IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 7, 2020, accepted October 20, 2020, date of publication November 5, 2020, date of current version November 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035638

Application of the Quasi-Static Memdiode Model in Cross-Point Arrays for Large Dataset Pattern Recognition

FERNANDO LEONEL AGUIRRE[1,2,3], (Member, IEEE), SEBASTIÁN MATÍAS PAZOS[1,2], (Member, IEEE), FÉLIX PALUMBO[1,2], (Member, IEEE), JORDI SUÑÉ[3], (Fellow, IEEE), AND ENRIQUE MIRANDA[3], (Senior Member, IEEE)

[1]Unidad de Investigación y Desarrollo de las Ingenierías (UIDI), Universidad Tecnológica Nacional (UTN-FRBA), Facultad Regional Buenos Aires, Buenos Aires C1179AAQ, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires C1425FQB, Argentina
[3]Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain

Corresponding author: Fernando Leonel Aguirre (aguirref@ieee.org)

This work was supported in part by the Argentine Ministerio de Ciencia, Tecnología e Innovación (MINCyT) under Contract PICT2013/1210, Contract PICT 2016/0579, and Contract PME 2015-0196; in part by the CONICET under Project PIP-11220130100077CO; and in part by the UTN-FRBA under Project PID-UTN EIUTIBA4395TC3, Project PID-UTN CCUTIBA4764TC, Project PID-UTN MATUNBA4936, Project PID-UTN CCUTNBA5182, and Project PID-UTN CCUTNBA6615. The work of Jordi Suñé and Enrique Miranda was supported by the TEC2017-84321-C4-4-R and WAKeMeUP 783176 projects, co-funded by grants from the Spanish Ministerio de Ciencia e Innovación and the Electronic Components and Systems for European Leadership-European Union (ECSEL-EU) Joint Undertaking.

ABSTRACT We investigate the use and performance of the quasi-static memdiode model (QMM) when incorporated into large cross-point arrays intended for pattern classification tasks. Following Chua's memristive devices theory, the QMM comprises two equations, one equation for the electron transport based on the double-diode circuit with single series resistance and a second equation for the internal memory state of the device based on the so-called logistic hysteron or memory map. Ex-situ trained memdiodes with different MNIST-like databases are used to establish the synaptic weights linking the top and bottom wire networks. The role played by the memdiode electrical parameters, wire resistance and capacitance values, image pixelation, connection schemes, signal-to-noise ratio and device-to-device variability in the classification effectiveness are investigated. The confusion matrix is used to benchmark the system performance metrics. We show that the simplicity, accuracy and robustness of the memdiode model makes it a suitable candidate for the realistic simulation of large-scale neural networks with non-idealities.

INDEX TERMS RRAM, resistive switching, cross-point, memory, memristor, neuromorphic, pattern recognition.

I. INTRODUCTION

Resistive memory (RRAM) or memristor-based cross-point arrays (CPA, see Fig. 1) are nowadays in the spotlight as their properties as programmable non-volatile memory (NVM) devices have enormous potential application in fields such as artificial intelligence (AI) [1], [2] and information storage [3]. Moreover, the newly introduced paradigm of Internet of Everything (IoE) requires the processing of large amounts of data with a very reduced power consumption. In this regard, the Matrix-Vector-Multiplication (MVM) method used by many of these applications is particularly suitable for being performed by a CPA [4]. Additionally, the CPA structure may be scaled down to $4F^2$, with F the feature size of the technology node [5], which enables a large scale integration of memory units. These features allow CPAs to solve a number of computationally intensive specific AI tasks such as the classification of a variety of patterns (sounds, images, electrocardiograms, etc.) with a lower energy consumption than conventional Von Neumann systems [4]. Such applications have been extensively studied in previous works [1], [6]–[11] considering various CPA architectures as well as different memristor models. Hu *et al.* reported in [1] a simulation-based case study of a CPA for character recognition with added noise using two CPAs of 256 × 26 (i.e. 256 rows by 26 columns, totalling ∼13k devices) to represent

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Rahimi Azghadi.

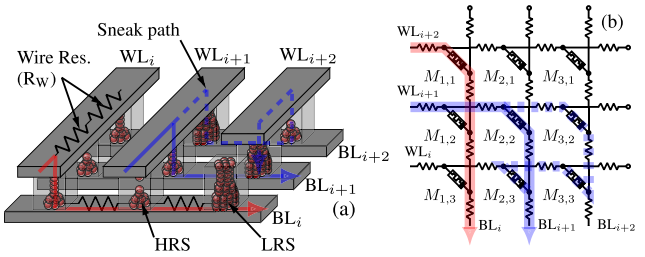


FIGURE 1. (a) Sketch of the CPA structure. Red and blue arrows exemplify the electron flow through the memdiodes connecting the top (Word lines -WL-) and bottom lines (Bit lines -BL-). Different resistance states are schematically represented (High Resistance State -HRS- to Low Resistance State -LRS-). The dashed blue line depicts the so-called sneak path problem. The parasitic wire resistance is indicated for WL_i and BL_i. (b) Equivalent circuit representation of the CPA sketched in (a).

both the positive and negative synaptic weights using a Verilog-A nonlinear memristor model [12]. To reduce both the area and power consumption arising from having two CPAs, Truong *et al.* presented in [6] a CPA architecture (64×26, $\sim$1.6k devices) using the same memristive device model, but with a single memristor array and a constant-bias circuit for representing both the plus- and minus-polarity synaptic weights. This model was also used in [8] for voice recognition using a set of CPAs summing up to $\sim$2.5k memristors.

Nevertheless, beyond this promising capability, CPAs are not exempted from practical limitations such as the line or wire resistances (R_W), the resistance window of the devices (R_{ON} and R_{OFF}), the degraded Signal-to-Noise ratio, the inference latency, the device-to-device variability (D2D) as well as the inherent conducting features of CPAs like the so-called sneak path problem (see Fig. 1). While the formers are mainly a consequence of the increase of R_W as the fabrication technology scales down [10], [13] which in combination with a reduced resistance window or low R_{ON} causes a significant voltage drop across the CPA lines, the latter refers to the non-negligible current flowing through the unselected devices. This is the origin of errors in the read and write processes [13]. Both hardware [14] and software [1], [6]–[11], [13], [15], [16] approaches have been proposed to address these challenges. Although hardware-based techniques include compensation methods that improve the system performance, they are in general time and cost demanding [14]. Software solutions allow a more systematic study and can be split into three groups: first, several authors [13], [15]–[20] opted for solving the system of coupled differential equations which arises from considering the current Kirchoff's law at each junction in the CPA assuming that the programmed memory devices act as resistors of fixed value. Although promising, this approach neither accounts for the CPA control circuitry nor it considers the nonlinear conducting behaviour of memristors, thus limiting their applicability. Second, Python-based approaches and similar [21], [22] allow incorporating realistic RRAM models, but normally ignore the parasitic CPA effects or do not account for the peripheral circuitry. Third and last, SPICE simulation appears as the most suitable approach, as it allows studying the full system (CPA and control electronics) [1], [6]–[11]. However, this approach is constrained to the limitations of the memristor model considered and to the size of the memristor-based CPA given the high computational requirements [23]. As an example, the NVM-SPICE simulator [24] is capable of simulating CPAs of up to 32×32 devices within reduced times, but it seems unable to handle step functions as well as if-statements in the memristor model [25].

Given its importance for reliable SPICE simulations, great attention has been put on the memristor model considered, though no general consensus has been reached on which conduction mechanism and memory equation (ME) better represent the wide spectrum of memristive behaviours [26]–[28]. The ME is a first order differential equation that links the current flowing or the voltage applied to the structure with its internal memory state. This results in a variety of both behavioural and physical-phenomenological models, enabling a trade-off between simplicity and accuracy. Roughly, memristor models can be classified into three groups: first, simple behavioural models [12], [29], [30] are useful in the early stages of circuit design, i.e. when a quick proof of concept is required. Second, device specific models such as the physical-phenomenological Pickett's [31] and the simple-phenomenological Bayat's [32] models for TiO_2-based MIM structures provide the highest accuracy. However, given their high computational cost, they may not be suitable in a scenario involving a large number of devices [33]. Last but not least, general phenomenological models such as the Yakopcic [34], TEAM [35], VTEAM [36], and Eshraghian [37] models can successfully fit certain experimental data. Nevertheless, the latter two groups rely on various internal equations or the introduction of an artificial window function in the ME (commonly used for modelling the SET/RESET transitions) which pose serious mathematical drawbacks causing convergence problems [33], [38]. A promising memristor compact model providing high simulation accuracy at reduced computational cost was presented by Miranda *et al.* in [39], [40]. Its closed-form expression for the I-V curve (continuous and differentiable) and the recursive nature of the state variable computation, makes it suitable for dealing with arbitrary input signals (continuous and discontinuous, differentiable and non-differentiable). This model is called the quasi-static memdiode model (QMM) and is the central subject of our analysis. Remarkably, the QMM has been explored so far as a single device or as simple series/anti-series/parallel/anti-parallel connections (just two devices) [39]–[41], and its application to the case of large CPAs for pattern recognition tasks is still to be addressed. This is the central topic of this work.

In this article, we demonstrate that the QMM not only accurately fits the experimental I-V loops for a wide range of memristive devices, but that it can also be used for the SPICE simulation of large-scale memristor-based CPAs intended for pattern recognition tasks without increasing the computational cost. By considering *ex-situ* training of a single layer perceptron as a case study and the classification of

grayscale images (among them the hand-written characters of the MNIST database [42]) for benchmarking, we present a comprehensive exploratory analysis of the CPA and QMM performances, addressing the dependence of the inference accuracy on: *i*) the R_{ON}/R_{OFF} ratio, *ii*) the line resistance (R_W), *iii*) the R_W/R_{ON} ratio, *iv*) the CPA size and image resolution, *v*) the partitioning schemes of the full CPA structure and *vi*) the device-to-device (D2D) variability. *vii*) The inference latency as a function of the feature size is also studied, as well as *viii*) the power consumption and *ix*) signal-to-noise ratio as a function of the R_{ON}/R_{OFF}, R_W and CPA size. In addition, we report *x*) a comparison between the QMM and a strictly linear model in terms of the CPA accuracy for different applied voltages and *xi*) a comparison of the computational complexity of CPAs comprising the QMM model against other memristor models. To the best of the authors' knowledge, such a detailed and comprehensive study within a unified framework and considering a realistic memristor model has not been published before. The rest of this article is organised as follows: Section II describes the fundamentals of the QMM: the *I*-*V* characteristic and the memory equation. Section III explains the CPA's training and simulation procedures. Section IV discusses the obtained simulation results in terms of the aforementioned features, providing useful design considerations and trade-offs. Finally, in Section V the general conclusions of this article are presented.

II. QUASI STATIC MEMDIODE MODEL

RRAM devices are based on the resistive switching (RS) mechanism, which in the case of CBRAMs and OxRAMs relies on the displacement of metal ions/oxygen vacancies within the dielectric film in a Metal-Insulator-Metal (MIM) structure. The movement is originated by the application of an external electrical stimulus, current or voltage [49]–[52]. This causes the alternate completion and destruction of a conductive filament (CF) spanning across the insulating film. The CF acts as a bridge allowing or blocking the pass of electrons in one or the opposite direction. For a ruptured CF, the device is in the high resistance state (HRS), often characterised by an exponential *I*-*V* relationship, while the completion of the CF leads to the low resistance state (LRS), which often exhibits a linear current or voltage dependence [48], [53]. Within these two extreme situations, the modulation of the CF transport properties by voltage-controlled redox reactions renders intermediate states. From the modelling viewpoint, the compact model originally proposed by Miranda in [39] and later extended by Patterson *et al.* in [40] is able to describe these major (LRS) and minor (HRS) *I*-*V* loops and the gradual transitions in bipolar resistive switches. This is accomplished by considering a nonlinear transport equation based on two identical opposite-biased diodes in series with a resistor, as shown in the inset of Fig. 2a. The *I*-*V* relationship resembles a diode with memory and that is why this device was termed memdiode. For the sake of completeness, the QMM is succinctly reviewed in the next paragraphs.

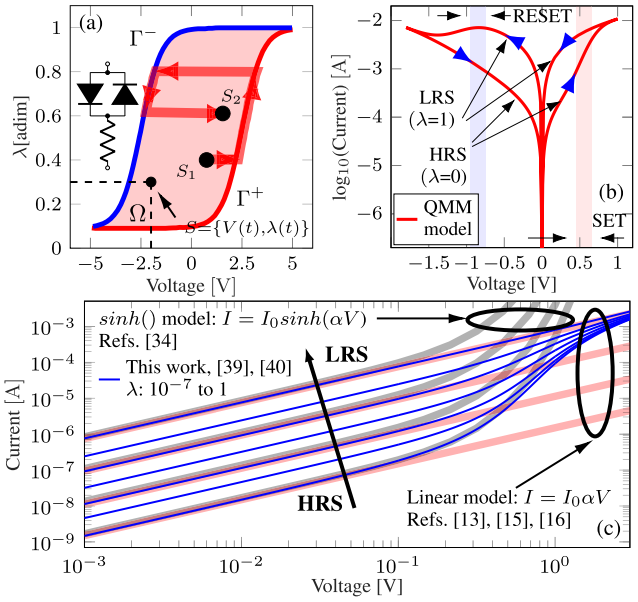


FIGURE 2. (a) Hysteron model with logistic ridge functions Γ^+ (Eq. 3) and Γ^- (Eq. 4). Ω is the space of feasible states S. The red thick faded line superimposed to the hysteron model indicates the trajectory of the state variable λ inside Ω from an initial (S_1) to a final (S_2) state. The inset in the left shows the equivalent circuit model for the current equation (1) including the series resistance. The diodes are driven by the memory state of the device and one diode is activated at a time. (b) Typical *I*-*V* characteristic for a memdiode obtained via simulation of the proposed model. Superimposed blue arrows indicate the current evolution. (c) Detail of the HRS (exponential) to LRS (linear) transition, showing a few intermediate states. Linear [13], [15], [16] and *sinh()*-like [34] models are superimposed for comparison.

Physically, the memdiode is associated with a potential barrier that controls the electron flow in the CF. The conduction properties of this non-linear device change according to the variation of this barrier. Because of the uncertainty in the area of the CF, instead of the potential barrier height, the diode current amplitude is used as the reference variable. Following Chua's memristive approach, the proposed model comprises two equations, one for the electron transport and a second equation for the memory state of the device (ME) which is based on a hysteresis operator. The equation for the *I*-*V* characteristic of a memdiode is given by the expression:

$$I = sgn(V)\left\{\frac{W\left(\alpha R I_0(\lambda) e^{\alpha(abs(V)+RI_0(\lambda))}\right)}{\alpha R} - I_0(\lambda)\right\} \quad (1)$$

where $I_0(\lambda) = I_{min}(1-\lambda) + I_{max}\lambda$ is the diode current amplitude, α a fitting constant and R a series resistance. Eq. 1 is the solution of a diode with series resistance and W is the Lambert function. I_{min} and I_{max} are the minimum and maximum values of the current amplitude, respectively, $abs(V)$ is the absolute value of the applied bias and $sgn()$ the sign function. As I_0 increases in Eq. 1, the *I*-*V* curve changes its shape from exponential to linear through a continuum of states as experimentally observed for this kind of devices. λ is a control parameter that runs between 0 (HRS) and 1 (LRS)

and is given by the recursive operator (Eq. 2):

$$\lambda(V) = min\left\{\Gamma^{-}(V), max\left[\lambda(\overleftarrow{V}), \Gamma^{+}(V)\right]\right\} \tag{2}$$

where *min*() and *max*() are the minimum and maximum functions, respectively and $\overleftarrow{V}$ is the voltage a timestep before V. The positive and negative ridge functions in Eq. 2, $\Gamma^{+}(V)$ and $\Gamma^{-}(V)$ represent the transitions from HRS to LRS (SET) and *vice versa* (RESET) and can be physically linked to the completion and destruction of the CF [48], [53], respectively. They are defined by Eqs. 3 and 4

$$\Gamma^{+}(V) = \left\{1 + e^{-\eta^{+}(V - V^{+})}\right\}^{-1} \tag{3}$$

$$\Gamma^{-}(V) = \left\{1 + e^{-\eta^{-}(V - V^{-})}\right\}^{-1} \tag{4}$$

where η^{+} and η^{-} are the transition rates and V^{+} and V^{-} the threshold voltages for SET and RESET, respectively. $\lambda(V)$ defines the so-called logistic hysteron or memory map of the device and keeps track of the history of the device as a function of the applied voltage (see Fig. 2a). λ calculated from Eq. 2 yields the transition from HRS to LRS and *vice versa* through a change in the properties of the diodes depicted in the inset of Fig. 2a. The combination of Eq. 1 and 2 results in a I-V loop as that illustrated in Fig. 2b, which starts in HRS ($\lambda = 0$) and evolves as indicated by the superimposed blue arrows. The name quasi-static comes from the fact that the characteristic switching time is assumed to be infinite for a state within the hysteron structure. The QMM can be transformed into a dynamic model by incorporating the time module described in [40].

Fig. 2c shows the HRS (exponential) to LRS (linear) transition, altogether with some intermediate states (solid blue lines). For comparison purposes, a linear [13], [15], [16] (faded-thick red lines) and *sinh()*-like models [34] (faded-thick black lines) are also plotted in Fig. 2c. Although the three models coincide at low voltages and exhibit a clear linear behaviour, significant discrepancies arise as the voltage increases. As it can be seen, first, the linear model is not able to capture the departure of the HRS curves at intermediate voltages. Second, the *sinh()*-like model, requires the simultaneous modification of multiple parameters to mimic the smooth linear-exponential to linear transition or even separate expressions for the HRS and LRS regimes [16], [54]. On the contrary, the memdiode model can accurately describe both HRS and LRS curves by solely changing a single parameter in the transport equation. As λ is swept from 10^{-7} to 1, I_0 in Eq. 1 varies between I_{min} and I_{max}, causing the I-V curve to gradually change its shape from linear-exponential (HRS regime) to linear (LRS regime). This is a consequence of the potential drop in the series resistance which linearizes the transport equation. Another relevant feature of the proposed model is that it can be described by a simple SPICE sub-circuit as shown in Table 1.

The model was put under test by fitting the experimental data extracted from different published works. In particular, Fig. 3 shows the results obtained for HfO_2 [43], Al_2O_3 [44], MnO_3 [45], CuO_2 [46], $La_{1-x}Ca_xMnO_3$ [47] and TaO_x [48] structures at room temperature under DC voltage sweeps. The experimental data were fitted by the SPICE model depicted in Table 1 based on Eqs. 1 and 2, and applying driving signals as described in the corresponding references. The fitting parameters are listed in each of the sub-figures of Fig. 3 as reference, as well as the details of the stack structure. It should be mentioned that the proposed QMM does not only provide a simple SPICE-compatible implementation for the resistive memory devices but also a versatile one, as it can accurately fit the I-V loops experimentally measured in a wide range of RRAM devices. Note that by the proper parameter selection, the QMM is capable of accounting for both gradual or abrupt transitions in the SET (see the SET in Figs. 3a and 3c) or RESET (see the RESET in Figs. 3d and 3f).

TABLE 1. Memdiode SPICE model code.

```
.subckt memdiode p n
  .param H0=0 CH0=1e-4 beta=0.5
  *Transition parameters
  .param etaset=10.5 vset=0.78 etares=7.2
  + vres=-0.79
  *I-V parameters
  .param imax=6.06e-3 imin=1.16e-4 alphamax=3.5
  + alphamin=5.6 rsmax=47 rsmin=47
  *Auxiliary functions
  .param I0(x)='imax*x+imin*(1-x)'
  .param A(x)='alphamax*x+alphamin*(1-x)'
  .param RS(x)='rsmax*x+rsmin*(1-x)'
  .param R(x)='1/(1+exp(-etares*(x-vres)))'
  .param S(x)='1/(1+exp(-etaset*(x-vset)))'
  *********H-V*********
  GH gnd! H cur='min(R(V(p,n)),max(S(V(p,n)),V(H)))'
  Rpar gnd! H R=1
  CH H gnd! C='CH0' IC='H0'
  *********I-V*********
  RS p D R='RS(V(H))'
  GD D n cur='I0(V(H))*(exp(beta*A(V(H))*V(D,n))-
  + exp(-(1-beta)*A(V(H))*V(D,n)))'
.ends memdiode
```

The model also allows the device to be set to a given conductance (resistance) value by using a Write-Verify iterative loop approach as the one schematically depicted in Fig. 4a. In such method, pulses of incremental amplitude are applied to the devices (Write) until the required conductance is reached (Verify) [56]. If the target conductance is exceeded, then increasing pulses with the opposite polarity are applied in a similar fashion to gradually reach the target conductance value (within an error margin). This writing methodology implies a transition as the one depicted in Fig. 2a by the red-thick faded line, where the incremental pulses cause the system to evolve from the initial state S_1 up to the final state S_2 following Γ^{+}. If the conductance target is exceeded, then the system moves down along Γ^{-} by the application of voltage pulses with the appropriate polarity. The latter procedure is experimentally presented in Fig. 4b for a RRAM stack comprising a SiO_x dielectric layer [55], and accurately modelled by the QMM model. In order to fully represent the intermediate states from LRS to HRS altogether with the major I-V loops, 7 successive ramped voltage pulses were considered omitting the verify step, as shown in the

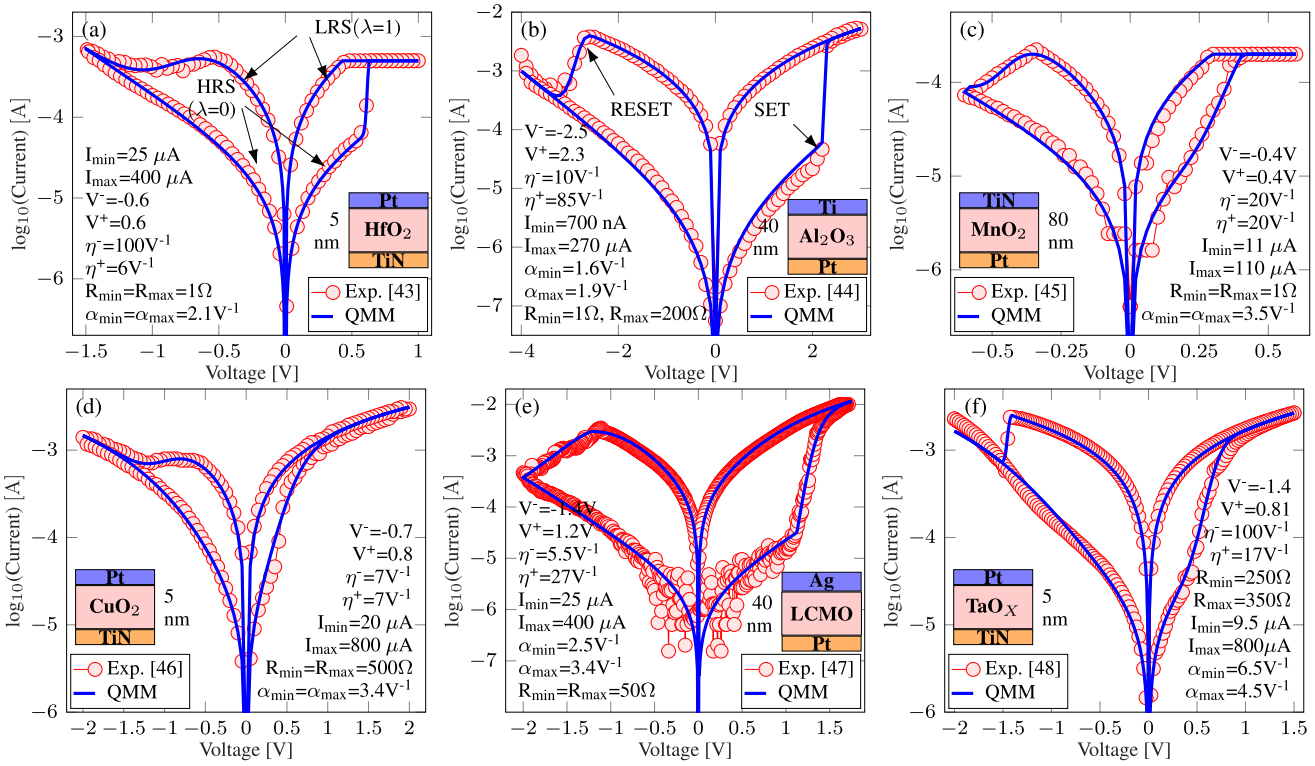


FIGURE 3. Experimental *I-V* loops of different materials reported in the literature fitted with the QMM model: (a) HfO_2 [43], (b) Al_2O_3 [44], (c) MnO_3 [45], (d) CuO_2 [46], (e) $La_{1-x}Ca_xMnO_3$ [47] and (f) TaO_X [48]. The QMM fitting parameters are shown for each case. As reference, the HRS and LRS curves are indicated in (a) and the SET and RESET points in (b). Note that in (a) a current compliance of 200 μA was imposed to prevent permanent dielectric breakdown, which can be also represented by the QMM.

inset of 4b. Additional details on the writing procedure are beyond the scope of this work and thereby will not be further discussed.

III. PROCEDURE FOR MEMDIODE CPA CREATION, TRAINING AND SIMULATION

To systematically evaluate the feasibility of the memdiode model when implemented in CPAs for large dataset pattern recognition tasks, a procedure for creating and simulating the single-layer feed-forward Artificial Neural Network (ANN) (single-layer perceptron, SLP) used as a case study is proposed. For the sake of simplicity, *ex-situ* supervised learning will be considered here. The recognition of patterns from different databases (MNIST [42], MINST-F [57] and MNIST-K [58]) is considered for benchmarking. The workflow is summarised in the chart depicted in Fig. 5. The tasks can be split into two parts: the first one comprises a set of MATLAB sub-routines for creating, training and writing the SPICE netlist for an ideal feed-forward ANN, while the second part relates to the SPICE simulation of the proposed circuit during the inference phase. It is worth mentioning that although a simpler approach than the more complex RRAM based neural networks explored in the literature (Multi-layer Perceptron [56], [59], [60], Convolutional Neural Networks [61], Spike Neural Networks [62], etc., see Supplementary Table 1), the SLP allows studying and clarifying the ANN limitations caused by parasitic effects and non-idealities occurring in the synaptic layers implemented with CPAs, as well as benchmarking the computational costs of the QMM based simulations against other available models.

Regarding the MATLAB-implemented part of the procedure, the first step consists in creating the image ($n \times n$ pixels) database. This includes rescaling each of the images of the original database (item 1 in the flowchart shown in Fig. 5). For the sake of brevity, the MNIST database will be considered for the vast majority of the studies in this article, yet the same procedures apply to all the previously mentioned datasets. The detailed results obtained with the MNIST-F and MNIST-K datasets can be found in the Supplementary Material (Supplementary Figs. 2-5). The MNIST (Modified National Institute of Standards and Technology) is a large database of handwritten digits from 0 to 9 commonly used for training and testing image processing systems including ANN in the field of machine learning. This database contains 60,000 training images and 10,000 testing images, both in grayscale and with a 28×28 pixels resolution [42]. A few examples of these images can be seen in Fig. 6a where the x and y axis stand for the pixel index. Pixel's brightness is codified in 256 gray levels between 0 (fully OFF, black) and 1 (fully ON, white). Databases MNIST-F (Fig. 6b) and MNIST-K (Fig. 6c) are similar to the MNIST database but comprising 10 different classes of fashion articles and handwritten Japanese Kanji ideograms, respectively.

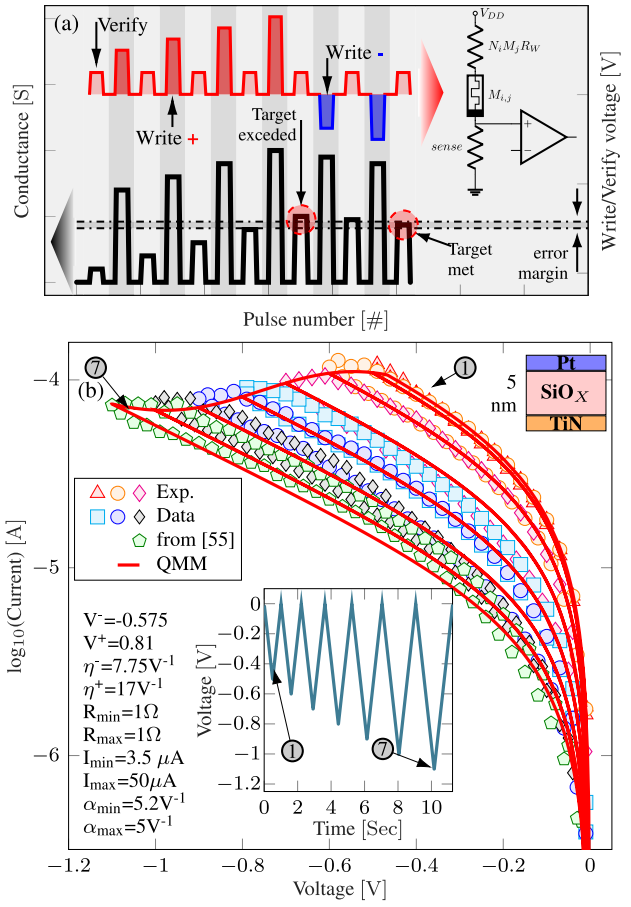


FIGURE 4. **(a) Schematic representation of the Write-Verify procedure for programming the CPA memdiodes. The upper waveform stands for the alternating write and verify voltage pulses, while the lower plot represents the associated conductance changes. A simplified circuit schematic for the writing connection is indicated in the upper right inset. (b) Experimental and simulation results for the reset characteristics of SiO$_X$ from UCL (Data from [55]) using the QMM model. Notice the control of the intermediate memory states. The inset shows the input signal.**

Then a software-based single layer perceptron of size $n^2 \times 10$ (with therefore 10 n^2 synapses) is created (2) and trained (3) using the previously rescaled database of training images (4). The ANN is *ex-situ* trained with the built-in functions of the MATLAB software for supervised learning, in this case considering the Scaled Conjugate Gradient (SCG) [63] as the training algorithm, as it provides a good trade-off between accuracy and learning time for the different datasets considered. Moreover, although the Levenberg-Marquardt (LM) learning algorithm [64] provides the highest accuracy at the cost of the highest CPU run-time among those considered [63]–[70], the difference between the inference accuracy obtained with this model and the SCG is not statistically significant, as shown in the 5-fold cross validation study with 10 repeats reported in Supplementary Fig. 6 and Supplementary Tables 2-7. Further details concerning the training function lie beyond the scope of this work, as we focus on the CPA-based implementation of the ANN. This training stage produces a $n^2 \times 10$ weight matrix $W_M \in \mathbb{R}$ (5). To allow

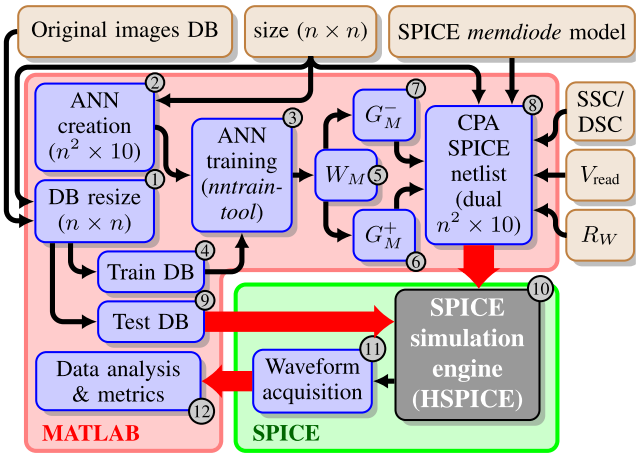


FIGURE 5. **Flowchart diagram for the simulation procedure. Starting with the image size specification, R_W, V_{read}, and connection scheme, the routine creates the database, trains the ANN (single layer perceptron), translates it into a CPA, adds the peripheral control circuit, performs the simulations and processes the results. MATLAB tasks are grouped by the red box and SPICE operations by the green box.**

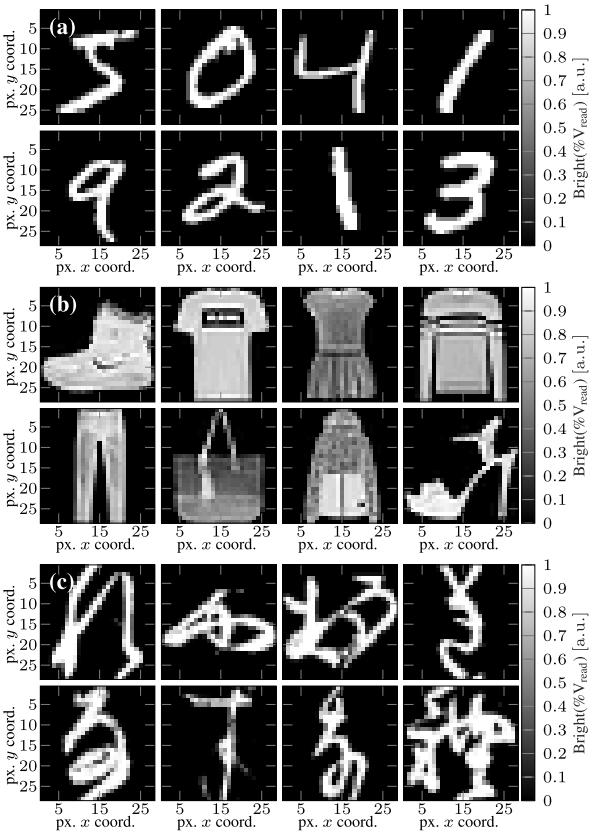


FIGURE 6. **Samples of the three different images databases considered in this article. In all cases images are represented in 28 × 28 px. Pixel brightness (or intensity) is codified in 256 levels ranging from 0 (fully OFF, black) to 1 (fully ON, white). (a) MNIST database of handwritten numeric digits, (b) MNIST-F database [57] of fashion articles and (c) MNIST-K database [58] of handwritten Kanji Japanese ideograms.**

rendering both the positive and negative elements of W_M with the always positive conductance of the CPA, each synaptic weight is implemented using two memdiodes as suggested in [71], [72] resulting in two CPAs of $n^2 \times 10$ (20 n^2 synapses).

TABLE 2. Conductance ranges used in the literature.

Work	Device Stack	R_{OFF} $(1/G_{min})$	R_{ON} $(1/G_{max})$	ratio	Model play
[56]	TiN/Ti/HfAlO/ TiN RRAM	∼1 MΩ	∼5 kΩ	200	∼C2
[14]	Ta/HfO_2/Pt RRAM	∼100 kΩ	∼2.5 kΩ	40	∼A2
[59]	GST-PCM	∼1 MΩ	∼50 kΩ	20	∼A1
[75]	Ta/Al_2O_3/ZrTe CBRAM	∼1 MΩ	∼5 kΩ	200	∼C2
[60]	Ta/HfO_2/Pd RRAM	∼10 kΩ	∼1 kΩ	10	∼A3

This representation method has been chosen as it doubles the dynamic range of the CPA, making it less susceptible to noise and variability [73]. Thereby W_M is split into two matrices W_M^+ and W_M^- as:

$$w^+_{M_{i,j}} \begin{cases} w_{M_{i,j}}, & w_{M_{i,j}} > 0 \\ 0, & w_{M_{i,j}} \leq 0 \end{cases} \quad (5)$$

$$w^-_{M_{i,j}} \begin{cases} 0, & w_{M_{i,j}} \geq 0 \\ -w_{M_{i,j}}, & w_{M_{i,j}} < 0 \end{cases} \quad (6)$$

each of them containing only positive weights, so that $W_M = W_M^+ - W_M^-$. In the next step, the conductance matrices G_M^+ and G_M^- (6 and 7) to be mapped into the CPAs are calculated by the linear transformation [20], [74]:

$$G_M^{+,-} = \frac{G_{max} - G_{min}}{max\{W_M\} - min\{W_M\}} W_M^{+,-} + \left[G_{max} - \frac{(G_{max} - G_{min})\ max\{W_M\}}{max\{W_M\} - min\{W_M\}} \right] \quad (7)$$

where $[G_{min}, G_{max}]$ is a selected conductance range for a linear computation in matrix-vector calculations. For simplicity, we consider $G_{max} = G_{LRS} = 1/R_{ON}$ and $G_{min} = G_{HRS} = 1/R_{OFF}$, where $max\{W_M\}$ and $min\{W_M\}$ are the maximum and minimum synaptic weight values in the software obtained W_M. In this way, the synaptic weights in the W_M^+ and W_M^- matrices are converted to conductance values within the range $[G_{HRS}, G_{LRS}]$.

The subsequent sub-routines generate the circuit netlist for the dual-$n^2 \times 10$ memdiode CPA-based ANN (8), adding the parasitic wire resistance, connection scheme, and control logic necessary to perform the inference phase. Each memdiode in the CPAs is set to the corresponding conductance value from the G_M^+ and G_M^- matrices by adjusting the control parameter λ (H0 in the sub-circuit). The required value of λ is obtained by solving Eq. 1 for $I = g_{i,j} \cdot V$, with $g_{i,j}$ being each of the elements of G_M^+ (G_M^-). In our implementation, a Dual Side Connection (DSC) scheme was considered, as shown in the simplified equivalent circuit from Fig. 7. Despite the increased peripheral circuitry complexity, this scheme improves the voltage delivery to each synapse [13] by connecting the wordline terminals to the same input

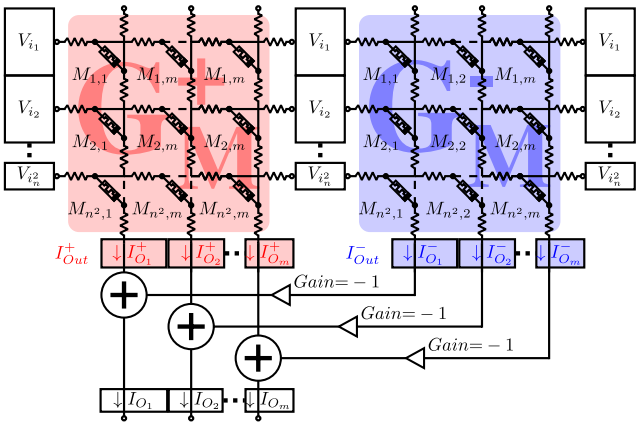


FIGURE 7. Simplified equivalent circuit of the single layer perceptron with n^2 inputs and m outputs ($n^2 \times m$). 2 CPAs are considered, one of them is used to represent the positive weights and the other the negative weights. The Dual Connection Scheme (DSC) implies biasing the wordline terminals. V_{i_1}-$V_{i_{n^2}}$ indicates the elements of the input image vectors ($n^2 \times 1$).

stimuli. The input stimuli are obtained by unrolling each of the rescaled grayscale $n \times n$ images of the test database (9) into an equivalent $n^2 \times 1$ vector and scaling it by a voltage V_{read}. V_{read} is chosen such as to prevent altering the memdiode states during the inference simulation. In this way, during the inference process each of the test images is presented to the CPA as a vector of analogue voltages in the range $[0, V_{read}]$. Once the circuit netlist has been generated, it is passed to a SPICE simulator (10) which evaluates the voltage and current distributions in the CPA circuit while it processes and classifies the input images (11), and then passes the resulting waveform back to the MATLAB routine for metrics extraction (12). For brevity, in this article we present only the total accuracy metric as well as the confusion matrix. Other relevant metrics such as the Sensitivity, Specificity, Precision, F-1 score and κ-coefficient are summarised in the Supplementary Material (Supplementary Figs. 1-5). In this article we consider the HSPICE simulator from the Synopsys® CAD Suite.

IV. SIMULATION RESULTS AND DISCUSSION

The inference accuracy of the memdiode-based CPA considered in this work was evaluated using the MNIST-like testing sets. These sets contain 10,000 images not used during the training phase. An exploratory analysis of the design space was carried out with the aim of evaluating the classification performance in terms of three different variables with ranges within the values reported in the literature. These variables are: the memristor resistance window (R_{ON} and R_{OFF}) [14], [56], [59], [60], [75], the wire resistance of the CPA interconnections (R_W) [15], [76] and the representation size of the dataset images ($n \times n$ pixels) [56], [59], [60]. Additionally, the accuracy sensitivity to device-to-device variability, and Signal-to-Noise ratio are also analysed. Besides the classification accuracy, the CPA power consumption and inference latency are also reported in this study. For clarity these aspects are organised in Sub-Sections IV-A to IV-D, respectively.

Finally in Sub-Section IV-E the inference accuracy and computational cost is presented for different datasets and memristor models.

A. INFLUENCE OF THE DEVICE RESISTANCE WINDOW (R_{ON} AND R_{OFF})

Low-power operation of memdiode-based CPAs requires RRAM devices with reduced I_{HRS} and I_{LRS} currents, or in other words, large R_{OFF} and R_{ON} resistances, respectively. This can be achieved by limiting the CF size [77], at the cost of introducing significant variability in the resistance value, especially in R_{OFF} [56], [77]. This trade-off between device variability and low-current operation can be partially solved by using memory cells with large resistance windows. From this standpoint, a number of device stacks and RS mechanisms (RRAM, CBRAM, PCM) associated with different R_{ON} and R_{OFF} values were studied for neuromorphic applications (see Table 2) [14], [56], [59], [60], [75]. In order to address how the memdiode model is able to cope with such variety of resistance windows and in turn how this variety affects the classification accuracy, different I-V loops were considered by carefully tuning the fitting parameters of the model described in Sec. II to render a series of scaled I-V curves with different $I_{HRS} - I_{LRS}$ (R_{OFF}/R_{ON}) ratios.

Starting from a reference curve, Fig. 8a shows four I-V loops (including SET and RESET) where both I_{HRS} and I_{LRS} are simultaneously increased by a factor of ∼10, thereby keeping the I_{LRS}/I_{HRS} ratio constant (∼10). Each loop corresponds to a different play of the model referred to as $A1$-$A4$ (constant R_{OFF}/R_{ON} ratio of 10 for R_{OFF} equal to 1 MΩ, 100 kΩ, 10 kΩ and 1 kΩ, respectively). Similarly, Fig. 8b and 8c explore the case of scaling the current trace for a single resistance state (I_{HRS} or I_{LRS}): First, in Fig. 8b, I_{HRS} is systematically increased while keeping I_{LRS} constant, causing the reduction of the I_{LRS}/I_{HRS} ratio by a factor of ∼10 for each model play (named $B1$-$B4$, with R_{ON} fixed at 1 kΩ and R_{OFF}/R_{ON} ratios of $\sim 10^4,10^3,10^2,10^1$, respectively). The complementary case (scaling I_{LRS} while keeping I_{HRS} fixed) is evaluated in Fig. 8c and modelled by plays $C1$-$C4$ (constant R_{OFF} of 1 MΩ and R_{OFF}/R_{ON} ratios of $\sim 10^1,10^2,10^3,10^4$, respectively). To provide a guide to the eye, the values of I_{HRS} and I_{LRS} evaluated at V_{read} are pinpointed for model plays $A2$, $B2$ and $C2$ in Figs. 8a, 8b and 8c, respectively.

CPA simulations were performed for each model play ($A1$-$A4$, $B1$-$B4$ and $C1$-$C4$) considering DSC, $R_W = 0.1 - 10$ Ω, $V_{read} = 300$ mV and assuming full-size images (28×28 px., i.e. the resulting ANN comprises 15,680 synapses). For brevity only the MNIST dataset is considered in this subsection. The obtained classification accuracy is presented in Figs. 8d and 8e against the ON (R_{ON}) and OFF (R_{OFF}) resistance of each model play, respectively. Figure 8d shows a clear inference accuracy degradation as R_{OFF} decreases (the resistance window shifts upward), for model plays $A1$ to $A4$, with the recognition performance for $A1$ being close to the software results for a single layer perceptron (90.9%) [78]. On the contrary, no clear dependence exists between the

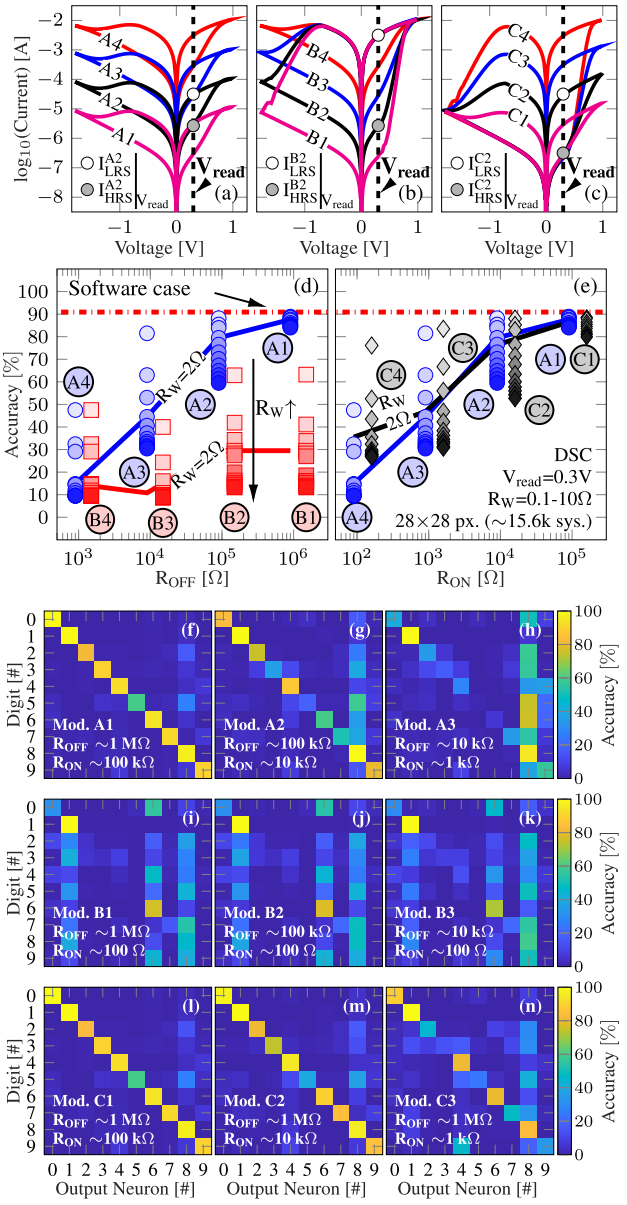


FIGURE 8. Impact of the memdiode resistance window on the detection performance. Different resistance windows are considered (R_{ON} and R_{OFF}) by scaling the LRS and HRS curves (a) HRS and LRS are scaled, (b) only HRS is scaled and (c) only LRS is scaled. Recognition accuracy as function of the models OFF resistance (d) and ON resistance (e). For every model each point represents a different value of R_W, swept from 100 mΩ to 10 Ω (darker markers indicate higher values or R_W). CPAs are connected from both sides (DSC) and images are not downsized (img. resolution: 28 × 28 px.). Confusion matrix with R_W =4.53 Ω for model plays (f) *A1*, (g) *A2*, (h) *A3*, (i) *B1*, (j) *B2*, (k) *B3*, (l) *C1*, (m) *C2* and (n) *C3*.

classification accuracy and R_{OFF} when testing model plays $B1$-$B4$ (the resistance window widens for a fixed R_{ON}). Last, the inference performance obtained for model plays $C1$-$C4$ (widening resistance window for a fixed R_{OFF}) is studied as function of R_{ON} and compared against the results for $A1$-$A4$ in Fig. 8e. Both $A1$-$A4$ and $C1$-$C4$ present an almost identical dependence on R_{ON} despite the remarkable differences in the resistance window scaling. To shed more light on these aspects, Figs. 8f to 8n show the so-called confusion matrices

for models *Ai-Ci* respectively (models *A*4-*C*4 are not shown as the error is too high to remain a case of interest) and considering the case $R_W = 4.53\ \Omega$. The confusion matrix is a useful tool for capturing the ability of a neural network to associate each input pattern with its corresponding class (in this case a digit from 0 to 9) and allows to graphically represent the inference accuracy for each possible input. Despite the large inference error reported for all digits when considering model plays *B*1-*B*3 (Figs. 8i-8k), plays *A*1-*A*3 and *C*1-*C*3 show a similar behaviour, with *C*2 showing a better classification performance than *A*2, due to a higher R_{ON}/R_{OFF} ratio for the same R_{ON}.

All these observations indicate that although the increase in the resistance window is expected to improve the classification accuracy, the simulated results are strongly limited by the value of the ON resistance, as shown by the very similar behaviour of model plays *A*1-*A*4 and *C*1-*C*4. The reason behind this observation is that the parasitic voltage drop along the selected wordline and bitlines substantially increases as the ON state resistance (R_{ON}) reduces. As a result, the RRAM devices "see" a much smaller effective read voltage than the applied voltage to the CPA inputs (namely the read margin, defined as V_{cell}/V_{read}) [76]. This is well in agreement with previous results by Liang *et al.* [76] showing that the read margin is mainly governed by R_{ON} with a much reduced dependency on the R_{OFF}/R_{ON} ratio, showing a clear degradation of the read margin as R_{ON} reduces. In fact, if the inference accuracy data from Figs. 8d-8e is plotted against the corresponding read margin, there is a strong correlation between them for average read margin values below 10% (see Fig. 9a). As the voltage drop in the interconnect wires is jointly determined by both the memdiodes and wire resistance, accuracy dependence on the wire resistances will be specifically addressed in the following section.

It is worth mentioning that the read margin changes among the devices in the CPA, with the memdiodes located close to the input drivers or output terminals having a greater read margin than those being further away, as depicted in Fig. 9b. Given that RRAM devices normally present a linear-exponential characteristic [81] in the HRS regime, this may lead to errors in the computation of the cell's current when using a simple linear model for the RRAM device [13], [15], [16] as illustrated in Fig. 9c. For a given linear model, fit for I_{LRS} and I_{HRS} at a nominal V_{read}, the reduction of the effective read voltage applied to the cells as the read margin decreases lead to an overestimation of the device current, causing errors in the pattern recognition and increasing power consumption. Instead, the model considered in this work [39], [40] can easily handle the effects of the read voltage reduction, given the linear-exponential – linear *I*-*V* characteristic considered. This is represented in Fig. 9d, where two different linear approximations of the HRS characteristic of the model play *C*2 (fitted for nominal V_{read} values of 0.2 and 0.6V) are used to compute the CPA accuracy under different read voltages. As expected, the simulation results show no dependence on the read voltage, since such model does not account

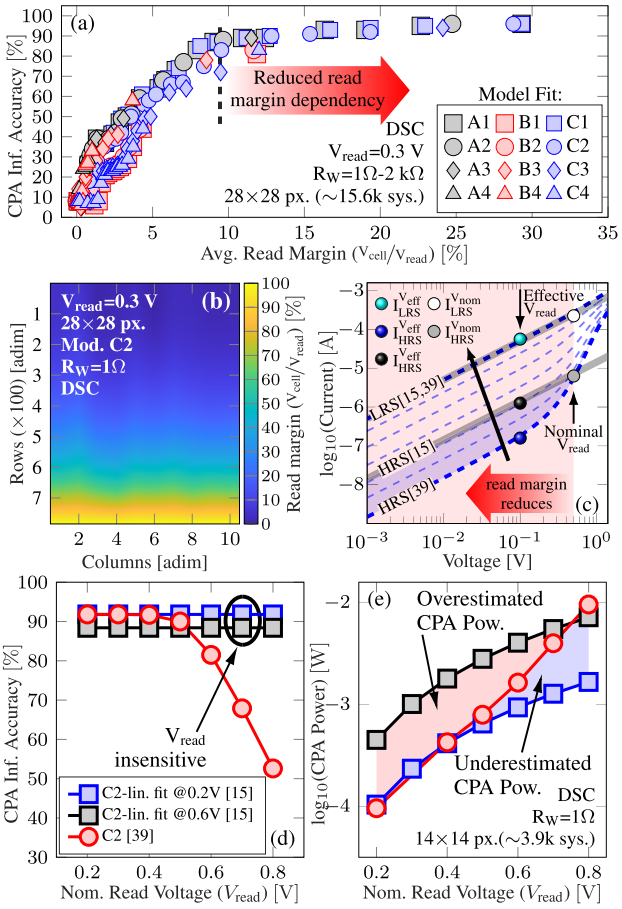


FIGURE 9. (a) CPA Inference accuracy for different model plays (R_{ON}/R_{OFF}) and wire resistances (R_W) plotted against its associated read margin. Avg. Read margins below 10% increasingly affects the CPA inference accuracy. (b) Map of read margin as function of the spatial location for a 784 × 10 CPA, with R_W=1 Ω. Significant voltage variation can be seen. (c) *I*-*V* characteristics of the memdiode showing the exponential (HRS) to linear (LRS) transition by varying λ. The red shaded region indicates the possible voltages applied to the device as the read margin reduces. I_{HRS} and I_{LRS} currents are pinpointed at nominal V_{read} with the grey and white circle markers, respectively. Overestimation of I_{HRS} may occur when considering a linear model for the HRS regime and lower effective V_{read} voltages as indicated by the cyan, blue and black ball markers. The Linear model does not account for the R_{ON}/R_{OFF} ratio reduction with V_{read}, and thereby cannot account for its impact on the accuracy (d), as well as on the CPA Power consumption (e).

for the HRS-LRS transition. Instead, the QMM model shows how the increment of the read voltage generates a reduction of the devices resistance window, clearly indicating a limit for the read voltages (in this case, V_{read}<~0.5V). In the same way, the fully-linear approach causes an overestimation or underestimation of the CPA power consumption, as shown in Fig. 9e. Thereby, they are only valid for a limited range of the possible read voltages.

Lastly, it is worth evaluating the CPA power consumption as a function of the memdiode R_{ON}/R_{OFF} ratio and CPA size, considering the framework proposed in this article. This is shown in Fig. 10a for model plays *C*1-*C*4. It can be seen that, as expected, as the R_{ON} value decreases from *C*1 to *C*4, the power consumption increases. Regarding the CPA size, there is an almost linear dependence between the power

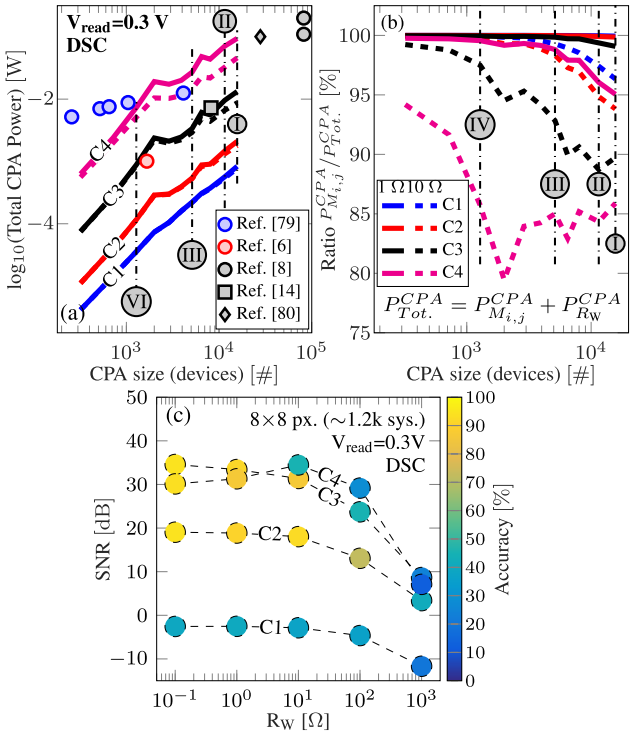


FIGURE 10. **(a) CPA Power consumption vs. CPA size (in number of devices). Note that Power increases as R_{ON} reduces from model plays (*C*1-*C*4). As reference, reported CPA power for other works [6], [8], [14], [79], [80] is also plotted. (b) Ratio between the memdiode power and the total CPA power consumption. (c) Obtained Signal-to-Noise ratio (SNR) for different model plays (*C*1-*C*4) as function of the line resistance. The marker color indicates the inference accuracy for each simulation case. Note that reduced R_{ON} in *C*1 comes at the cost of a reduced SNR and thereby limits accuracy.**

consumption and the number of devices in the CPA. The influence of the wire resistance is also another parameter to have in mind, specially when considering resistive devices with a reduced R_{ON}. Also, when considering the parasitic wire resistance, it is important to address the ratio between power dissipated in the memdiodes and the total CPA power, which also includes the power in the interconnections. This provides a possible metric of the power efficiency. In this context, the efficiency reduces for large CPAs implemented with devices with low R_{ON}. Although Figs. 8e, 10b and 10a jointly point toward $C1$ being the best case scenario, as it combines low power operation and high inference accuracy, if the thermal and flicker noise sources form the CPA circuit are considered, the Signal-to-Noise ratio (SNR) of the output signal is expected to degrade and thereby cause classification errors. This is shown in Fig. 10c, where the extracted SNR from the simulations is plotted against different values of R_W for the $C1$-$C4$ model plays. The marker color indicates the associated accuracy for each simulation based on the colorbar scale in the right side. Interestingly, the classification accuracy from model play $C1$ significantly reduces when considering the added noise, as the SNR severely degrades (the high values of R_{ON} implies reduced CPA currents, in fact below the noise floor). Thereby there is a trade-off between improving the SNR and reducing the parasitic voltage drop in the line interconnections. In this work, model play $C2$ (R_{ON} ∼ 10 kΩ, R_{OFF} ∼1 MΩ) shows the optimal case among those considered.

B. INFLUENCE OF THE WIRE PARASITICS (R_W AND C_W)

In a realistic scenario, the metallic wordlines (WL) and bitlines (BL) interconnections of the CPA are characterised by a parasitic wire resistance (R_W) and a parasitic wire capacitance (C_W), both a function of the wire geometry and material properties. While the first one severely degrades the read margin of the CPA, the second one affects the operational speed of the CPA, often measured as the CPA latency, defined as the settling time of the output vector after a change in the input pattern. In this Sub-Section both limiting phenomena are addressed within the framework of the QMM-based modelling of the CPA.

Regarding the first point, the resistance between the nearest cells ($R = \rho \cdot L/(W \cdot T)$, L and W are the wire length between adjacent cells and wire width respectively, and taken equal to the feature size (F) and T is the wire thickness) ranges from 1 to 10 Ω when T >10 nm is assumed, as the resistivity of conventional metal wires (ρ) ranges from 10^{-8} to $10^{-7}\Omega\cdot$m. Thereby, for a $4F^2$ cross-point structure, the resistance of the interconnect wires (R_W) between two adjacent cells can be estimated to be ∼4.53, 2.97, and 1.55 Ω under the 16, 22 and 32 nm technology nodes, respectively [15]. Nevertheless, for the novel technology nodes (10 nm and below) both surface and grain boundary scattering causes a size-dependent resistivity of Cu wires [82]–[84] as the mean free path of electrons becomes comparable to the wire dimensions. These two effects are well-known and can be quantified using the Fuchs-Sondheimer (FS) [85] and the Mayadas-Shatzkes (MS) [86] models, revealing that for highly scaled nodes (∼5 nm) R_W can be as large as ∼100 kΩ [76]. The increase in ρ is shown by Eq. 8:

$$\frac{\rho}{\rho_{Cu}} = \frac{3}{4}(1-p)\frac{l_0}{W} + 3\left[\frac{1}{3} - \frac{\alpha}{2} + \alpha^2 - \alpha^3 ln\left(1+\frac{1}{\alpha}\right)\right]^{-1}, \quad \alpha = \frac{l_0}{d}\frac{R}{1-R} \tag{8}$$

where ρ_{Cu} is the bulk Cu resistivity (1.9 $\mu\Omega\cdot$cm), p is the specular scattering fraction, l_0 is the bulk mean free path for electrons in Cu (39 nm at room temperature), W is the wire width, R is the probability for electrons to reflect at the grain boundaries, and d is the average grain size. In this article we consider $p = 0.25$ and $R = 0.3$ based on the average values reported in the literature, and d is assumed equal to the wire width [82], [84]. An aspect ratio of 1 is assumed (W equal to the feature size) with a barrier thickness of 2nm on each side of the wire [76]. To study the impact of the voltage drop across the wire interconnects, in this Sub-Section the memdiode-based CPA performance is addressed for the 28 × 28 px. images (∼15.6k synapses) with R_W

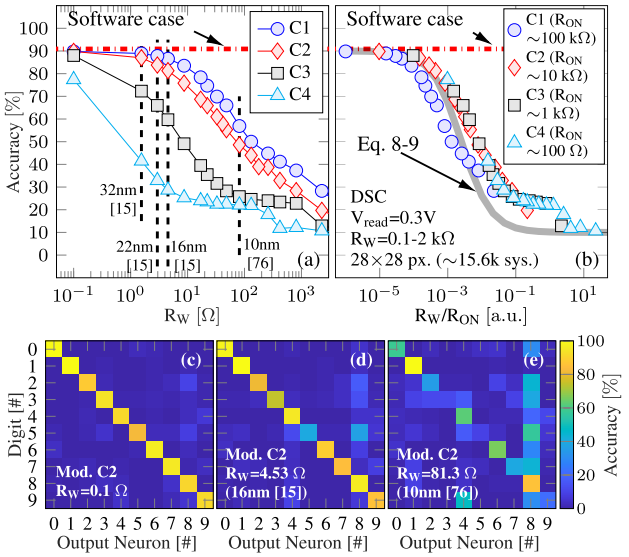


FIGURE 11. (a) Impact of the wire resistance (R_W) on the detection performance considering four different resistance windows corresponding to models plays $C1$-$C4$. (b) Inference accuracy is plotted against the R_W/R_{ON} ratio showing a unified trend among all model plays. Confusion matrix for model $C2$, with R_W equal to (c) 0.1 Ω, (d) 4.53 Ω (16 nm [15]) and (e) 81.3 Ω (10 nm [76]). In all cases CPAs are connected from both sides (DSC) and images are not downsized (image resolution: 28 × 28 px.).

ranging between 10^{-1}-10^4 Ω and compared against the ideal case ($R_W = 0$ Ω). For the sake of brevity, only model plays $C1$-$C4$ are considered as they exhibit the best performance as shown in Sub-Section IV-A.

As it can be seen in Fig. 11a, detection accuracy shows a sustained reduction as R_W increases from 10^{-1} to ∼10^4 Ω, downshifted as R_{ON} scales down from 100 kΩ to 100 Ω in model plays $C1$-$C4$. For the $C2$ case ($R_{ON} = 10$ kΩ and $R_{OFF} = 1$ MΩ) the accuracy remains above 80% when considering the wire resistance expected for the 32 (1.55Ω), 22 (2.95Ω) and 16 (4.53Ω) nm technology nodes (87.00%, 83.84% and 81.35% respectively). Nevertheless, severe degradation is to be expected for further scaling as the wire resistance rapidly increases for the 10 nm node and beyond. For the latter, a resistance $R_W = 80$ Ω is predicted in [76], causing the inference accuracy to drop to 48.63%. Then, intensive research is needed in the material engineering domain to take advantage of the $4F^2$ scalability of the CPA without penalising the inference performance.

As discussed in Sub-Section IV-A, the CPA inference accuracy is severely affected by the CPA's read margin. Given that each memdiode is in series connection with a number of interconnect resistors with value R_W, it can be demonstrated that the read margin is proportional to the ratio R_W/R_{ON}. Thus, this metric is used to represent the inference accuracy data in Fig. 11b. Interestingly, data from different model plays ($C1$-$C4$) and R_W exhibit a unique trend. For values of R_W/R_{ON} below 10^{-4} there is no influence of R_W, as it results negligible against the value of R_{ON} and the entire input voltage is applied to each RRAM cell. Similarly, when the R_W/R_{ON} ratio surpasses the 10^{-1} threshold, the voltage drop across R_W dominates the voltage distribution, causing significant recognition errors. For values in between these two limits, there is a constant increment in the part of the applied voltage that drops across the wire resistance, and therefore a sustained decrease in the detection accuracy is observed. Such a behaviour can be approximately captured by the following empirical model derived from assuming a simple voltage divider between resistors R_{ON} and $\langle S_{ij}\rangle.R_W$

$$Acc = \frac{Acc_{R_W=0\Omega} + Acc_{min}\frac{R_W}{R_{ON}}\langle S_{ij}\rangle}{1 + \langle S_{ij}\rangle\frac{R_W}{R_{ON}}} \tag{9}$$

$$\langle S_{ij}\rangle = \frac{n^2 + M}{2} + 1 \tag{10}$$

where $\langle S_{ij}\rangle$ is the average number of interconnect resistances of value R_W in series connection to each memdiode in a $n^2 \times M$ CPA, $Acc_{R_W=0\Omega}$ is the inference accuracy for the ideal case ($R_W = 0\Omega$) and Acc_{min} is the minimum accuracy (∼10% as there are 10 possible outputs). As an example of the RW impact on the inference accuracy, the confusion matrices obtained from model play $C2$ with R_W being equal to 0.1, 4.53 and 81.3 Ω are shown in Figs. 11c-11e, respectively. The gradual decrease of the detection accuracy is visible for each individual digit. Similar trends were obtained for the MNIST-F and MNIST-K datasets and they are reported in the Supplementary Material (Supplementary Figs. 2 and 4). Although showing a slightly lower inference accuracy than the MNIST case, the CPA implementation also matches the ideal software-based model when R_W tends to 0.

With regard to the second point, and in addition to the wire resistance, the parasitic line capacitance (C_W) has a major role in determining the CPA latency, which is an essential performance metric of interconnect. Moreover, with the dramatic increase of R_W at the single-digit-nm scaling regime, wire latency severely deteriorates and has become a primary limit to achieving fast CPA operation. According to Liang *et al.* and Meindl [76], [87] the wire latency (τ) can be modelled as in Eq. 11:

$$\tau = R_W C_W L^2 \tag{11}$$

in which L is the total wire length, R_W is the wire resistance per unit length calculated according to Eq. 8 and C_W is the wire capacitance per unit length calculated as in [88]:

$$\begin{aligned} C_L = {} & \varepsilon\frac{1}{2}\left[1.15\frac{W}{H} + 2.8\left(\frac{T}{H}\right)^{0,222}\right] \\ & + \varepsilon 2\left[0.03\frac{W}{H} + 0.83\frac{T}{H} - 0.07\left(\frac{T}{H}\right)^{0.222}\right] \\ & \times \left(\frac{H}{S}\right)^{1,34} \end{aligned} \tag{12}$$

where W and T are the width and thickness of the wire (assume $W = T$), S is the inter-wire spacing (assume $S = W$), and H is the inter-metal layer spacing (assumed to be 20nm). ϵ is the dielectric permittivity of the inter-metal layer

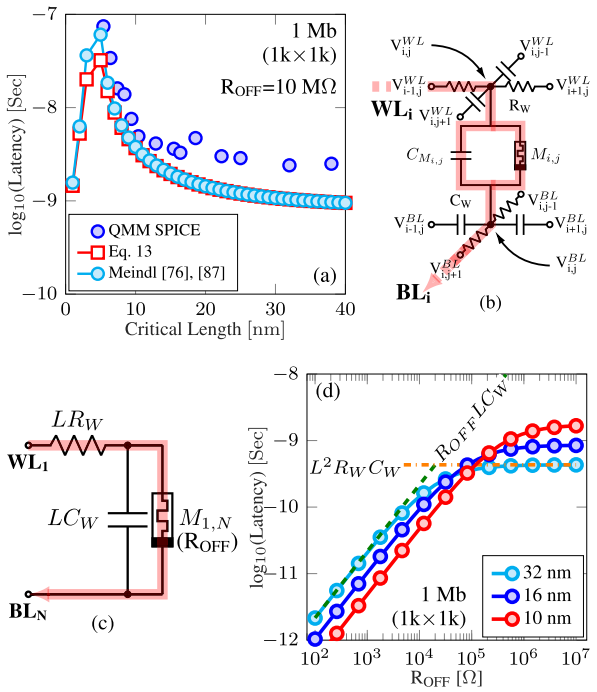


FIGURE 12. (a) Expected inference latency for a CPA of 1Mb (1,000 × 1.000) as function of the process feature size, considering the model proposed by Meindl *et al.* in [76], [87], SPICE simulations considering the QMM, and the simplified circuital model depicted by Eq. 13. (b) Circuit schematic of a RRAM cell in a CPA structure considering the associated wire parasitic resistance and capacitance. (c) Schematic representation of the simplified model of Eq. 13. (d) Estimated Latency based on Eq. 13 as function of R_{OFF}.

and is assumed to be 20 times the vacuum permittivity. It can be inferred from Equation 11 that the larger the memory array size and the smaller the wire width (higher R_W), the larger the wire latency. The wire latency calculated with Eq. 11 is presented in Figure 12a for a 1Mb memory CPA as a function of the feature size showing how it severely degrades in highly scaled processes.

Given its relevance, in this article we studied the inference latency within the framework of SPICE simulations using the QMM model. To do so, the inter-wire and wire-to-ground capacitance are included in the SPICE netlist. This results in a RRAM cell as the one schematically represented in Fig. 12b. In this way, each memdiode terminal is connected to two adjacent wire resistors, and two inter-line wire capacitors. Simulation results properly follows the results from Liang *et al.* [76] and Meindl [87], capturing the latency increase for the smaller feature sizes.

An interesting point to notice about the approach followed by Meindl [87], is that it does not take into account the resistance of the memristor devices. To do so, we propose an equivalent simplified circuit for the latency calculation in a CPA structure, as shown in Fig. 12c. Then the latency is studied as the settling time of the output current at the BL terminal, which is indicated in Eq. 13.

$$\tau = \frac{L^2 R_W C_W}{1 + L R_W g_{min}} = \frac{L^2 R_W R_{OFF} C_W}{L R_W + R_{OFF}} \tag{13}$$

Following this approach, the impact of the memristor resistance on the inference latency is studied in Fig. 12d. The resistance in the OFF state is considered, as it supposes the worst case scenario (higher latency). As the OFF resistance increases, the latency value calculated by Eq. 13 approaches the limit defined by Eq. 11. On the contrary, as the OFF state resistance decreases, the latency also decreases asymptotically to the OFF resistance.

C. INFLUENCE OF THE IMAGE SIZE (n × n PX.)

The MNIST database has been widely used in the literature to benchmark the inference accuracy of CPA-based ANNs. To deal with the hardware limitations imposed by the size of the available CPAs, it is a usual practice to downscale the pixelation of the images in the database. For example, both the training and testing images are resized to 8 × 8 px. in [60], 14 × 14 px. in [56] and 22 × 24 px. in [59] by using the bicubic interpolation method. However, it is clear from Eq. 9 and 10 that for a given R_W/R_{ON} ratio, the size of the CPA (determined by the image size) directly affects the inference accuracy. Thereby it is reasonable to expect the classification accuracy to increase for down-scaled images. Nevertheless, MNIST input images become barely recognisable for the human eye when resolution is reduced beyond 12 × 12 px, as shown in Fig. 13a. This issue also affects the accuracy of the software-based ANNs, as shown in Fig. 13b for images smaller than 8 × 8 px (CPA size ∼1.2k synapses). This suggests that there may be a trade-off between readability loss and read margin improvement, resulting in an optimal image (CPA) size. To investigate how the representation size of the images affects the general performance of the memdiode-based CPA ANN, the inference accuracy has been studied for image sizes ranging from 3 × 3 px. to the full representation size (28 × 28 px.). Images were down-scaled with the bicubic interpolation method and the wire resistance was parametrically swept from 1 to 100 Ω. Cases corresponding to 28 × 28 (I), 20 × 20 (II), 12 × 12 (III) and 8 × 8 px. (IV) image resolution used as example in Fig. 13a are pinpointed in Fig. 13b for comparative purposes.

Simulation results in Fig. 13b reveal that variations in R_W have more impact in large arrays than in the small ones. In this way, Case I exhibits a clear degradation of the inference accuracy, as it goes from 88.21% down to 46.43% for R_W 1-100 Ω. Instead, case IV presents a more reduced variation of the inference accuracy, as it changes between 88.91% and 84%. This could be explained if we consider that for a given value of the R_W/R_{ON} ratio, larger arrays are subjected to a higher total voltage drop across the wire resistances, reducing the read margin and hence the pattern recognition accuracy. In this regard, for each value of R_W there is a CPA size (image resolution) that maximises the inference accuracy. Such values are indicated in Fig. 13b, being 2000, 3380 and 5780 devices, for R_W equal to 100, 10 and 1 Ω, respectively, clearly following a decreasing trend as R_W grows. Deeper analysis of the CPA size influence is provided in Fig. 13c, where the accuracy results obtained for

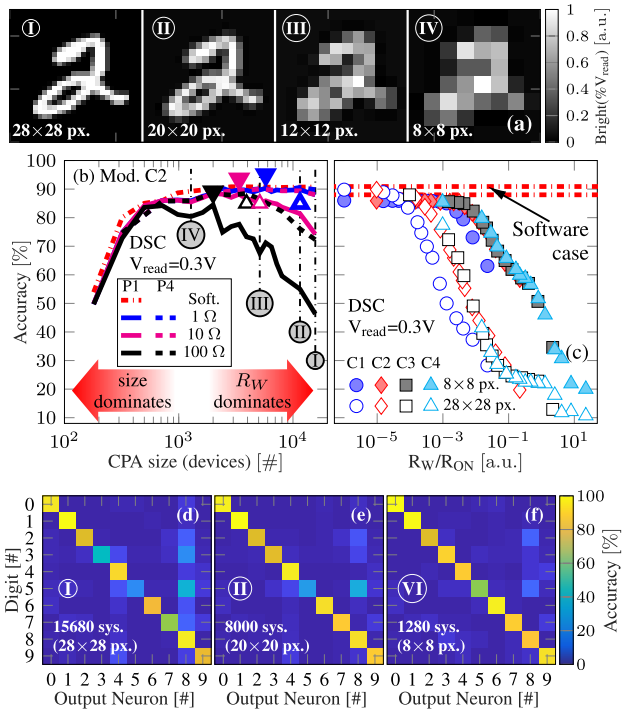


FIGURE 13. (a) Readability loss as the resolution decreases from 28 × 28 px (case I) to 8 × 8 (case IV). (b) Impact of the CPA size on the detection performance for model *C*2 and with R_W swept from 1 Ω to 100 Ω. Two different partitions schemes are considered: P1 indicates non-partitioned arrays whereas P4 stands for CPA partitioned in 4 sub-arrays. Markers indicate max. size with max. accuracy. Partitioned arrays allow higher accuracy in larger CPAs. (c) Recognition accuracy against the ratio R_W/R_{ON} for two different array sizes: 28 × 28 images (two 784 × 10 arrays, ∼15.6k devices, empty markers) and 8 × 8 images (two 64 × 10 arrays, ∼1.2k devices, filled markers). Resulting confusion matrix considering model play *C*2 and R_W=10 Ω for (d) 28 × 28, (e) 20 × 20 and (f) 8 × 8 px. image sizes, considering non-partitioned CPAs.

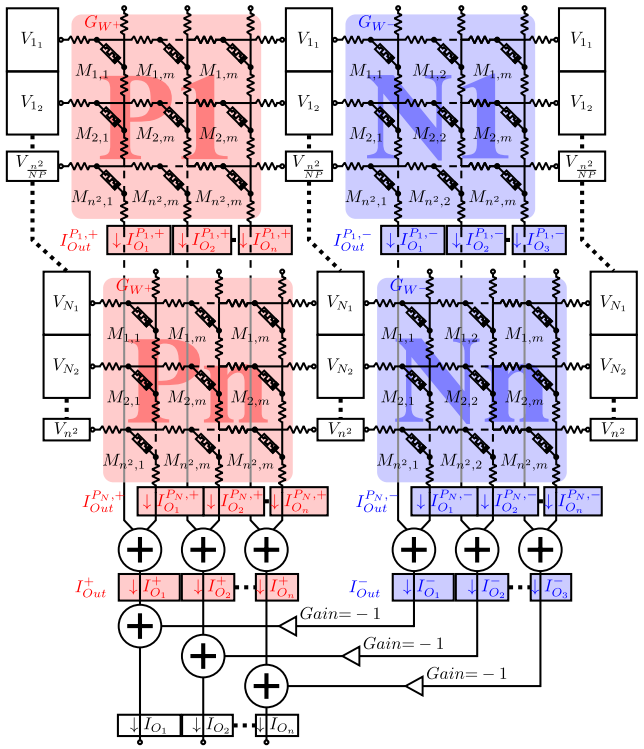


FIGURE 14. Simplified equivalent circuit schematic for a partitioned CPA based single layer perceptron. Each CPA is subdivided into N identically sized partitions to minimise the parasitic voltage drops. Partial output current vectors are indicated in the output of each partition.

the smallest CPA (1280 devices) simulated with model plays *C*1-*C*4 are plotted altogether with the data from Fig. 11b. This figure shows that the Accuracy–R_W/R_{ON} relationship follows a common trend right shifted as the CPA size decreases. In this way, the R_W/R_{ON} ratio ranges for which the CPA behaviour remains independent of the wire resistance extends up to one order of magnitude when changing the CPA size from ∼15.6 k deices (28 × 28 px. image) to ∼1.2 k devices (8 × 8 px image).

From these results, and as expected, it becomes clear that it is not efficient to implement large matrices using one single cross-point array. Given that both R_W and R_{ON}/R_{OFF} are normally defined by the selected fabrication node and RS mechanism, respectively, a widely accepted [73], [76] design alternative consists in dividing the large matrices into small partitions, whose reduced size improves their read margin. Fig. 14 shows the simplified circuit schematic of the partitioned CPAs and the interconnections required to realise the complete Matrix-Vector Multiplication. Exploding the integrability of the CPA with CMOS circuitry, vertical interconnects used to connect the outputs of the vertical CPA partitions may be placed under the partitioned structure, as well as the analogue sensing electronics, allowing the partitioned CPA to maintain a similar area consumption than the original non-partitioned case [73]. The vertical interconnects are grounded through the sensing circuit to absorb the currents within the same vertical wire. When compared against the original non-partitioned arrays, subdivided CPAs show a clear improvement in terms of classification accuracy, as shown in Fig. 13b for the three wire resistance values considered (1,10 and 100 Ω). For example, the inference accuracy for the 28 × 28 px. and $R_W = 100\ \Omega$ image increases from 46.33% to 72.63%. Similarly, pinpointed Accuracy-size maximums shift to the right for the partitioned arrays, being the corresponding values 3920, 5120, 11520 synapses for R_W equal to 100, 10 and 1 Ω, respectively. It is worth mentioning that for the other MNIST-like datasets considered, the same dependencies of the inference accuracy (in terms of partitioning scheme, CPA size and line resistance) were found as it is shown in the Supplementary Material (Supplementary Figs. 3 and 5), altogether with the sensitivity, specificity, precision, F1-score and κ-coefficient metrics.

D. DEVICE-TO-DEVICE AND PROGRAMMING VARIABILITY

The RS technology has been successfully demonstrated in amorphous and poly-crystalline materials. These materials have the advantage of low temperature deposition, so multiple matrix layers can be manufactured without disturbing the digital circuitry below. However, the uncontrolled high density

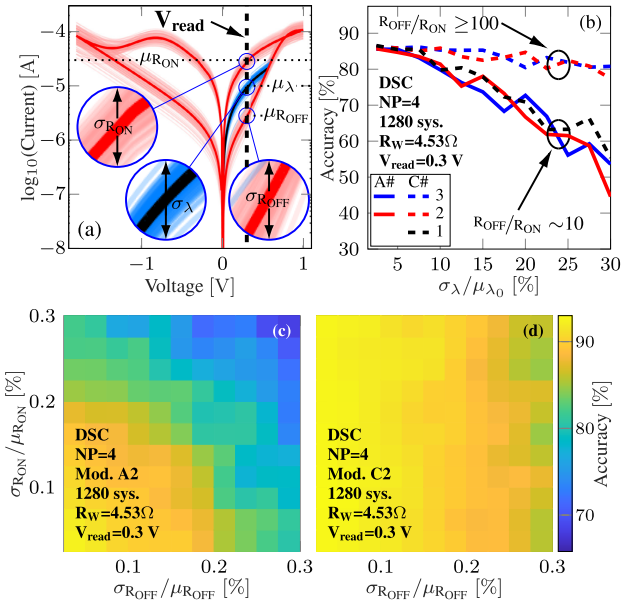


FIGURE 15. Impact of the device variability on the test accuracy. (a) Different sources of variability considered: R_{ON}, R_{OFF} and λ. (b) Impact of the λ variability on the accuracy. Model plays *A*1 and *C*1 are equivalent and thereby only one is considered (See. Fig 8a-8c). Inference accuracy is studied as function of the combined variability of R_{ON} and R_{OFF} for model plays *A*2 (c) and *C*2 (d) showing a similar behaviour to the one exhibit by (b).

of defects in addition to the intrinsic stochastic nature of the switching mechanism can induce a high degree of variability [89]. Along with the read margin reduction caused by the limited resistance window (R_{ON}/R_{OFF}) and interconnection line resistance (R_W), such variability in the memristive structures also increases the possibility of a deviation of the weighted sum from the target value [90]. Moreover, if each device performs slightly different and its characteristics are allowed to vary in time, programming to a desired state becomes a challenging task.

The normalised device variability is expressed as σ/μ, where σ is the standard deviation and μ the mean value of the ON and OFF resistance (R_{ON} and R_{OFF}) distributions. Variability can happen from cycle-to-cycle (intra-device) and from device-to-device (inter-device). The variability of the resistance states R_{ON} and R_{OFF} across a matrix is largely influenced by the choice of the stack's materials (e.g., single material HfO_X vs. bilayer HfO_X+TaO_X) [91], [92], as well as the device scaling. Extreme scaling seems to reduce the variability, probably because of a reduction of the area where the switching occurs [93]. Furthermore, the conductance value set during the Write-Verify procedure also presents variability [56], [90]. As in this article we exclusively focus on the inference phase, in this Sub-Section we show that the QMM is suitable to study the device variability in terms of the R_{ON} and R_{OFF} dispersion, as well as due to errors in Ω the conductance programming, expressed by the variability of the QMM control parameter λ (see Fig. 15a). In this short study, different model plays are considered to compare their susceptibilities to device-to-device variations. For brevity, only the MNIST dataset is considered.

In Fig. 15b the influence of the control parameter λ variability ($\sigma_\lambda/\mu_\lambda$ ranging from 0 to 30%) over the inference accuracy is studied for different model plays (*A*1-*A*3, *C*1-*C*3) and a wire resistance corresponding to a 16nm technology node (R_W=4.5 Ω). No variability in the major *I*-*V* loop is considered ($\sigma_{R_{OFF}} = \sigma_{R_{ON}}$=0). Two trends are clearly observed. On one hand, model plays having an R_{OFF}/R_{ON} equal to or greater than 100 (*C*2 and *C*3) exhibit a very reduced sensitivity to λ variations (accuracy loss is below 5% for variabilities up to 30%). On the other hand, there is a sustained accuracy reduction for model plays *A*1-*A*3 (*A*1 and *C*1 are equivalent) over the same range of $\sigma_\lambda/\mu_\lambda$. Both cases were then more thoroughly studied by considering the joint variability of R_{ON} and R_{OFF} ($\sigma_{R_{ON}}/\mu_{R_{ON}}$ and $\sigma_{R_{OFF}}/\mu_{R_{OFF}}$ respectively). As the variability is normally higher in HRS than in LRS [61], they were swept independently, resulting in the accuracy maps illustrated in Figs. 15c and 15d for model plays *A*2 and *C*2, respectively. The trend observed in Fig. 15b is repeated among Figs. 15c and 15d with a reduced sensitivity to the R_{ON} and R_{OFF} variations for model play *C*2 (higher R_{OFF}/R_{ON} ratio). Interestingly, for the case *C*2, $\sigma_{R_{OFF}}/\mu_{R_{OFF}}$ has a higher impact on the inference accuracy, likely due to a higher number of memdiodes mapped close to the R_{OFF} value.

The clear differences in the sensitivity of the inference accuracy to the device variability between model plays with varying R_{OFF}/R_{ON} ratios shows that the QMM is also able to deal with the device variability within the SPICE simulation framework. In addition, the obtained results indicate that apart from the high R_{ON} value required to minimise the parasitic voltage drop that limits the read margins, a high resistance window is necessary to improve both the tolerance to errors in the mapping of the CPA conductances and the device variability ($\sigma_{R_{ON}}/\mu_{R_{ON}}$ and $\sigma_{R_{OFF}}/\mu_{R_{OFF}}$).

E. COMPUTATIONAL COMPLEXITY COMPARED TO OTHER MEMRISTOR MODELS

In this article we focus on the inference phase of the CPA-based Neural Network, considering *ex-situ* training. In this phase, the time complexity for the vector-matrix multiplication performed on each input vector is $O(1)$ (defined in terms of the $big-O$ notation [98]), regardless of the RRAM-algorithm considered (RRAM model) or learning algorithm. The total accuracy obtained with the memdiode-based CPA is shown to meet that one obtained with an ideal ANN implemented in MATLAB, when the wire parasitics becomes negligible as shown in Fig. 16, for all the datasets considered (the CIFAR-10 [94] and SVHN [95] datasets were included for the sake of comprehensiveness). In fact, for the cases with $R_W \leq 1$ the differences between the accuracy of the idealised network and the QMM implementation are not statistically significant [99]. Note that although poor, the classification results for the CIFAR-10 and SVHN results obtained with an idealised single-layer perceptron are

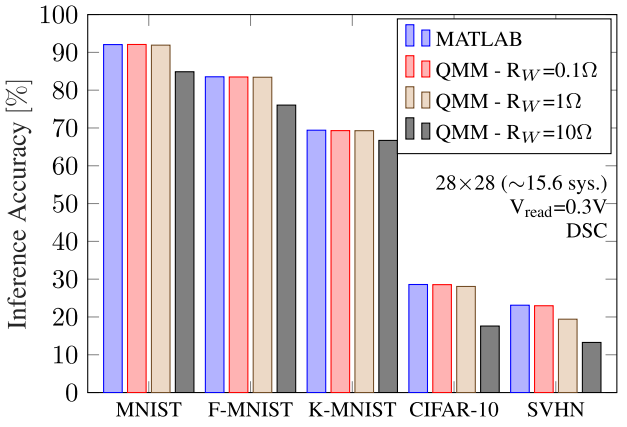


FIGURE 16. Performance metrics of the memdiode-based CPA simulated with the QMM, compared against the ideal Software case computed in MATLAB. 5 different databases are considered: MNIST [78], MNIST-F [57], MNIST-K [58], CIFAR-10 [94] and SVHN [95]. Each CPA is divided into 4 partitions to further reduced the impact of R_W. Note that as R_W tends to 0, the CPA mimics the MATLAB results.

matched by the memdiode-based CPA classifier. Therefore, the limitation is not in the CPA implementation but in the Neural Network itself, i.e. for this kind of datasets more complex networks are required, as shown in [90], [100], where a Deep Neural Network comprising 6 Convolutional Layers and 2 Fully Connected layers are considered. However, in both studies no parasitic effects taken into account.

Nevertheless, studying both the time complexity and space complexity of the resulting SPICE code for different RRAM models is a powerful metric to compare their performance in terms of computational complexity in electrical simulation platforms. As an analytic determination of the time complexity is not plausible, we have opted for empirically measuring the running time and RAM memory usage for the SPICE simulation of CPA circuits involving 320 to 15680 memristive devices. To minimise errors induced by the hardware where the simulations are performed, we have carried out multiple simulations for each case and reported the mean values. Additionally, other 4 different RRAM models proposed in the literature (Yakopcic model [34], Laiho-Biolek model [96], the University of Michigan Model [97] and a linear model [15], i.e. a fixed resistance defined upon the corresponding synaptic value) were considered apart from the QMM model for comparison purposes. The time and memory requirements are shown in Figs. 17a and 17b, respectively. It can be seen, that regardless of the model considered the CPU usage increases proportionally to the square of the number of devices suggesting an $O(n^2)$ time complexity, while the RAM usage increases linearly with the number of devices, indicating an $O(n)$ space complexity.

Further details regarding the comparison are provided for two CPA sizes (1280 and 15680 devices) and two different circuit simulators (HSPCE and FineSim -A Fast SPICE circuit simulator-). Both the run-time and RAM memory usage are reported in Figs. 17c and 17d, respectively, normalised respect to the metrics from the linear model case (simplest

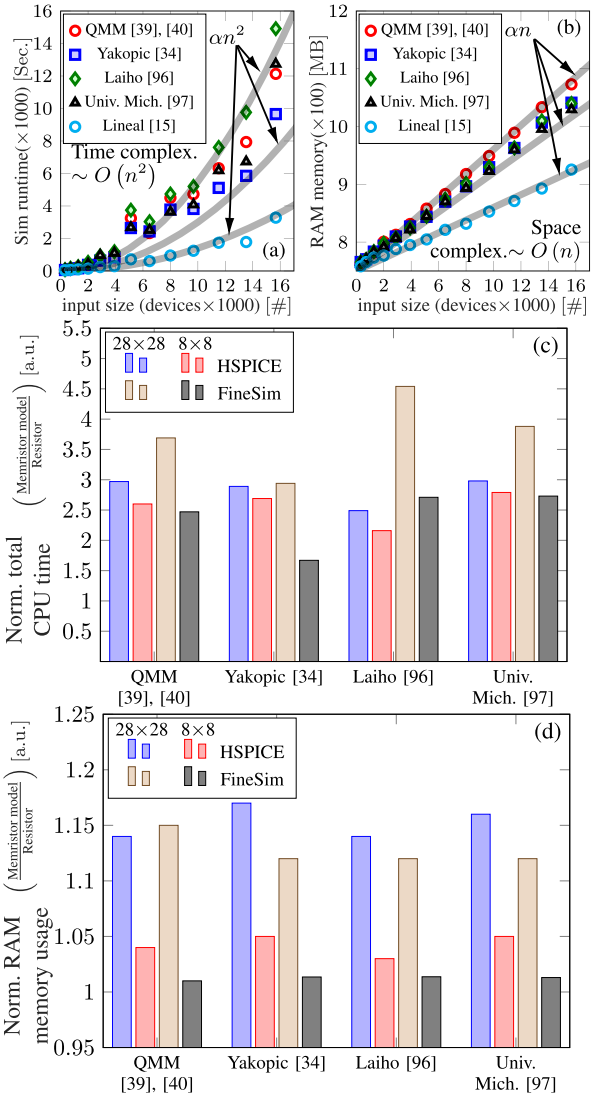


FIGURE 17. The computational cost (run-time and RAM memory usage) of the memdiode-based CPA implemented with the QMM model is compared against other memristor models reported by Yakopcic *et al.* [34], Laiho *et al.* [96] and the University of Michigan [97]. (a) Simulation run-time as function of the CPA size (in number of devices) increases quadratically. (b) Simulation RAM memory usage increases linearly with the CPA size. (c) Total CPU time normalised against the case of a resistor of fixed value to represent the programmed memristor. (d) Total RAM memory usage normalised against the resistor case (linear model). In both cases simulations were performed both in HSPICE and Fast-SPICE (FineSim) environments and for different image sizes.

possible model), for each simulator-CPA size scenario. From the comparison among the QMM and other memristor models, considering both the normalised CPU and RAM memory usage, we conclude that the QMM is capable of providing a very accurate fit to the experimental I-V loops as shown in Section II without increasing the time and space complexity in the simulation of large circuits.

V. CONCLUSION

In this article, we have demonstrated the viability of the Quasi-Static memdiode compact model (QMM) for realistic SPICE simulations of large (up to ∼15.6 k synapses)

RRAM-based cross-point arrays (CPA) intended for neuromorphic applications. A single layer perceptron and different datasets (including the MNIST database of greyscale, handwritten digits) were considered as case study. Although a simplistic approach when compared with more sophisticated multi-layer Artificial Neural Networks (ANN), the single layer perceptron allows studying and clarifying the ANN limitations caused by parasitic effects and non-idealities occurring in the synaptic layers implemented with CPAs, as well as benchmarking the computational costs of the QMM based simulations against other available models.

Apart from its versatility, which allows it to accurately fit the experimental results from multiple devices, the proposed model provides a number of intrinsic features which are quite helpful in connection with the Matrix-Vector Multiplication (MVM) method. First, the unified expression for the HRS and LRS I-V characteristics simplifies the conductance mapping stage: once the synaptic weights are obtained by an *ex-situ* training procedure, they can be translated into the G_{HRS}-G_{LRS} range by tuning one single control parameter (λ) in the memdiode model. Second, as the mapped conductance runs from G_{HRS} to G_{LRS} (for λ ranging between 0 and 1) the shape of the I-V curve progressively changes from linear-exponential to linear. In this way, the memdiode model allows accounting for both the non-linear behaviour in HRS and the linear I-V characteristic in LRS. Capturing the commonly exponential dependence of the high resistance state is relevant as the reduction in the read margins causes different voltages to be applied to the CPA memristors depending on their spatial location. Last but not least, the proposed model admits a simple sub-circuit description independent of the simulation timestep parameter. This is of utmost importance for circuit simulations in which the time evolution of the system is under the control of the simulator itself and not in hands of the user. Consequently, the QMM provides high fitting accuracy of the device experimental results without increasing the computational complexity (run-time and RAM memory usage) with regard to other memristor models.

An exploratory analysis of the main features governing the CPA performance was conducted considering the classification accuracy of MNIST-like datasets. Different I-V loops were generated with the memdiode model to account for the variety of resistance windows (R_{OFF}/R_{ON}, or equivalently G_{LRS}/G_{HRS} ratio) reported in the literature. Simulation results for different wire resistances (R_W) calculated for different technological nodes are in line with previous analytical results showing that the R_W/R_{ON} ratio plays a major role in the read margin of the CPA devices and thereby in the classification accuracy. As R_W is expected to maintain an increasing trend in the upcoming technology nodes, RRAM devices with improved ON resistance and reasonable resistance windows (to reduce the sensitivity to device variability) are mandatory to minimise the parasitic voltage drops. However, higher values of R_{ON} (and thereby lower power operation) comes at the cost of degrading the Signal-to-Noise Ratio (SNR), which also limits the inference accuracy, suggesting a trade-off between niminising the parasitic voltage-drop and increasing the SNR. The increase in R_W also causes a sensitive increment in the inference latency, which severely limits the high frequency operation of highly-scaled CPAs. In addition, for a given read margin/classification accuracy we have shown that the R_W/R_{ON} ratio also limits the maximum size of the realisable CPA. In this regard, partitioning techniques are mandatory for large size CPAs to maintain an acceptable classification accuracy. It is also worth mentioning that the usual downscaling procedure to match the input database with the size limitation of the CPA should be carried out having in mind the possible classification accuracy reduction arising from the readability loss of the input images. In this regard, an optimum CPA size is shown for the case of the MNIST-like datasets classification using the single layer perceptron.

REFERENCES

[1] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1864–1878, Dec. 2014, doi: 10.1109/TNNLS.2013.2296777.

[2] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S.-P. Wong, "A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation," *Adv. Mater.*, vol. 25, no. 12, pp. 1774–1779, Mar. 2013, doi: 10.1002/adma.201203680.

[3] R. F. Freitas and W. W. Wilcke, "Storage-class memory: The next storage system technology," *IBM J. Res. Develop.*, vol. 52, no. 4.5, pp. 439–447, Jul. 2008, doi: 10.1147/rd.524.0439.

[4] N. K. Upadhyay, S. Joshi, and J. J. Yang, "Synaptic electronics and neuromorphic computing," *Sci. China Inf. Sci.*, vol. 59, no. 6, Jun. 2016, Art. no. 061404, doi: 10.1007/s11432-016-5565-1.

[5] Y. Sasago, M. Kinoshita, T. Morikawa, K. Kurotsuchi, S. Hanzawa, T. Mine, A. Shima, Y. Fujisaki, H. Kume, H. Moriya, N. Takaura, and K. Torii, "Cross-point phase change memory with $4f^2$ cell size driven by low-contact-resistivity poly-si diode," in *Dig. Tech. Papers Symp. VLSI Technol.*, Jul. 2009, pp. 24–25.

[6] S. N. Truong and K.-S. Min, "New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing," *JSTS, J. Semicond. Technol. Sci.*, vol. 14, no. 3, pp. 356–363, Jun. 2014, doi: 10.5573/JSTS.2014.14.3.356.

[7] S. Truong, S.-J. Ham, and K.-S. Min, "Neuromorphic crossbar circuit with nanoscale filamentary-switching binary memristors for speech recognition," *Nanosc. Res. Lett.*, vol. 9, no. 1, p. 629, 2014, doi: 10.1186/1556-276X-9-629.

[8] S. N. Truong, S. Shin, S.-D. Byeon, J. Song, and K.-S. Min, "New twin crossbar architecture of binary memristors for low-power image recognition with discrete cosine transform," *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 1104–1111, Nov. 2015, doi: 10.1109/TNANO.2015.2473666.

[9] S. Park, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *IEDM Tech. Dig.*, Dec. 2012, pp. 10–12, doi: 10.1109/IEDM.2012.6479016.

[10] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, Nov. 2014, pp. 63–70, doi: 10.1109/ICCAD.2014.7001330.

[11] S.-J. Ham, H.-S. Mo, and K.-S. Min, "Low-power V_{DD}/3 write scheme with inversion coding circuit for complementary memristor array," *IEEE Trans. Nanotechnol.*, vol. 12, no. 5, pp. 851–857, Sep. 2013, doi: 10.1109/TNANO.2013.2274529.

[12] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, May 2008, doi: 10.1038/nature06932.

[13] C. Chen, S. Gao, G. Tang, H. Fu, G. Wang, C. Song, F. Zeng, and F. Pan, "Effect of electrode materials on AlN-based bipolar and complementary resistive switching," *ACS Appl. Mater. Interface*, vol. 5, no. 5, pp. 1793–1799, Mar. 2013, doi: 10.1021/am303128h.

[14] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Analogue signal and image processing with large memristor crossbars," *Nature Electron.*, vol. 1, no. 1, pp. 52–59, Jan. 2018, doi: 10.1038/s41928-017-0002-z.

[15] Y. K. Lee, J. W. Jeon, E.-S. Park, C. Yoo, W. Kim, M. Ha, and C. S. Hwang, "Matrix mapping on crossbar memory arrays with resistive interconnects and its use in in-memory compression of biosignals," *Micromachines*, vol. 10, no. 5, p. 306, May 2019, doi: 10.3390/mi10050306.

[16] R. Han, P. Huang, Y. Zhao, X. Cui, X. Liu, and J. Kang, "Efficient evaluation model including interconnect resistance effect for large scale RRAM crossbar array matrix computing," *Sci. China Inf. Sci.*, vol. 62, no. 2, pp. 1–11, Feb. 2019, doi: 10.1007/s11432-018-9555-8.

[17] B. Zhang, N. Uysal, D. Fan, and R. Ewetz, "Handling stuck-at-faults in Memristor Crossbar Arrays using Matrix Transformations," in *Proc. Asia South Pacific Des. Automat. Conf.*, New York, NY, USA, Jan. 2019, pp. 474–479, doi: 10.1145/3287624.3287707.

[18] B. Zhang, N. Uysal, D. Fan, and R. Ewetz, "Handling stuck-at-fault defects using matrix transformation for robust inference of DNNs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 10, pp. 2448–2460, Oct. 2020, doi: 10.1109/tcad.2019.2944582.

[19] L. Xia, M. Liu, X. Ning, K. Chakrabarty, and Y. Wang, "Fault-tolerant training with on-line fault detection for RRAM-based neural computing systems," in *Proc. Des. Autom. Conf.*, vol. 12828, Jun. 2017, pp. 1–6, doi: 10.1145/3061639.3062248.

[20] C. Liu, M. Hu, J. P. Strachan, and H. Li, "Rescuing memristor-based neuromorphic design with high defects," in *Proc. 54th Annu. Design Autom. Conf.*, Jun. 2017, pp. 1–6, doi: 10.1145/3061639.3062310.

[21] A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell, and A. J. Kenyon, "Simulation of inference accuracy using realistic RRAM devices," *Frontiers Neurosci.*, vol. 13, pp. 1–15, Jun. 2019, doi: 10.3389/fnins.2019.00593.

[22] C. Lammie, S. Member, W. Xiang, and S. Member, "MemTorch : An open-source simulation framework for memristive deep learning systems," 2020, *arXiv:2004.10971*. [Online]. Available: https://arxiv.org/abs/2004.10971

[23] C. Yakopcic, R. Hasan, T. M. Taha, M. R. McLean, and D. Palmer, "Efficacy of memristive crossbars for neuromorphic processors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 15–20, doi: 10.1109/IJCNN.2014.6889807.

[24] H. Yu and Y. Wang, "Nonvolatile state identification and NVM spice," in *Design Exploration of Emerging Nano-scale Non-volatile Memory*. Springer, 2014, pp. 45–83, doi: 10.1007/978-1-4939-0551-5_3.

[25] C. Yakopcic, T. M. Taha, G. Subramanyam, and R. E. Pino, "Memristor SPICE modeling," in *Advances in Neuromorphic Memristor Science and Applications*. Amsterdam, The Netherlands: Springer, 2012, pp. 211–244, doi: 10.1007/978-94-007-4491-2_12.

[26] D. Panda, P. P. Sahu, and T. Y. Tseng, "A collective study on modeling and simulation of resistive random access memory," *Nanosc. Res. Lett.*, vol. 13, no. 1, Dec. 2018, doi: 10.1186/s11671-017-2419-8.

[27] J. Zha, H. Huang, T. Huang, J. Cao, A. Alsaedi, and F. E. Alsaadi, "A general memristor model and its applications in programmable analog circuits," *Neurocomputing*, vol. 267, pp. 134–140, Dec. 2017, doi: 10.1016/j.neucom.2017.04.057.

[28] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, "Device and SPICE modeling of RRAM devices," *Nanoscale*, vol. 3, no. 9, pp. 3833–3840, 2011, doi: 10.1039/c1nr10557d.

[29] Y. N. Joglekar and S. J. Wolf, "The elusive memristor: Properties of basic electrical circuits," *Eur. J. Phys.*, vol. 30, no. 4, pp. 661–675, Jul. 2009, doi: 10.1088/0143-0807/30/4/001.

[30] T. Prodromakis, B. P. Peh, C. Papavassiliou, and C. Toumazou, "A versatile memristor model with nonlinear dopant kinetics," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 3099–3105, Sep. 2011, doi: 10.1109/TED.2011.2158004.

[31] M. D. Pickett, D. B. Strukov, J. L. Borghetti, J. J. Yang, G. S. Snider, D. R. Stewart, and R. S. Williams, "Switching dynamics in titanium dioxide memristive devices," *J. Appl. Phys.*, vol. 106, no. 7, Oct. 2009, Art. no. 074508, doi: 10.1063/1.3236506.

[32] F. Merrikh Bayat, B. Hoskins, and D. B. Strukov, "Phenomenological modeling of memristive devices," *Appl. Phys. A*, vol. 118, no. 3, pp. 779–786, Mar. 2015, doi: 10.1007/s00339-015-8993-7.

[33] D. Biolek, Z. Biolek, V. Biolkova, and Z. Kolka, "Modeling of TiO_2 memristor: From analytic to numerical analyses," *Semicond. Sci. Technol.*, vol. 29, no. 12, pp. 2–7, 2014, doi: 10.1088/0268-1242/29/12/125008.

[34] C. Yakopcic, T. M. Taha, G. Subramanyam, and R. E. Pino, "Generalized memristive device SPICE model and its application in circuit design," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 8, pp. 1201–1214, Aug. 2013, doi: 10.1109/TCAD.2013.2252057.

[35] S. Kvatinsky, E. G. Friedman, A. Kolodny, and U. C. Weiser, "TEAM: ThrEshold adaptive memristor model," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 1, pp. 211–221, Jan. 2013, doi: 10.1109/TCSI.2012.2215714.

[36] S. Kvatinsky, M. Ramadan, E. G. Friedman, and A. Kolodny, "VTEAM: A general model for voltage-controlled memristors," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 8, pp. 786–790, Aug. 2015, doi: 10.1109/TCSII.2015.2433536.

[37] K. Eshraghian, O. Kavehei, K.-R. Cho, J. M. Chappell, A. Iqbal, S. F. Al-Sarawi, and D. Abbott, "Memristive device fundamentals and modeling: Applications to circuits and systems simulation," *Proc. IEEE*, vol. 100, no. 6, pp. 1991–2007, Jun. 2012, doi: 10.1109/JPROC.2012.2188770.

[38] Z. Biolek, D. Biolek, V. Biolkova, and Z. Kolka, "Reliable modeling of ideal generic memristors via state-space transformation," *Radioengineering*, vol. 24, no. 2, pp. 393–407, Jun. 2015, doi: 10.13164/re.2015.0393.

[39] E. Miranda, "Compact model for the major and minor hysteretic I–V loops in nonlinear memristive devices," *IEEE Trans. Nanotechnol.*, vol. 14, no. 5, pp. 787–789, Sep. 2015, doi: 10.1109/TNANO.2015.2455235.

[40] G. A. Patterson, J. Sune, and E. Miranda, "Voltage-driven hysteresis model for resistive switching: SPICE modeling and circuit applications," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 12, pp. 2044–2051, Dec. 2017, doi: 10.1109/TCAD.2017.2756561.

[41] S. Petzold, E. Miranda, S. U. Sharath, J. Muñoz-Gorriz, T. Vogel, E. Piros, N. Kaiser, R. Eilhardt, A. Zintler, L. Molina-Luna, J. Suñé, and L. Alff, "Analysis and simulation of the multiple resistive switching modes occurring in HfOx-based resistive random access memories using memdiodes," *J. Appl. Phys.*, vol. 125, no. 23, Jun. 2019, Art. no. 234503, doi: 10.1063/1.5094864.

[42] Y. LeCun, C. Cortes, and C. J. Burges. (1998). *Mnist Handwritten Digit Database*. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[43] J. Blasco, P. Jančovič, K. Fröhlich, J. Su né, and E. Miranda, "Modeling of the switching I-V characteristics in ultrathin (5 nm) atomic layer deposited HfO_2 films using the logistic hysteron," *J. Vac. Sci. Technol. B, Microelectron.*, vol. 33, no. 1, Jan. 2015, Art. no. 01A102, doi: 10.1116/1.4900599.

[44] C.-Y. Lin, C.-Y. Wu, C.-Y. Wu, C. Hu, and T.-Y. Tseng, "Bistable resistive switching in Al[sub2]O[sub3] memory thin films," *J. Electrochem. Soc.*, vol. 154, no. 9, p. G189, 2007, doi: 10.1149/1.2750450.

[45] M. K. Yang, J.-W. Park, T. K. Ko, and J.-K. Lee, "Resistive switching characteristics of TiN/MnO_2/Pt memory devices," *Phys. Status Solidi (RRL) Rapid Res. Lett.*, vol. 4, nos. 8–9, pp. 233–235, Sep. 2010, doi: 10.1002/pssr.201004213.

[46] *Resistance Random Access Memory (RRAM)*. Accessed: Sep. 30, 2010. [Online]. Available: http://archive.today/c6PS

[47] E. Miranda, W. Román Acevedo, D. Rubi, U. Läders, P. Granell, J. Suñé, and P. Levy, "Modeling of the multilevel conduction characteristics and fatigue profile of Ag/$La_1/_3Ca_2/_3MnO_3$/Pt structures using a compact memristive approach," *J. Appl. Phys.*, vol. 121, no. 20, May 2017, Art. no. 205302, doi: 10.1063/1.4984051.

[48] K. Fröhlich, I. Kundrata, M. Blaho, M. Precner, M. Åapajna, M. Klimo, O. Šuch, and O. Škvarek, "Hafnium oxide and tantalum oxide based resistive switching structures for realization of minimum and maximum functions," *J. Appl. Phys.*, vol. 124, no. 15, Oct. 2018, Art. no. 152109, doi: 10.1063/1.5025802.

[49] A. R. Lee, Y. C. Bae, H. S. Im, and J. P. Hong, "Complementary resistive switching mechanism in ti-based triple TiOx/TiN/TiOx and TiOx/TiOxNy/TiOx matrix," *Appl. Surf. Sci.*, vol. 274, pp. 85–88, Jun. 2013, doi: 10.1016/j.apsusc.2013.02.100.

[50] W. J. Duan, J. B. Wang, X. L. Zhong, H. J. Song, and B. Li, "Complementary resistive switching in single sandwich structure for crossbar memory arrays," *J. Appl. Phys.*, vol. 120, no. 8, Aug. 2016, Art. no. 084502, doi: 10.1063/1.4961222.

[51] M. Yang, H. Wang, X. Ma, H. Gao, and Y. Hao, "Voltage-amplitude-controlled complementary and self-compliance bipolar resistive switching of slender filaments in Pt/HfO_2/HfO_X/Pt memory devices," *J. Vac. Sci. Technol. B, Microelectron.*, vol. 35, no. 3, May 2017, Art. no. 032203, doi: 10.1116/1.4983193.

[52] A. Chen, "A comprehensive crossbar array model with solutions for line resistance and nonlinear device characteristics," *IEEE Trans. Electron Devices*, vol. 60, no. 4, pp. 1318–1326, Dec. 2013, doi: 10.1109/TED.2013.2246791.

[53] F. L. Aguirre, A. Rodriguez-Fernandez, S. M. Pazos, J. Sune, E. Miranda, and F. Palumbo, "Study on the connection between the set transient in RRAMs and the progressive breakdown of thin oxides," *IEEE Trans. Electron Devices*, vol. 66, no. 8, pp. 3349–3355, Aug. 2019, doi: 10.1109/ted.2019.2922555.

[54] W. Choi, K. Moon, M. Kwak, C. Sung, J. Lee, J. Song, J. Park, S. A. Chekol, and H. Hwang, "Hardware implementation of neural network using pre-programmed resistive device for pattern recognition," *Solid-State Electron.*, vol. 153, pp. 79–83, Mar. 2019, doi: 10.1016/j.sse.2018.12.018.

[55] A. Mehonic, A. L. Shluger, D. Gao, I. Valov, E. Miranda, D. Ielmini, A. Bricalli, E. Ambrosi, C. Li, J. J. Yang, Q. Xia, and A. J. Kenyon, "Silicon oxide (SiO_X): A promising material for resistance switching?" *Adv. Mater.*, vol. 30, no. 43, pp. 1–21, 2018, doi: 10.1002/adma.201801187.

[56] V. Milo, C. Zambelli, P. Olivo, E. Pérez, M. K. Mahadevaiah, O. G. Ossorio, C. Wenger, and D. Ielmini, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Mater.*, vol. 7, no. 8, Aug. 2019, Art. no. 081120, doi: 10.1063/1.5108650.

[57] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*. [Online]. Available: https://arxiv.org/abs/1708.07747

[58] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," 2018, *arXiv:1812.01718*. [Online]. Available: https://arxiv.org/abs/1812.01718

[59] G. W. Burr, R. M. Shelby, C. Di Nolfo, J. W. Jang, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," in *IEDM Tech. Dig.*, Feb. 2015, pp. 29.5.1–29.5.4, doi: 10.1109/IEDM.2014.7047135.

[60] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks," *Nature Commun.*, vol. 9, no. 1, pp. 1–8, Dec. 2018, doi: 10.1038/s41467-018-04484-2.

[61] Z. Dong, Z. Zhou, Z. Li, C. Liu, P. Huang, L. Liu, X. Liu, and J. Kang, "Convolutional neural networks based on RRAM devices for image recognition and online learning tasks," *IEEE Trans. Electron Devices*, vol. 66, no. 1, pp. 793–801, Jan. 2019, doi: 10.1109/TED.2018.2882779.

[62] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, "Immunity to device variations in a spiking neural network with memristive nanodevices," *IEEE Trans. Nanotechnol.*, vol. 12, no. 3, pp. 288–295, May 2013, doi: 10.1109/TNANO.2013.2250995.

[63] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, no. 4, pp. 525–533, Nov. 1993, doi: 10.1016/S0893-6080(05)80056-5.

[64] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. for Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, Jun. 1963, doi: 10.1137/0111030.

[65] R. Battiti, "First- and second-order methods for learning: Between steepest descent and Newton's method," *Neural Comput.*, vol. 4, no. 2, pp. 141–166, Mar. 1992, doi: 10.1162/neco.1992.4.2.141.

[66] M. J. D. Powell, "Restart procedures for the conjugate gradient method," *Math. Program.*, vol. 12, no. 1, pp. 241–254, Dec. 1977, doi: 10.1007/BF01593790.

[67] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1996, doi: 10.1137/1.9781611971200.

[68] R. Fletcher, "Function minimization by conjugate gradients," *Comput. J.*, vol. 7, no. 2, pp. 149–154, Feb. 1964, doi: 10.1093/comjnl/7.2.149.

[69] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, pp. 586–591, doi: 10.1109/icnn.1993.298623.

[70] M. Hagan and H. Demuth. (2014). *Neural Network Design*. [Online]. Available: http://scholar.google.com/scholar?hl=en&sugexp=gsih&pq=badminton+% training&xhr=t&q=neural+network+design&cp=16&qe=bmV1cmFsIG5ldHdvcm% sgZA&qesig=k9_6OUSCnOMtzbLRV7Bxag&pkc=AFgZ2tnXZWxdMzsRdm7bQlAZ9Ouzahw% -8F-IapNFqR9QR-QaxLC0MI5wrX4F_gMeseaytVjRLbReEMmERZgMf

[71] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, May 2015, doi: 10.1038/nature14441.

[72] M. Hu, H. Li, Q. Wu, G. S. Rose, and Y. Chen, "Memristor crossbar based hardware realization of BSB recall function," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–7, doi: 10.1109/IJCNN.2012.6252563.

[73] M. E. Fouda, S. Lee, J. Lee, A. Eltawil, and F. Kurdahi, "Mask technique for fast and efficient training of binary resistive crossbar arrays," *IEEE Trans. Nanotechnol.*, vol. 18, no. 5, pp. 704–716, Dec. 2019, doi: 10.1109/tnano.2019.2927493.

[74] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing," in *Proc. 53rd Annu. Design Autom. Conf.*, New York, NY, USA, 2016, pp. 1–6, doi: 10.1145/2897937.2898010.

[75] Y. Shi, L. Nguyen, S. Oh, X. Liu, F. Koushan, J. R. Jameson, and D. Kuzum, "Neuroinspired unsupervised learning and pruning with subquantum CBRAM arrays," *Nature Commun.*, vol. 9, no. 1, pp. 1–11, Dec. 2018, doi: 10.1038/s41467-018-07682-0.

[76] J. Liang, S. Yeh, S. S. Wong, and H.-S.-P. Wong, "Effect of Wordline/Bitline scaling on the performance, energy consumption, and reliability of cross-point memory array," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 1, pp. 1–14, Feb. 2013, doi: 10.1145/2422094.2422103.

[77] D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling," *Semicond. Sci. Technol.*, vol. 31, no. 6, Jun. 2016, Art. no. 063002, doi: 10.1088/0268-1242/31/6/063002.

[78] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998.

[79] J. Liang and H.-S.-P. Wong, "Cross-point memory array without cell Selectors—Device characteristics and data storage pattern dependencies," *IEEE Trans. Electron Devices*, vol. 57, no. 10, pp. 2531–2538, Oct. 2010, doi: 10.1109/TED.2010.2062187.

[80] C. Liu, Q. Yang, B. Yan, J. Yang, X. Du, W. Zhu, H. Jiang, Q. Wu, M. Barnell, and H. H. Li, "A memristor crossbar based computing engine optimized for high speed and accuracy," *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Sep. 2016, pp. 110–115, 2016, doi: 10.1109/ISVLSI.2016.46.

[81] X. Lian, M. Lanza, A. Rodriguez, E. Miranda, and J. Sune, "On the properties of conducting filament in ReRAM," in *Proc. 12th IEEE Int. Conf. Solid-State Integr. Circuit Technol. (ICSICT)*, Oct. 2014, pp. –5, doi: 10.1109/ICSICT.2014.7021484.

[82] S. M. Rossnagel and T. S. Kuan, "Alteration of Cu conductivity in the size effect regime," *J. Vac. Sci. Technol. B, Microelectron.*, vol. 22, no. 1, pp. 240–247, 2004, doi: 10.1116/1.1642639.

[83] D. Josell, S. H. Brongersma, and Z. Tőkei, "Size-dependent resistivity in nanoscale interconnects," *Annu. Rev. Mater. Res.*, vol. 39, no. 1, pp. 231–254, Aug. 2009, doi: 10.1146/annurev-matsci-082908-145415.

[84] W. Steinhögl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100 nm and smaller," *J. Appl. Phys.*, vol. 97, no. 2, Jan. 2005, Art. no. 023706, doi: 10.1063/1.1834982.

[85] K. Fuchs, "The conductivity of thin metallic films according to the electron theory of metals," *Math. Proc. Cambridge Phil. Soc.*, vol. 34, no. 1, pp. 100–108, Jan. 1938, doi: 10.1017/S0305004100019952.

[86] A. F. Mayadas and M. Shatzkes, "Electrical-resistivity model for polycrystalline films: The case of arbitrary reflection at external surfaces," *Phys. Rev. B, Condens. Matter*, vol. 1, no. 4, pp. 1382–1389, Feb. 1970, doi: 10.1103/PhysRevB.1.1382.

[87] J. D. Meindl, "Interconnect opportunities for gigascale integration," *IEEE Micro*, vol. 23, no. 3, pp. 28–35, May 2003, doi: 10.1109/MM.2003.1209464.

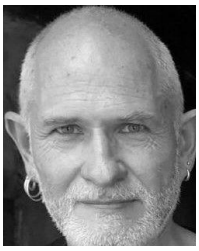
[88] T. Sakurai and K. Tamaru, "Simple formulas for two- and three-dimensional capacitances," *IEEE Trans. Electron Devices*, vol. 30, no. 2, pp. 183–185, Feb. 1983, doi: 10.1109/T-ED.1983.21093.

[89] G. C. Adam, A. Khiat, and T. Prodromakis, "Challenges hindering memristive neuromorphic hardware from going mainstream," *Nature Commun.*, vol. 9, no. 1, pp. 1–4, Dec. 2018, doi: 10.1038/s41467-018-07565-4.
[90] W. Yi, Y. Kim, and J.-J. Kim, "Effect of device variation on mapping binary neural network to memristor crossbar array," in *Proc. Des., Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 320–323, doi: 10.23919/DATE.2019.8714817.
[91] A. Chen and M.-R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," in *Proc. Int. Rel. Phys. Symp.*, Apr. 2011, p. 7, doi: 10.1109/IRPS.2011.5784590.
[92] Q. Luo, X. Xu, T. Gong, H. Lv, D. Dong, H. Ma, P. Yuan, J. Gao, J. Liu, Z. Yu, J. Li, S. Long, Q. Liu, and M. Liu, "8-layers 3D vertical RRAM with excellent scalability towards storage class memory applications," in *IEDM Tech. Dig.*, Dec. 2017, pp. 2–7, doi: 10.1109/IEDM.2017.8268315.
[93] S. Pi, P. Lin, and Q. Xia, "Cross point arrays of 8 nm × 8 nm memristive devices fabricated with nanoimprint lithography," *J. Vac. Sci. Technol. B, Microelectron.*, vol. 31, no. 6, 2013, Art. no. 06FA02, doi: 10.1116/1.4827021.
[94] A. Krizhevsky. (2009). *Learning Multiple Layers of Features from Tiny Images*. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html
[95] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS*, 2011, pp. 1–8.
[96] M. Laiho, E. Lehtonen, A. Russell, and P. Dudek, "Memristive synapses are becoming reality," in *Proc. Neuromorphic Eng.*, 2010, pp. 10–12, doi: 10.2417/1201011.003396.
[97] T. Chang, S.-H. Jo, K.-H. Kim, P. Sheridan, S. Gaba, and W. Lu, "Synaptic behaviors and modeling of a metal oxide memristive device," *Appl. Phys. A, Solids Surf.*, vol. 102, no. 4, pp. 857–863, Mar. 2011, doi: 10.1007/s00339-011-6296-1.
[98] D. Z. Du and K. I. Ko, *Theory Computing Complexity*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014, doi: 10.1002/9781118595091.
[99] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.
[100] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018, doi: 10.1109/JPROC.2018.2790840.

FERNANDO LEONEL AGUIRRE (Member, IEEE) received the degree in electronics engineering from the Universidad Tecnológica Nacional, Buenos Aires, Argentina (UTN-FRBA), in 2016, having done part of his studies in Germany thanks to the DAAD grant. He is currently pursuing the Ph.D. degree with UTN-FRBA/CONICET, working on the field of new materials and CMOS circuits reliability, under a full Ph.D. grant from CONICET. He has made a Ph.D. stay at the Universitat Autònoma de Barcelona (UAB) in the field of memristor model applications. He is currently a Teaching Assistant and an Interim Adjunct Professor with UTN.BA, teaching analogue IC design. He has coauthored the 2019 IEEE-IRPS Best Paper Award winning article, as well as other awarded conference papers. He has (co)authored several articles in the field of electron devices published in both international journals and conferences. His research interests include CMOS IC design, neuromorphic computing, semiconductor reliability, and electrical characterization.

SEBASTIÁN MATÍAS PAZOS (Member, IEEE) received the degree in electronic engineering from the Universidad Tecnológica Nacional, Buenos Aires, Argentina, in 2016, where he is currently pursuing the Ph.D. degree with UIDI-CONICET, working on the field of novel materials and RF CMOS circuit reliability, with a full Ph.D. grant from CONICET. He is currently a TA and an Interim Adjunct Professor with UTN.BA, teaching electron devices and analog IC designing. He has coauthored articles in several internationally recognized journals. His research interests include CMOS IC design, RF circuits, semiconductor device and circuit reliability, and electrical characterization. He was a recipient of the DAAD Grants for Undergrad abroad exchanges.

FÉLIX PALUMBO (Member, IEEE) received the M.Sc. and Ph.D. degrees in physics from the University of Buenos Aires, Argentina, in 2000 and 2005, respectively. He is currently an Active Researcher in the field of semiconductor device physics and reliability, with experience in the academy and industry. From 2001 to 2002, he was with the Research and Development Group, Tower Semiconductor, Migdal HaEmek, Israel, making research on the reliability of gate oxides and flash memories. Since 2003, he has been a repeat visiting Scientist of different Institutions: the IMM-CNR Catania, Italy; the Universitat Autónoma de Barcelona, Spain; the Minatec Institute, Grenoble, France; and Soochow University, China. From 2012 to 2014, he was a Visiting Scientist with the Technion Institute, Israel, under a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme. He is also a Research Staff of the National Council of Science and Technology (CONICET) and a Full Professor with National Technological University (UTN), Buenos Aires, Argentina. He managed several funded research projects in the field of radiation effects of CMOS technology and reliability of ultra-thin gates oxides, supported by CONICET, The Ministry of Science and Technology of the Argentinean Government and the Ministero degli Affari Esteri (MAE) of the Italian Government. To date, he has authored/coauthored close to 80 international peer-reviewed publications and four review articles. In 2019, he was awarded by the IEEE-IRPS as the Best Paper.



JORDI SUÑÉ (Fellow, IEEE) is currently a Full Professor of electronics with the Universitat Autònoma de Barcelona (UAB). He is also the coordinator of the NANOCOMP Research Group, dedicated to the modeling and simulation of electron devices with a multi-scale approach. His main contributions are in the area of gate oxide reliability for CMOS technology. In terms of research achievements in this field, he was upgraded to IEEE Fellow for contributions to the understanding of gate oxide failure and reliability methodology. Since 2008, he has been working in the area of memristive devices and their application to neuromorphic circuits. He also launched a New Research Group/Network (neuromimeTICs.org) dedicated to the application of neuromorphic electronics to artificial intelligence and to dissemination activities. He has (co)authored more than 400 papers (H-index = 44) in international journals and relevant conferences, including 14 IEDM papers, several invited papers, and five tutorials on oxide reliability at the IEEE-IRPS. He is also the local UAB Coordinator of a European project on emerging non-volatile memories embedded in microprocessors for automotive, secure, and general electronics applications. His current research interests include transition metal oxide-based filamentary RRAM memristors; complex perovskite oxide based memristors; bio-realistic compact modeling of memristors for neuromorphic applications; RRAM fabrication, characterization and modeling; biomimetic electrical circuit simulation with SPICE; analog circuits based on the combination of CMOS and memristors; and, in general, artificial neural networks for artificial intelligence applications. In 2008, he received the IBM Faculty Award for a long-lasting collaboration with IBM Microelectronics in this field and the ICREA ACADEMIA Award in 2010. In 2012 and 2013, he was awarded the Chinese Academy of Sciences Professorship for Senior International Scientists, for collaboration with IMECAS, Beijing, China.

ENRIQUE MIRANDA (Senior Member, IEEE) received the Ph.D. degree in electronics engineering from the Universitat Autònoma de Barcelona (UAB), Spain, in 1999, and the Ph.D. degree in physics from the Universidad de Buenos Aires, Argentina, in 2001. He is currently a Professor with UAB. He was a Visiting Scientist at IIT-India, Technical University Darmstadt, IHP-Germany, the Universita di Napoli, Modena, Cagliari-Italy, and Soochow University, China. He has authored or coauthored around 250 peer-review journal articles most of them devoted to the electron transport mechanisms in thin dielectric films. He received numerous scholarships and awards, including the INTERCAMPUS (Universidad de Zaragoza, Spain), the MUTIS (UAB), the RAMON y CAJAL (UAB), the DAAD (Technical University Hamburg-Harburg), from the Italian Government (Universita degli Studi di Padova), the MATSUMAE (Tokyo Institute of Technology, Japan), the TAN CHIN TUAN (Nanyang Technological University, Singapore), the WALTON Award from Science Foundation Ireland (Tyndall National Institute), the Distinguished Visitor Award (Royal Academy of Engineering, U.K.), the CESAR MILSTEIN (CNEA, Argentina), the Visiting Professorships from the Abdus Salam International Centre for Theoretical Physics, Slovak Academy of Sciences, Politecnico di Torino, and the Leverhulme Trust (University College London, U.K.). He serves as a member of the Distinguished Lecturer Program of the Electron Devices Society (EDS-IEEE) since 2001 and an Associate Editor of *Microelectronics Reliability* since 2003.