge to asymptotic values, which are proportional to the mathematical eigenvector (the factor is linked to the one to convert the matrix to a conductance matrix) [10]. Figure 2e sketches the equivalent circuits which implements the power iteration algorithm: the current coming from a first MVM product is converted in voltage, which feed again

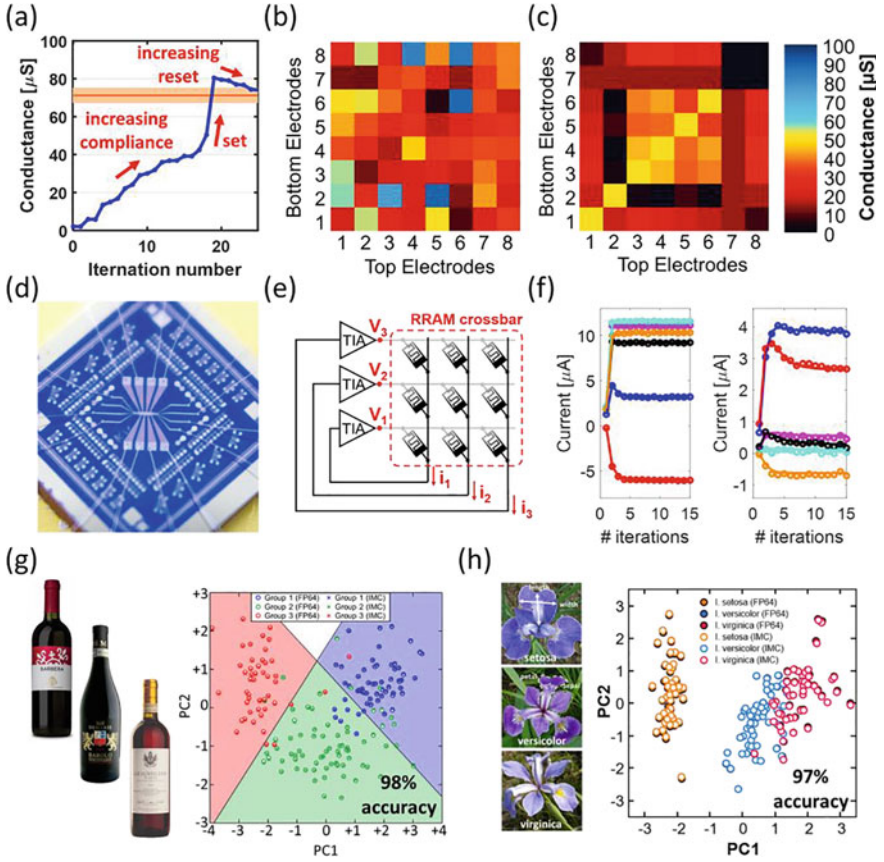

Fig. 2 Program and verify operation for In-Memory Computing and PCA. **a** Tuning of the conductive state using set and reset operations. **b** Initial state of an 8 × 8 crossbar after fabrication. **c** Conductance matrix programmed through the program and verify algorithm. The matrix encodes the Wine dataset covariance matrix [9]. **d** Optical image of an 8 × 8 crossbar bonded. **e** Conceptual circuit to implement the power iteration algorithm. **f** Eigenvector computation for PC1 (on the left) and PC2 (on the right). The curves stand for the evolution of the current values. **g** Wine dataset projection along the first two PCs [9]. **h** Iris dataset projection [7]

the MVM. The extraction of the first and second greater eigenvectors, also called principal components (PC), can be followed in Fig. 2f. Within 10 iterations, the currents converge to the asymptotic values [9]. Finally, the extracted PCs are used to project the dataset along the components and to group the different wines, as seen in Fig. 2g, with an accuracy of 98%, comparable with the floating point 64 software-based computation [9]. The Wine dataset contains 3 different wines classified with 6 properties, like chemical values, color and sugar content. The eigenvectors are proportional to these values and all the wines can be written as a combination of such components. To validate the approach, the same experiment is repeated in Fig. 2h

by looking at the Iris dataset [7], containing three different Iris flowers labelled with petal and sepal length and width. These results support RRAM crosspoint arrays for accelerating advanced machine learning with IMC.

3 Volatile RRAMs

Filamentary switching memories are a different class of RRAMs which rely on a metallic filament to change the electrical properties, where high mobility metal ions migrate from one electrode to the other creating a conductive bridge [11, 12]. Silver-based RRAMs exhibit spontaneous disruption of the metallic conductive filament with a lifetime ranging from few microseconds to several seconds, thus by controlling and predicting the filament lifetime, devices can be engineered for a wide range of applications. When a positive bias is applied to the Ag electrode, the electric field leads the Ag ions to migrate across the oxide and the resistivity drops down, creating a conductive path made of nanoclusters [11]. Reducing the voltage, the filament spontaneously disrupts, the resistivity rises-up and a gap occurs, which is responsible for the absence of conductance. Figure 3a reports the electrical response associated to the mechanisms described.

Because of the spontaneous disruption of the filament [11–13], it is important to study the temporal evolution of the devices, by switching on the memory and then monitoring the state until it switches off. The time window in which the filament remains stable is called retention time. Figure 3b collects the cumulative distribution curves of the retention time as a function of the maximum current reached during the switching [12], current which is limited by exploiting the saturation region of transistors. The larger is the current and the longer is the average retention time,

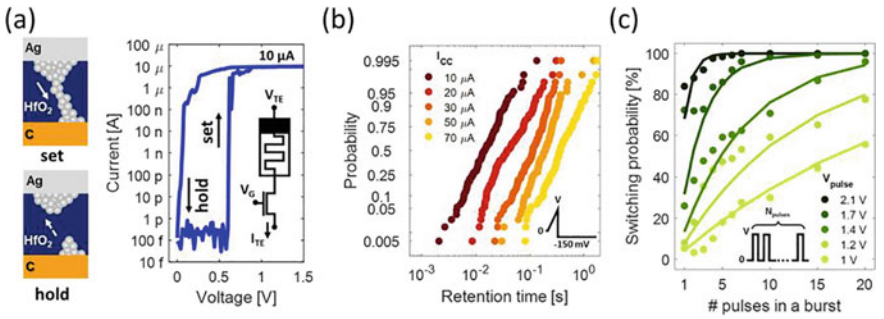

Fig. 3 Ag-based volatile 1T1R RRAM electrical characterization. **a** Quasi-static I–V sweep. **b** Retention time distributions for different maximum current. 3 V and 1 ms triangular pulse are applied to set the device, and a constant −150 mV bias voltage is applied to monitor the status of the device. **c** Impact of the number of pulses applied. Considering a group of pulses, the probability of finding the device in the ON state increases with the number of applied pulses

according to the fact that the filament has a greater diameter and thus it is naturally more robust.

Moreover, the devices result to be sensitive also to the pulse amplitude [12, 13], meaning that for small amplitudes the devices do not switch on while for large values (>3 V for example) the devices always switch on. The trends of the probability that the device switches on after a pulse as a function of the pulse amplitude and the number of pulses [13] are shown in Fig. 3c. By increasing the pulse amplitude (V_{pulse}), the probability increases and it starts saturating at 100% after few pulses (like in the case of 2.1 V, the darker curve). For low voltages (1 V pulses, the lighter green curve) the probability weakly increases with the number of pulses.

The fact that stochastic properties, retention time and switching probability, are tunable with the voltage and the temporal dynamic is adaptable according to the compliance current are explored in a simple neuromorphic circuit. Short-term memory is a primary concept in human life, since it is responsible of the storing of acquired information in the meantime that it is processed and evaluated. The proposed system has two main features: storing the information in the memory and later recognizing it, as depicted in.

Figure 4a: the memory has an item stored inside, for example an advertisement spot (marked with a specific color), which is linked to specific areas that are activated. When the true item arrives (the orange in the example), the system recognizes it, and a trigger signal is generated. When other items arrive, the system does not recognize them and thus it is not triggered. Being a short-term memory, if the system is not refreshed somehow, providing for example the true item, it forgets the stored information.

The circuit is implemented using 5 different devices (in Fig. 4b) where the total current is summed together, and the transistor share the same gate to have similar time responses. At each device is sent a signal which is calibrated to have a specific switching probability (P_{ON}), considering the device-to-device variability. This switching probability can be seen as the volume, thus the higher the volume the higher the relevance we give to the spot. Figure 4c shows the evolution of the system when a pattern is applied multiple times as soon it is impressed in the system (3 devices are switched on) and then random patterns are applied. When the right pattern arrives (marked with a dot) the system is triggered and recognizes it, otherwise no. Different experimental parameters, in terms of pattern rate, delay and amplitude, are tested, finding the best condition when the switching probability is low while the refresh is high (in Fig. 4d). This condition is in good agreement with what happens to the human brain during the advertisement: all the spots have a small relevance, but when the right one is on the tv the attention is high because the spot is recognized, thus we can distinguish what we like from the other spots. Differently, when the spot is less broadcasted (small spike rate, in Fig. 4e) the information is lost, and it is more difficult to recognize it. On the other hand, the volume plays a crucial role, because our attention changes drastically. For great P_{ON} (so large volume) the system easily changes the information stored and thus is not able anymore to recognize the first one.

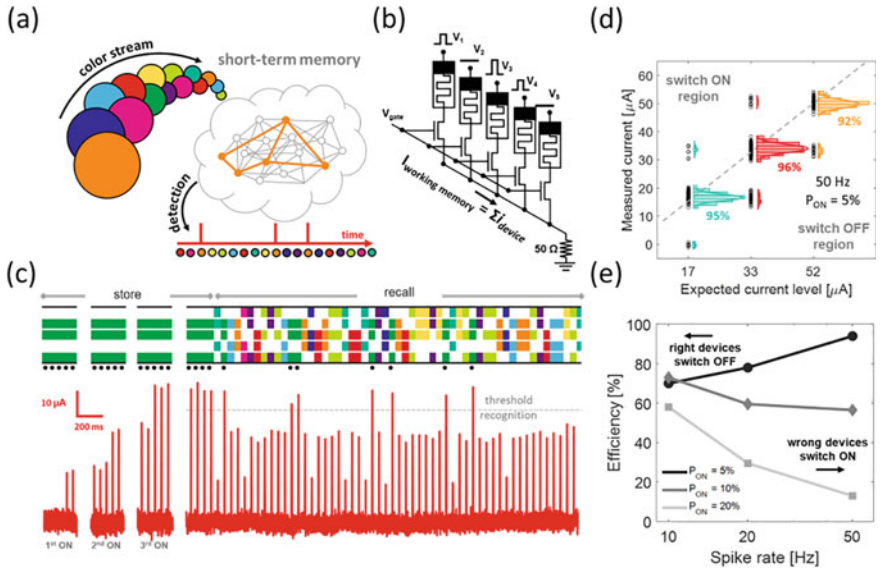

Fig. 4 Sketch of the memristive implementation of a working memory circuit emulator. **a** After an object (a color for example orange) is stored in the mnemonic architecture, a stream of objects is sent to the system. When the true object arrives, the system is refreshed and triggered. **b** Real implementation using 5 Ag-based volatile 1T1R RRAMs. The devices are connected in parallel to sum the currents and share the same gate voltage to have the similar electrical responses. **c** Example of an experimental trace. In the store phase the same pattern is sent multiple times to switch ON the right devices, until all the 3 RRAMs are in the ON state. In the recall phase random pattern are sent. The current is discretized in four levels, according to the number of ON devices. A suitable current threshold is used to discriminate when the true pattern arrives. **d** Correlation plot of the best experimental parameters to check the accuracy of the system. **e** Behavior of the memristive architecture by changing experimental parameters. The best results are achieved when the system is frequently refreshed

4 Conclusions

This Chapter aimed to give an overview on emerging memories and on the potentialities of resistive switching devices in the field of in-memory computing, hardware accelerators and brain-inspired architecture. In RRAMs with Pt/HfO_2/Ti stack the conductance value of the memory can be tuned in an analog way according to external parameters, such as the current flowing through the device. This gives the possibility to directly map mathematical weights into conductance values and, in a suitable crossbar configuration, to operate the MVM operation in one step. The eigenvalue/eigenvector calculation is experimentally demonstrated, and the extracted component is used in the PCA of a large dataset, showing not only a fast convergence but also an accuracy comparable to the FP64 software-based solution, with a value up to 98%. On the other side, by changing one of the electrodes, silver-based devices feature a spontaneous change of device conductance with a retention ranging from 1 ms to

several seconds, as it happens in the biological systems. This similarity is crucial to implement biological functions and tasks, as the short-term-memory typical of living animals. The simple structure combined with a wide flexibility in terms of electrical responses and properties, supports the RRAM technology as interesting candidate for accurate acceleration of machine learning, in-memory computing, and neuromorphic systems.

Acknowledgements The authors would like to thank M. Asa, A. Scaccabarozzi, C. Somaschini, C. Nava, S. Fasoli, S. Bigoni, E. Sogne and G. Cannetti for help in the fabrication process. This work was partially performed in Polifab, the micro and nanofabrication facility of Politecnico di Milano. This article received funding from the European Union's Horizon 2020 research and innovation program (grant agreement no. 824164).

References

1. Xia Q, Yang JJ (2019) Memristive crossbar arrays for brain-inspired computing. Nat Mater 18:309–323
2. Yang JJ, Strukov DB, Stewart DR (2013) Memristive devices for computing. Nat Nanotechnol 8(1):13–24
3. Kim H, Mahmoodi MR, Nili H, Strukov DB (2021) 4K-memristor analog-grade passive crossbar circuit. Nat Commun 12(1):5198
4. Milano G, Pedretti G, Montano K, Ricci S, Hashemkhani S, Boarino L, Ielmini D, Ricciardi C (2022) In materia reservoir computing with a fully memristive architecture based on self-organizing nanowire networks. Nat Mater 21(2):195–202
5. Pedretti G, Mannocci P, Li C, Sun Z, Strachan JP, Ielmini D (2021) Redundancy and analog slicing for precise in-memory machine learning—part II: applications and benchmark. IEEE Trans Electron Devices 68(9):4379–4383
6. Wang R, Shi T, Zhang X, Wei J, Lu J, Zhu J, Wu W, Liu Q, Liu M (2022) Implementing in-situ self-organizing maps with memristor crossbar arrays for data mining and optimization. Nat Commun 13:2289
7. Ricci S, Mannocci P, Farronato M, Hashemkhani S, Ielmini D (2022) Forming-free resistive switching memory crosspoint arrays for in-memory machine learning. Adv Intell Syst 4:2200053
8. Mannocci P, Baroni A, Melacarne E, Zambelli C, Olivo P, Pérez E, Wenger C, Ielmini D (2022) In-memory principal component analysis by crosspoint array of resistive switching memory: a new hardware approach for energy-efficient data analysis in edge computing. IEEE Nanatechnol Mag 16(2):4–13
9. Ricci S, Mannocci P, Farronato M, Ielmini D (2023) In-memory computing with crosspoint resistive memory arrays for machine learning. In: Proceedings of SIE 2022. SIE 2022. Lecture notes in electrical engineering, vol 1005. Springer
10. Jolliffe I (2005) Principal component analysis. In: Everitt BS, Howell DC (eds) Encyclopedia of statistics in behavioral science. Wiley, Chichester, UK, p bsa501
11. Covi E, Wang W, Lin Y, Farronato M, Ambrosi E, Ielmini D (2021) Switching dynamics of Ag-based filamentary volatile resistive switching devices—part I: experimental characterization. IEEE TED 2021, vol 68, No. 8
12. Wang W, Covi E, Milozzi A, Farronato M, Ricci S, Sbandati C, Pedretti G, Ielmini D (2021) Neuromorphic motion detection and orientation selectivity by volatile resistive switching memories. Adv Intell Syst 3:2000224

13. Ricci S, Kappel D, Tetzlaff C, Ielmini D, Covi E (2022) Decision making by a neuromorphic network of volatile resistive switching memories. In: 2022 29th IEEE international conference on electronics, circuits and systems (ICECS), Glasgow, United Kingdom, 2022, pp 1–4

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.