Conductance variability in RRAM and its implications at the neural network level

Aziza, H.; Fieback, M.; Hamdioui, S.; Xun, H.; Taouil, M.

DOI
10.1016/j.microrel.2025.115594

Publication date
2025

Document Version
Final published version

Published in
Microelectronics Reliability

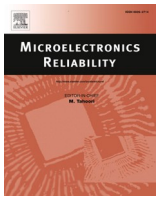
Citation (APA)
Aziza, H., Fieback, M., Hamdioui, S., Xun, H., & Taouil, M. (2025). Conductance variability in RRAM and its implications at the neural network level. *Microelectronics Reliability*, *166*, Article 115594. https://doi.org/10.1016/j.microrel.2025.115594

Microelectronics Reliability 166 (2025) 115594

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel



Conductance variability in RRAM and its implications at the neural network level☆

H. Aziza [a,*], M. Fieback [b], S. Hamdioui [b], H. Xun [b], M. Taouil [b]

[a] Aix-Marseille University, CNRS, IM2NP, 5 rue Enrico Fermi, 13451 Marseille Cedex 20, France
[b] Computer Engineering Laboratory, Delft University of Technology, Mekelweg 4, 2628CD Delft, the Netherlands

ARTICLE INFO

Keywords:
RRAM
Variability
Neuromorphic
Computing
Synaptic weights
Reliability

ABSTRACT

While Resistive RRAM (RRAM) provides appealing features for artificial neural networks (NN) such as low power operation and high density, its conductance variation can pose significant challenges for synaptic weight storage. This paper reports an experimental evaluation of the conductance variations of manufactured RRAMs memory cells at the memory array level. Variability is evaluated with respect to the RRAM low resistance state (LRS) and high resistance state (HRS) conductance ratio. This ratio is selected as the parameter of interest as it guarantees the proper operation of the RRAM: the larger the ratio, the more reliable and robust the RRAM cell is in storing and retrieving data. The measurement results show that conductance ratio is significantly influenced by variability. Using these findings, the performance of an artificial neural network that uses individual RRAM cells for synaptic weight storage is evaluated in relation to conductance variability. It is shown that RRAM variability can heavily affect the network behavior, resulting in a substantial decrease in the classification accuracy during inference.

1. Introduction

Resistive RAM (RRAM) is a promising technology not only for large data storage but also to enable energy efficient computing solutions which could facilitate the deployment of artificial intelligence at the edge (edge-AI) [1]. However, not solving the issues related to non-idealities such as the variability in the electrical parameters of RRAMs (e.g., conductance variability) may hinder the technology's continued advancement [2,3]. In RRAM-based neural networks (NN), conductance variability results in weight variability [4–6]. Weight variability can affect the network during training and inference, affecting the network ability to make precise predictions [7,8]. Therefore, there is an urgent need to analyze and quantify the conductance variability in RRAMs.

A solution to improve the network resilience against conductance fluctuation issues is to intentionally inject some noise into the synaptic weights during the training, exploiting a technique called variability-aware training (VAT) [9]. To obtain realistic results after the training process, such noise should be linked to the actual variability of the RRAM device, including device to device (D2D) and cycle to cycle (C2C) variabilities. However, this last point is neglected in many publications [10]. An alternative way to mitigate conductance fluctuations issues at the NN level is the mapping-aware biased training methodology [11] which consists in identifying RRAM conductance states inherently more immune to variation (favorable states). Then, a mapping-aware training technique is adopted where important weights are directly get mapped to such favorable states [12]. Mapping-aware techniques take into account the inherent non-idealities of RRAM devices, such as variations in conductance levels [12]. As a result, identifying devices affected by variability issues is a critical step before the practical implementation of a mapping-aware training methodology. However, in this case as well, this aspect is often ignored in many publications [13].

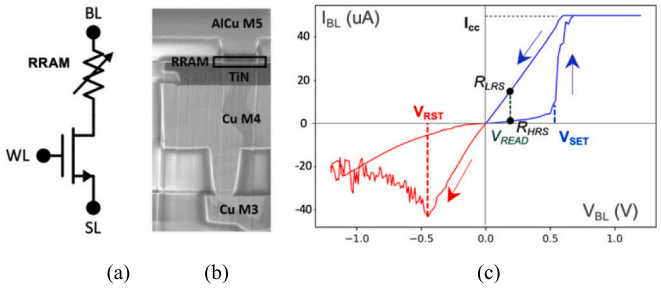
In this context, this paper advances the state-of the art by providing a silicon-based analysis of the conductance variability in RRAMs. Conductance variability is assessed quantitatively for each cell of a memory array test chip. Afterwards, a ranking of the cells more immune to variability is established. Finally, the study is extended to a basic NN used for image classification.

The main contributions of this study are summarized below:

☆ This article is part of a special issue entitled: LATS24 published in Microelectronics Reliability

* Corresponding author.
E-mail addresses: hassen.aziza@univ-amu.fr (H. Aziza), m.c.r.fieback@tudelft.nl (M. Fieback).

https://doi.org/10.1016/j.microrel.2025.115594
Received 17 July 2024; Received in revised form 13 January 2025; Accepted 14 January 2025
Available online 1 February 2025



(a) (b) (c)

Fig. 1. (a) Schematic view of a 1T1R memory cell. (b) SEM cross section of the RRAM stack [23]. (c) RRAM I-V hysteresis.

- RRAM conductance variation silicon data are collected at the test chip level.
- A deep analysis of the conductance variation over multiple cycles is provided.
- The impact of variability on a RRAM neural network performances is assessed.
- Outcomes of this work are supported by silicon results provided by ST-Microelectronics.

Considering that the limited precision of RRAM devices intended to map synaptic weights is addressed [14], outcomes derived from this study can be applied to any mapping technique currently used to implement RRAM-based NN accelerators, namely, (a) multilevel [15,16], (b) binary [17], (c) unary [18], (d) multilevel with redundancy [19] and (e) slicing [8]. Moreover, this study contributes to the understanding of the conductance variation in RRAMs [20] from an electrical standpoint, which is the first step before enabling accurate analogue computing with imprecise memory devices [21,22].

The remainder of this paper is organized as follows. Section 2 introduces the specifications of the manufactured RRAM cells. Section 3 presents the experimental setup. Section 4 reports the silicon measured data on RRAM conductance variability and analyze them. Section 5 shows how conductance variability can impact the performances of a RRAM-based NN used for image classification. Finally, Section 6 concludes the paper.

2. Specifications of the manufactured RRAMs

RRAM devices typically operate based on the reversible change in resistance caused by the formation and rupture of conductive filaments (CFs) [15]. When a voltage is applied across the cell (i.e., between the top and bottom electrodes), depending upon the voltage polarity, one or more CFs made out of oxygen vacancies are either formed or ruptured. Once the conductive filaments (CFs) are formed within the metal oxide, bridging the top and bottom electrodes, they establish a low-resistance state (LRS). Subsequent changes in resistance are achieved by rupturing the filaments. Applying a voltage with reversed polarity causes the filaments to break, leading to a high-resistance state (HRS). Fig. 1a presents the considered 1T1R RRAM memory cell where one transistor (W = 0.8 µm and L = 0.5 µm) is connected in series with one resistive element (RRAM). The resistive element, shown in Fig. 1b, is incorporated in the Back End Of Line (BEOL) of a 130 nm technology, between metal layers [23]. The stack is deposited using Physical Vapor Deposition (PVD) where a 10 nm Hafnium dioxide (HfO_2) layer is placed on the top of a TiN Bottom electrode (BE). A Ti/TiN bilayer stack is then deposited as a top electrode (TE) to form a capacitor-like structure. Fig. 1c illustrates the typical I-V characteristics of a 1T1R device, showcasing a hysteresis behavior. Based on this hysteresis, the memory cell operation can be understood as follows: after an initial electro-Forming (FMG) stage, the memory element can be switched reversibly between LRS and HRS. The resistance change is triggered by applying specific biases across the 1T1R cell: V_{SET} to switch to LRS after a SET operation and V_{RST} to switch to HRS after a RESET (RST) operation.

Table 1
Standard cell operating voltages.

	FMG	RST	SET	READ
WL	2 V	2.5 V	2 V	2.5 V
BL	3.3 V	0 V	1.2 V	0.1 V
SL	0 V	1.2 V	0 V	0 V
Resistance	10 kΩ	240 kΩ	15 kΩ	–
Conductance	100 µS	4 µS	66,6 µS	–

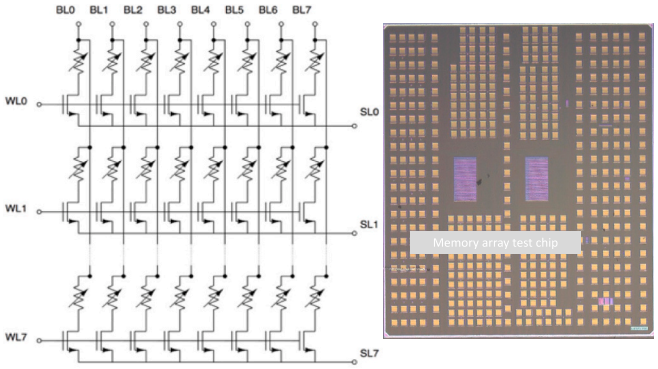


Fig. 2. (a) 8 × 8 RRAM memory array and (b) physical view of the fabricated memory array.

Table 1 presents the voltage levels applied during the various operating stages, along with the corresponding nominal resistance and conductance values. It should be noted that a nominal conductance ratio of approximately 16 is achieved for the targeted technology (66.6 µS divided by 4 µS). During the READ operation, a small read voltage (typically 0.1 V) is used to avoid disrupting the cell's state. Also, it is important to note that in the 1T1R configuration, the transistor regulates the current flowing through the cell based on its gate voltage bias. This controlled current is known as the compliance current (I_{CC}).

3. Experimental setup

Fig. 2a shows the test chip used for measurements, which is a typical 1T1R memory array. Memory cells are grouped to form eight 8-bit memory words. Word Lines (WL_X) signals are used to address a specific row, Bit Lines (BL_X) signals are used to address specific columns during a SET operation and Source Lines (SL_X) signals are used to RST a whole memory word or an addressed cell. To allow a full flexibility during characterization, BL, WL and SL nodes are externally available. During the RRAM cell characterization, the extraction of R_{LRS} and R_{HRS} is achieved using 1 ms DC voltage sweeps with a 1 mV voltage step; the applied voltage increases step by step and the current flowing through the cell is measured, allowing an extraction of the I-V characteristics of each cell. Fig. 2b shows the fabricated memory array. Due to the probe card's limited pinout, only a 7 × 7 memory array is accessible for our experiments, which represents a subset of the full 8 × 8 array.

Before any operation, each cell of the memory array is first formed. Then, memory cells are RST one by one to extract the R_{HRS} value at 0.1 V. After RST, cells are SET to extract the R_{LRS} value, also at 0.1 V. The RST/SET process is repeated 230 times for the whole array in order to catch C2C as well as D2D variability. A total of 230 cycles is used to evaluate the stability of the conductance ratio for each memory cell without addressing long-term degradation. This limited number of cycles is chosen to avoid the influence of reliability factors, such as endurance and retention, on the extracted resistance levels [23]. In other words, a "time-zero" robustness assessment is conducted before any stress effect is observed. The measurement protocol seen by each

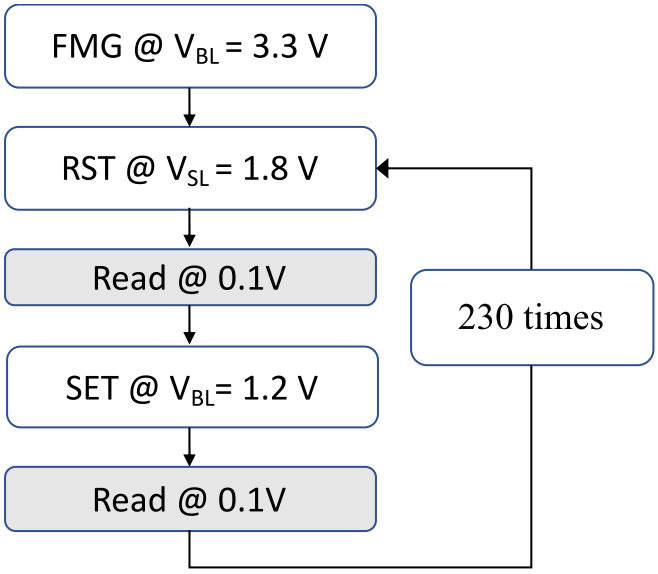


Fig. 3. Measurement protocol: after FMG, a RST/SET operation is repeated 230 times for each addressed cell. RST and SET operations are followed by a read operation to extract the cell resistances.

cell of the array is presented in Fig. 3.

4. Experimental results

This section examines the variability of RRAM devices, with a focus on both temporal and spatial variations in the LRS and HRS states. Also, a comprehensive evaluation of the conductance ratio variability is provided.

4.1. RRAM variability experimental evidence

Although RRAMs have shown interesting properties, one of the most important challenges of the technology is the control of the device variability (temporal and spatial) in both LRS and HRS states [24,25]. In fact, variations of R_{HRS}/R_{LRS} are so unpredictable that they have been employed as an entropy source in True Random Number Generators (TRNG) [26,27]. Fig. 4 shows the impact of D2D and C2C variability at the I-V characteristic level after RST/SET operations applied to each of the 49 cells of the memory array (D2D variability, Fig. 4a) and after a RST/SET operation applied only 49 times (for comparison purposes) to an isolated cell of the memory array (C2C variability, Fig. 4b). The nominal characteristic is highlighted in red (RST) and blue (SET) colors.

This qualitative analysis demonstrates that different I-V hysteresis signatures can arise when cycling the same cell (C2C) or when extracting different I-V hysteresis patterns for different cells within the memory array (D2D). To complement this analysis, a quantitative analysis of variability is conducted in the next sub-section. Based on these preliminary measurement results, it appears clearly that HRS and LRS resistance/conductance is affected by spatial and temporal variations. Hence, this non-ideality has to be considered when designing RRAM-based NN. In the next section, a cell tracking analysis will be conducted in order to monitor the evolution of the conductance ratio of each cell of the memory array presented in Fig. 2a over 230 programming cycles. The state of individual memory cells will be tracked to detect cells that deviate from their nominal behavior (i.e., deviation from the nominal conductance ratio of 16).

4.2. Conductance ratio variability evaluation

In Fig. 5, the evolution of the LRS/HRS conductance ratio of three different cells (i.e., located at three different addresses) is presented in the logarithmic scale. Cell (5;0), where '5' and '0' represent the WL and BL line numbers respectively, is the most affected by variability.

Large conductance fluctuations are reported with a conductance ratio standard derivation $\sigma = 97.6$ with respect to its mean value $\mu = 61.5$. In contrast, cell (1;0) and cell (3;0) are less impacted with standard derivation values equal to 6.3 and 13 respectively. Note that for cell

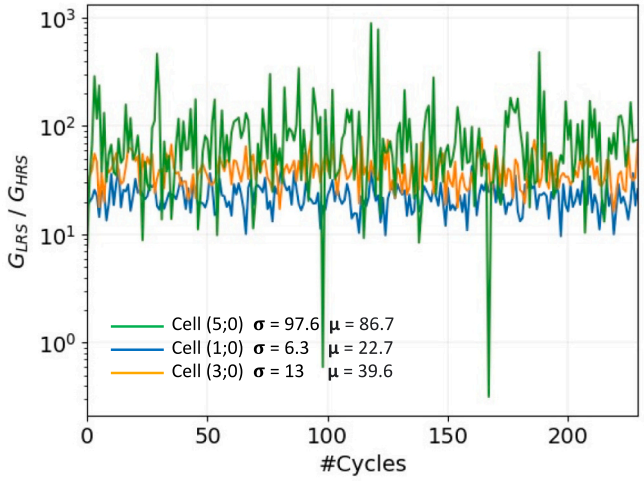


Fig. 5. Conductance ratio versus the number of RST/SET cycles for 3 different cells of the memory array presented in Fig. 2a.

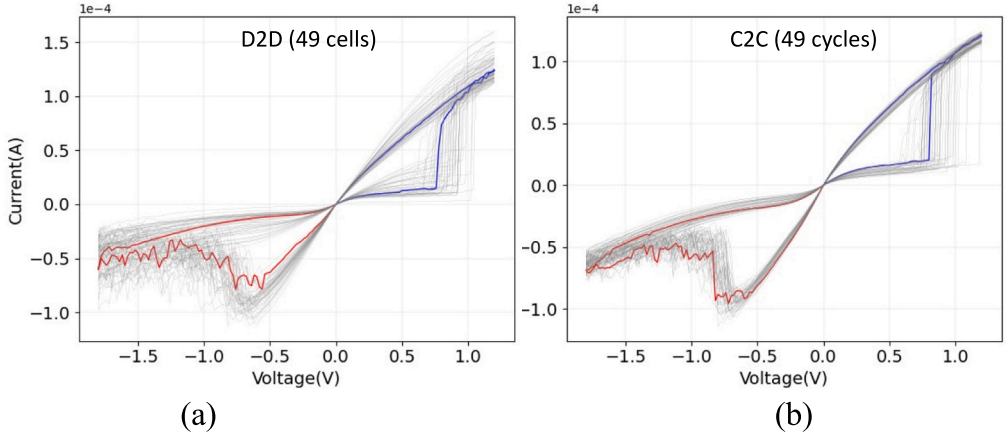


Fig. 4. Experimental evidence of (a) cell level D2D variability and (b) cell level C2C variability. The nominal characteristic is highlighted in color.

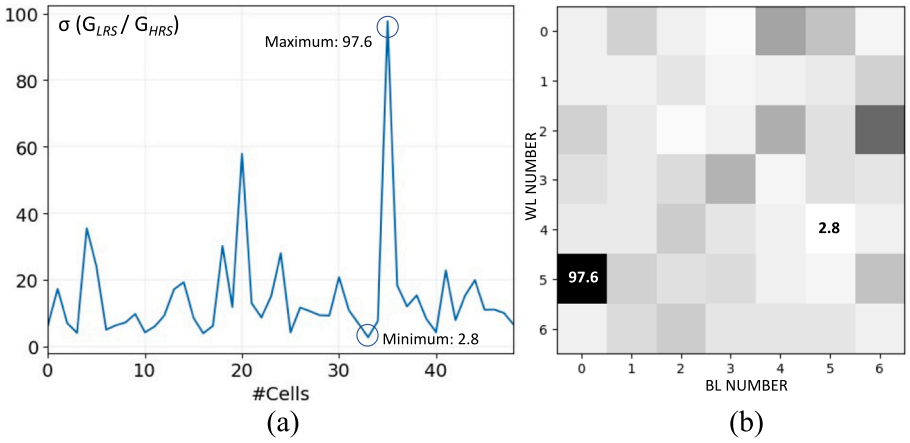


Fig. 6. (a) Evolution of the conductance ratio standard deviation of each cell of the memory array. (b) Topological representation of the standard deviation of each cell of the memory array. The values of the most impacted cell (97.6) and least impacted cell (2.8) are reported in (a) and (b).

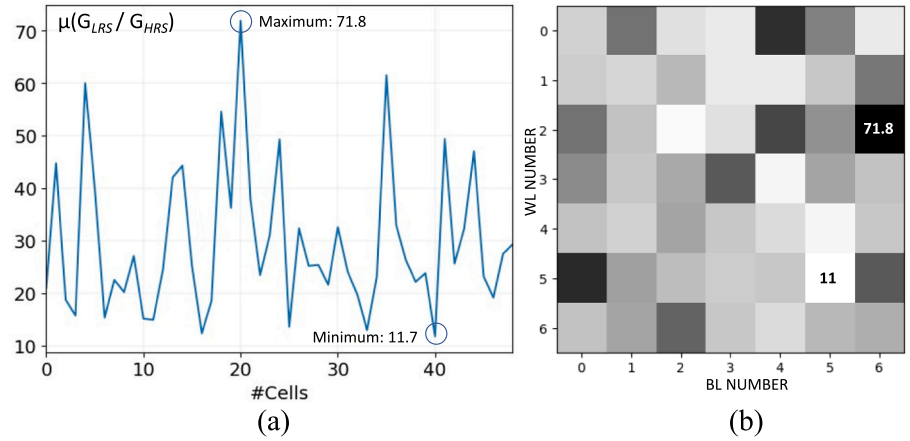


Fig. 7. (a) Evolution of the conductance ratio mean value of each cell of the memory array. (b) Topological representation of the mean value of each cell of the memory array. Largest and smallest values are reported in (a) and (b).

(5;0), the conductance ratio falls below one in two cycles, resulting in an overlap between LRS and HRS conductance levels. Hence, this cell needs to be avoided for synaptic weight storage. The evolution of the conductance ratio standard derivation of the 49 cells of the memory array is provided in Fig. 6a. The standard deviation ranges from 2.8 (min. value) to 97.6 (max. value).

A 2D representation of the standard deviation values over the memory array is presented in Fig. 6b. Each cell is associated with a variable degree of grey. The whiteness of a cell reflects lower standard deviations. The white color being associated with the minimal standard deviation and the black color with the maximal standard deviation. For instance, cell at location (5;0), associated with a black color, is the most affected by variability with a standard deviation of 97.6, while cell at location (4;5), associated with a white color, is the least affected by

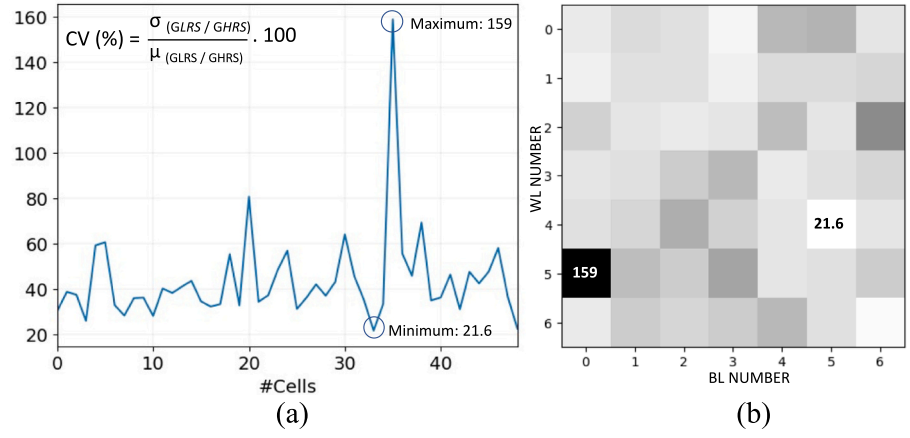


Fig. 8. (a) Evolution of the coefficient of variation CV of each cell of the memory array. (b) Topological representation of CV for each cell of the memory array.

Table 2
Favorable cells ranking.

#	CV (%)	σ	μ (S)	(WL; BL)
1	21.6	2.79	12.04	(4;5)
2	22.4	6.58	29.33	(6;6)
3	25.8	4.07	15.72	(0;3)
4	28.0	4.25	15.15	(1;3)
5	28.2	6.35	22.53	(1;0)
6	30.2	6.33	20.97	(0;0)
7	31.0	7.95	25.64	(6;0)
8	31.1	4.24	23.62	(3;4)
9	32.2	3.97	12.35	(2;2)
10	32.7	11.80	36.25	(2;5)
Worst cell	159	97.6	61.5	(5;0)

variability regarding its standard deviation of 2.8.

Fig. 7 presents the evolution of the mean value of the conductance ratio for each cell of the memory array. Interestingly, this parameter is also affected by variability, demonstrating that the conduction window differs across the cell in the array.

The fluctuation in the mean value of the conductance ratio is crucial when mapping NN weights, as a narrower conductance window leads to a substantial decrease in the cell's ability to modulate conductance (i.e., a reduction in the number of available analog conductance levels).

The objective of Figs. 5, 6, and 7 is to present a synthesis of the variability analysis of the memory array, highlighting the most robust cells based on their conductance ratio standard deviations and their conductance ratio mean values. Fig. 8 focuses on the evolution of the ratio of the standard deviation over the mean value (σ/μ) for each cell of the memory array. This parameter is a dimensionless quantity that is used to measure the relative variability of the conductance ratio dataset, even if the datasets have different scales (i.e., different mean values). It is referred to as the coefficient of variation CV. The formula for calculating CV is given in (1).

$$CV\ (\%) = \frac{Standard\ deviation}{Mean} \bullet 100 = \frac{\sigma}{\mu} \bullet 100 \quad (1)$$

Dividing the standard deviation by the mean value essentially standardize the measure of the variability. In Fig. 8a, the minimum CV value of 21.6 % indicates that the standard deviation is relatively small compared to the mean, while the maximum CV value of 159 % suggests a larger relative variability. As this parameter combine the influence of the standard deviation and the mean value, the latter will be considered in the upcoming discussion section.

4.3. Favorable cells ranking

Based on the CV parameter, a ranking of the most favorable cells (i.e., cells with lower μ/σ ratio) is proposed in Table 2. The CV parameter (column 2) accounts for both the stability (σ contribution, column 3) and the mean value (μ contribution, column 4) of the conductance ratio. The addresses of each cell are reported in column 5.

According to Table 2 and based on the NN application requirements, favorable conductance states presenting low CV values can be chosen to map significant weights [13]. Conversely, conductance states presenting high CV values (such as the worst cell in Table 2 last column) can be skipped during the weight mapping process due to less immunity to variations.

5. Variability aware neuromorphic computing

This section focuses on evaluating the implications of conductance variability for RRAM-based artificial neural networks (ANNs) during the inference stage. It describes the process of modeling and simulating variability effects on an ANN accuracy.

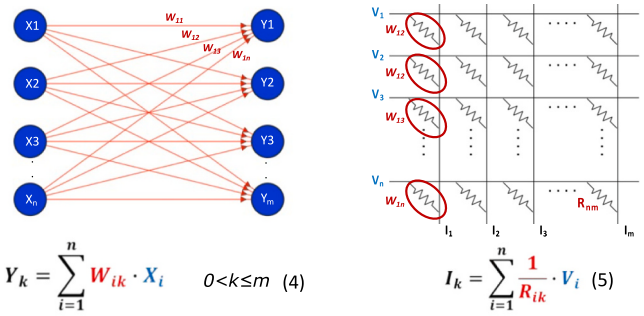


Fig. 9. (a) Two-layer feed-forward neural network and (b) neural network mapping to a crossbar RRAM array. The VMM algorithm is reported for the abstract NN (Eq. 4) and for the RRAM-based NN (Eq. 5).

5.1. Neural network level implications

Table 2 reveals that conductance variability is a major concern for RRAM-based computing. Hence, the performance of artificial neural networks (ANN) relying on individual RRAM cells to store the synaptic weights has to be assessed versus conductance variability.

In ANNs, there are two distinct phases, each with specific purposes and processes: training and inference. Inference refers to the process of using a trained model to make predictions on new data. It consists in retrieving stored weights during a read operation and performing computations based on them. In contrast, training is the process of learning the weights of a model from data. It involves iterative updates of the model's parameters using optimization algorithms such as gradient descent, requiring both read and write operations. RRAM is commonly used during inference owing to its fast and energy-efficient read operations, making it a good candidate for read-intensive applications. The latter are generally associated with in-memory computing and more particularly with ANN applications where synaptic weights are constantly and simultaneously read during inference. Additionally, using RRAM exclusively during inference prevents cycling endurance issues as the RRAM technology has a finite number of write cycles before the cells degrade, which can be a limitation in the context of training where frequent memory write operations are required.

Introducing conductance variability into a RRAM ANN during inference involves simulating the actual imperfections of the RRAM devices. Indeed, and as already mentioned, variability in conductance leads to inaccuracies in weight representations, thereby degrading the inference accuracy of the overall ANN. To describe the conductance variability, a mathematical model based on a Gaussian distribution is assumed and proposed in Eq. (2).

$$Greal = Gnominal + \Delta G \quad (2)$$

where, G_{real} is the actual conductance used during inference, which includes the variability, $G_{nominal}$ is the nominal conductance without variability and ΔG, given in (3), is a random variable following a normal distribution with a mean value referred to as μ and a standard deviation σ representing the conductance variation.

$$\Delta G \sim N(\mu, \sigma) \quad (3)$$

5.2. Neural network level evaluation

In this section, a comprehensive step-by-step description of a RRAM-based ANN for image classification is presented. The network is then evaluated against conductance variability. For clarity, we consider a relatively simple network (i.e., single layer perceptron).

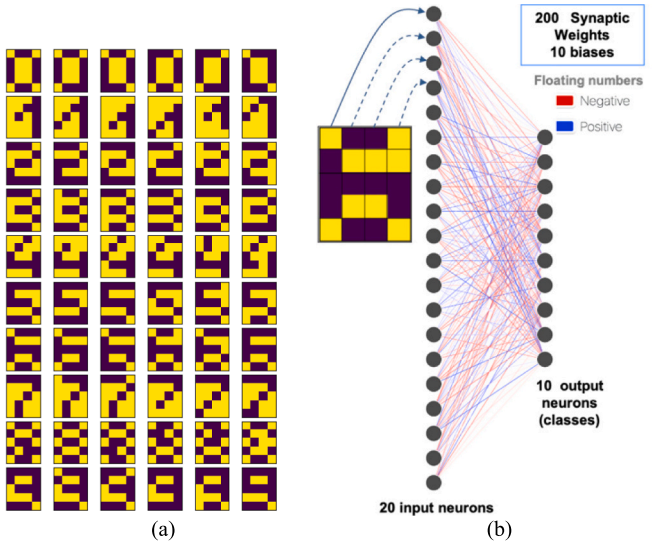


Fig. 10. 60 5 × 4 grayscale patterns representing digits (0–9). Each class gathers six different instances of the same digit. (b) Two-layer fully connected neural network made of 20 input neurons and 10 output neurons. The network is defined by 210 parameters (200 weights and 10 biases).

5.2.1. *Vector-matrix multiplication*

The most studied network architecture in the literature is the fully connected NN where each neuron in a layer is connected to every neuron in the previous layer. In this work, a two-layer fully connected NN presented in Fig. 9a is considered. This feed-forward network can be used for image classification based on a linear predictor function combining a set of weights W_{ik} with the input vector X_i. Outputs Y_k are computed using Vector-Matrix Multiplication (VMM), which is the fundamental computation algorithm in neural networks (see Eq. 4 of Fig. 9a).

An array of RRAM cells could naturally accomplish VMM within one step by collecting the output current of the array. Fig. 9b describes how the neural network of Fig. 9a can be mapped to a crossbar RRAM array. Input vectors (X_x) are mapped to input voltages (V_x) and the weight matrix (W_{ik}) is mapped to memory cell conductance values $1/R_{ik}$. VMM can be easily implemented following Eq. 5 of Fig. 9b: when voltages V_x are applied to the rows, the current through each cell is proportional to the product of the input voltage and the cell's conductance (which represents the weight). The weighted sum is obtained by measuring the total current I_k. It is worth mentioning that an activation function (not presented here) can be applied to the total current. Note that the RRAM crossbar structure allows a simultaneous computation of multiple dot products as currents in all columns are instantaneously summed by Kirchhoff's Current Law (KCL). Before performing a VMM, different weights are loaded into the crossbar matrix. Hence, the key point of this approach is the ability of the RRAM device to store data as different conductance levels.

5.2.2. *Image dataset and neural network architecture*

Given the complexity and technology-specific nature of actual RRAM hardware implementation, simulating the impact of variability on RRAM neural network performances can be effectively achieved through software. In this context, *TensorFlow* neural network libraries are used to define and train an ANN model using a custom image dataset.

In the context of image classification, an image is represented as a matrix of pixels with dimensions $n \times m$. We have considered 5 × 4 grayscale images representing digits (0–9), in which the color of pixels is codified by one single value: each pixel has a value between 0 (black) and 255 (white). Note that before use, a preprocessing step is needed to convert the images into a format suitable for NN training and inference

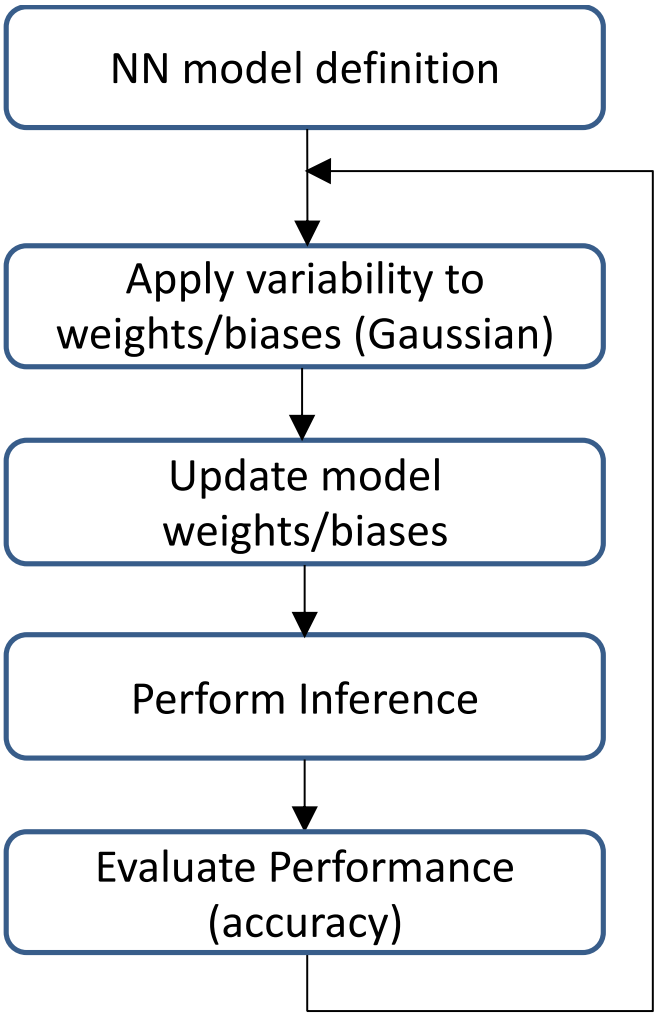


Fig. 11. Conductance variability simulation workflow under *TensorFlow* framework.

(pixel values are normalize to a range of 0 to 1). Fig. 10a shows the considered dataset made of 10 different classes. Each class gathers six different instances of the same digit.

Given the small size of the images, a simple feedforward NN with two layers is targeted (see Fig. 10b). The input layer is fed with 20 inputs encoding the 5 × 4 input pixels. The output layer is made of 10 output neurons and determines the final prediction of the model. The model parameters include 200 synaptic weights along with 10 biases. It is important to note that synaptic weights and biases can be represented using both positive and negative floating-point numbers.

5.2.3. *Impact of variability during inference*

Variability in RRAM is modeled under the *TensorFlow* framework by introducing noise (or variations) in the synaptic weights and biases during inference. More particularly, variability is incorporated into the ANN model while primary focusing on C2C variability, since the study examines resistance fluctuations occurring after a programming operation is applied to a specific cell.

Accuracy is chosen as the key performance metric to evaluate the robustness of the NN under different variability conditions. Fig. 11 presents the conductance variability simulation workflow: (*i*) After the definition of the targeted NN model, (*ii*) a mathematical model representing the variability characteristics of RRAM devices (e.g., Gaussian noise with a specific mean and variance) is used to generate weights and biases values that account for variability, (*iii*) the NN model is updated

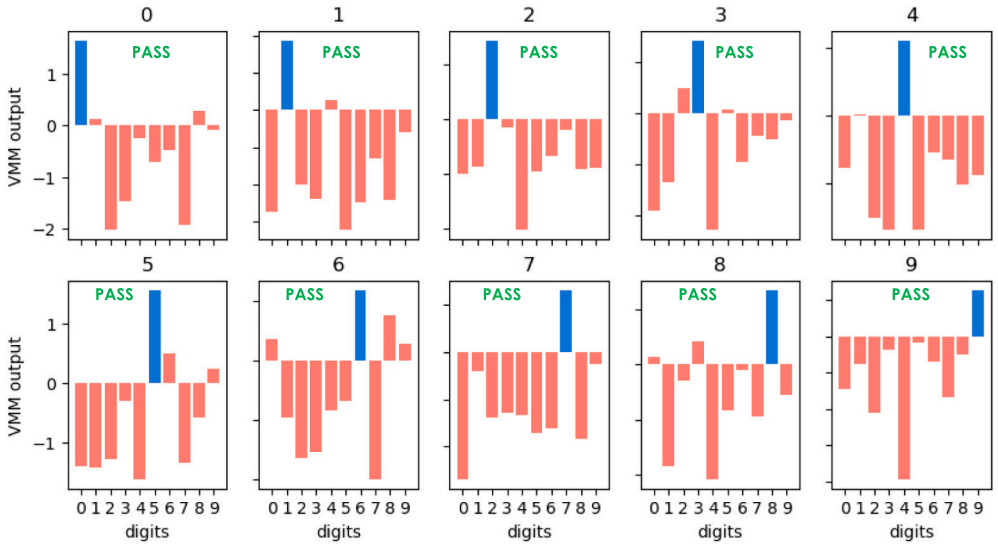


Fig. 12. Golden simulation: inference test achieved without variability. An overall accuracy of 100 % is reached.

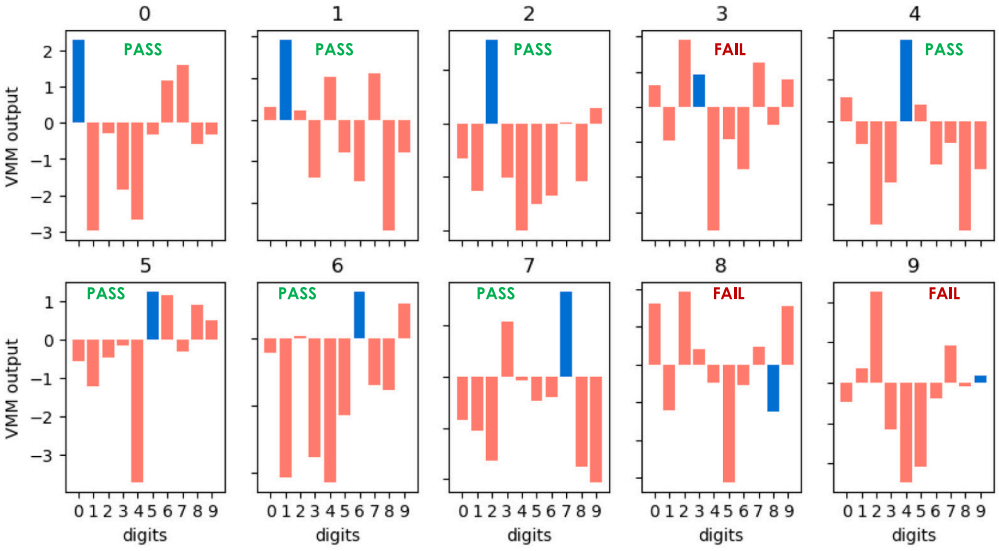


Fig. 13. Inference test achieved after nose injection in the NN. A standard deviation with respect to the mean value (σ/μ) of 0.25 is considered.

with new weight and bias values and (*iv*) the model is simulated during inference to extract the accuracy of the predictions. The same process is repeated with new input images to calculate the average accuracy across multiple inputs. A new image dataset (that was not used during training) is derived from the dataset presented in Fig. 10. In this new dataset, the color of half of the pixels of each image has been modified to values different from 0 (black) and 255 (white). 15 and 240 have been chosen for black and white respectively.

Fig. 12 shows simulation results related to the considered NN without considering variability (golden simulation). The *softmax* function is skipped to see the direct result of the linear VMM transformation. Indeed, *softmax* converts the raw output of the network into probabilities, hence, masking the VMM transformation result. In Fig. 12, for each digit, ranging from 0 to 9, the blue bar represents the VMM result for the considered digit. The higher the blue bar value, the better the inference accuracy. For each digit, the blue bar shows a higher value compared to the red bars, hence, a "pass" test is reported for each tested digit, thereby achieving an overall accuracy of 100 %.

The same protocol is used for the second inference test, presented in Fig. 13, expect that variability has been injected into the weights and biases of the NN model. A standard deviation with respect to the mean value (σ/μ) of 0.25 is considered (CV = 25). Under these conditions, a global accuracy of 70 % is reported.

Simulations with varying levels of introduced variability have been performed, and their impact on the NN accuracy is shown in Fig. 14. The impact of variability on the accuracy can be significant. Beyond 0.1 (CV = 10), the accuracy reaches 10 %. Beyond 0.5 (CV = 50), the accuracy drops below 4 %.

If we consider experimental variability results reported in Table 2, and more particularly the CV parameter variation range (changing from 20 to 150), it appears clearly that the actual RRAM conductance variation can lead to a noticeable drop in the classification accuracy.

6. Discussion

In this study, the conductance ratio has been chosen as the main criterion to assess the robustness of RRAMs used in computing applications for two reasons: (i) a stable conductance ratio is essential for

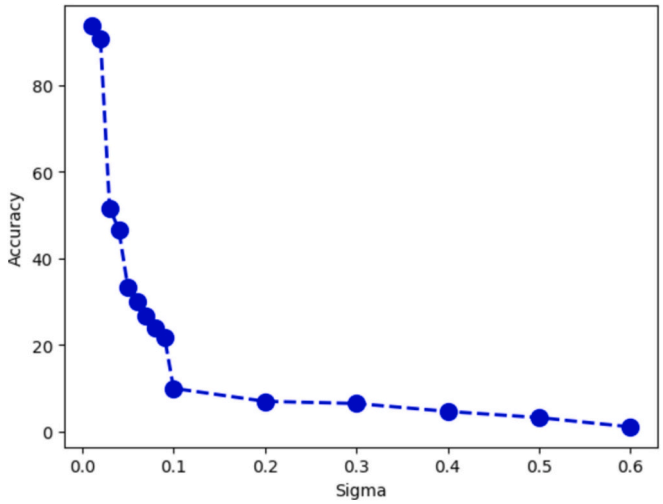


Fig. 14. Accuracy of the NN versus the standard deviation with respect to the mean value (σ/μ).

consistent learning processes, enabling the network to adapt to new information while updating the previously stored information (ii) a high conductance ratio provides a larger dynamic range for multi-level cell storage (MLC [28]) which enables better differentiation between different synaptic states, turning the NN more robust. The conductance ratio has been monitored against 230 RST/SET programming cycles. The standard deviation, the mean value and the CV parameters have been computed to analyze the behavior of each cell. By leveraging software simulations, the impact of RRAM variability on a NN dedicated to image classification has been evaluated during inference. It has been demonstrated that RRAM variability can significantly impact the accuracy of the NN. While the conductance variability is an important criterion at time zero [29], it is worth noting that time-dependent reliability metrics [6] such as endurance, retention and read/write stress also play a critical role in determining the robustness of RRAM-based NNs. Particularly, cycling and endurance can lead to hard errors (memory cell stuck at one conductance state forever, with a conductance ratio stuck at one [30]). Also, similarly to other emerging memory technologies, RRAMs is subject to defects that directly impact the conductance ratio [31]. Therefore, appropriate test mechanisms are required to detect RRAM-related failures due to these defects [32,33]. Beyond RRAMs, the NN CMOS subsystem variability [34] (including the neurons [35], the RRAM reading [36] and programming circuitry [37,38]) can also impact the conductance ratio. Hence, a complete analysis strategy [39] has to be defined to mitigate the impact of all these non-idealities on the conductance fluctuations in RRAM-based NN accelerators.

7. Conclusion

The existing of important fluctuations in the RRAM conductance has been experimentally established. The electrical behavior of each cell of an elementary array has been analyzed at the electrical level. After having computed the coefficient of variation CV of each cell of the array, a large variation of this parameter has been reported (from 21.6 to 159). Based on these results, it has been demonstrated that the impact of the RRAM conductance variability on the accuracy of a neural network can be significant and cannot be ignored, especially in precision-critical applications like image classification. Hence, investigating strategies to mitigate this variability is crucial in order to maintain high performances in RRAM-based neural networks.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Authors wish to acknowledge the support from the CEA-Leti and ST-Microelectronics (technology access as part of MAD200 project). This work is also partially funded by EU's Horizon research program (grant agreement No. 101070374).

Data availability

The data that has been used is confidential.

References

[1] A. Gebregiorgis, et al., Tutorial on memristor-based computing for smart edge applications, Memories-Materials, Devices, Circuits and Systems 4 (2023) 100025.
[2] A. Gebregiorgis, et al., Dealing with non-idealities in memristor based computation-in-memory designs, in: 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC), IEEE, 2022.
[3] C. Bengel, et al., Reliability aspects of binary vector-matrix-multiplications using ReRAM devices, Neuromorphic computing and engineering 2 (3) (2022) 034001.
[4] A. Gebregiorgis, et al., RRAM crossbar-based fault-tolerant binary neural networks (BNNs), in: 2022 IEEE European Test Symposium (ETS), IEEE Computer Society, 2022.
[5] H. Aziza, Oxide-based resistive RAM analog synaptic behavior assessment for neuromemristive systems, in: Memristors - the Fourth Fundamental Circuit Element - Theory, Device, and Applications, IntechOpen, 2023.
[6] H. Aziza, et al., On the reliability of RRAM-based neural networks, in: IFIP/IEEE International Conference on Very Large-Scale Integration (VLSI-SoC) 2023, 2016. In Press.
[7] W. Wang, et al., Integration and Co-design of Memristive Devices and Algorithms for Artificial Intelligence, iScience, 2020.
[8] A. Shafiee, et al., ISAAC:A Convolutional Neural Network Accelerator With In-Situ Analog Arithmetic in Crossbars, ISCA, 2016.
[9] A. Ankit, et al., PUMA: A Programmable Ultra-efficient Memristor-based Accelerator for Machine Learning Inference, ASPLOS, 2019.
[10] Q. Wang, et al., Device variation effects on neural network inference accuracy in analog in-memory computing systems, Adv. Intell. Syst. 4 (8) (2022) 2100199.
[11] H. Xiao, et al., Fashion-MNIST: a novel image dataset for bench-marking machine learning algorithms, arXiv (2017).
[12] H. Aziza, et al., Design considerations towards zero-variability resistive RAMs in HRS state, in: IEEE 22nd Latin American Test Symposium (LATS), 2021, pp. 1–5.
[13] S. Diware, et al., Mapping-aware biased training for accurate memristor-based neural networks, in: IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), 2023, pp. 1–5.
[14] G. Pedretti, et al., Conductance variations and their impact on the precision of in-memory computing with resistive switching memory (RRAM), in: IEEE International Reliability Physics Symposium (IRPS), 2021, pp. 1–8.
[15] V. Milo, et al., Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks, APL Mater. 7 (8) (2019) 081120.
[16] F. Alibart, et al., High Precision Tuning of State for Memristive Devices by Adaptable Variation-tolerant Algorithm, 2012, p. 8.
[17] S. Yu, et al., Binary neural network with 16 Mb RRAM macro chip for classification and online training, in: 2016 IEEE International Electron Devices Meeting (IEDM), 2016, pp. 16.2.1–16.2.4.
[18] C. Ma, et al., Go unary: a novel synapse coding and mapping scheme for reliable ReRAM-based neuromorphic computing, in: 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2020, pp. 1432–1437.
[19] I. Boybat, et al., Neuromorphic computing with multi-memristive synapses, Nat. Commun. 9 (1) (2018) 2514.
[20] H. Aziza, et al., Multi-level control of resistive ram (rram) using a write termination to achieve 4 bits/cell in high resistance state, Electronics 10 (18) (2021) 2222.
[21] Can Li, et al., Efficient and self-adaptive in-situ learning in multilayer memristor neural networks, Nat. Commun. 9 (1) (2018) 2385.
[22] Peng Yao, et al., Face classification using electronic synapses, Nat. Commun. 8 (1) (2017) 15199.
[23] A. Grossi, et al., Fundamental variability limits of filament-based RRAM, in: 2016 IEEE International Electron Devices Meeting (IEDM), 2016, pp. 4.7.1–4.7.4.
[24] H. Aziza, et al., Evaluation of OxRAM cell variability impact on memory performances through electrical simulations, in: IEEE Non-Volatile Memory Technology Symposium Proceeding, 2011, pp. 1–5.
[25] B. Hajri, et al., RRAM device models: a comparative analysis with experimental validation, IEEE Access 7 (2019) 168963–168980.
[26] H. Aziza, et al., True random number generator integration in a resistive RAM memory Array using input current limitation, IEEE Trans. Nanotechnol. 19 (2020) 214–222.
[27] H. Aziza, et al., Bipolar OxRRAM memory array reliability evaluation based on fault injection, in: IEEE 6th International Design and Test Workshop (IDT), 2011, pp. 78–81.

[28] H. Aziza, et al., Density enhancement of RRAMs using a RESET write termination for MLC operation, in: IEEE proc. of Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021, pp. 1877–1880.
[29] H. Aziza, et al., STATE: a test structure for rapid prediction of resistive RAM electrical parameter variability, in: IEEE International Symposium on Circuits and Systems (ISCAS), 2022, pp. 3532–3536.
[30] H. Aziza, et al., STATE: a test structure for rapid and reliable prediction of resistive RAM endurance, IEEE Transactions on Device and Materials Reliability 22 (4) (2022) 500–505.
[31] M. Fieback, et al., Defects, fault modeling, and test development framework for RRAMs, ACM Journal on Emerging Technologies in Computing Systems (JETC) 18 (3) (2022) 1–26.
[32] M. Fieback, et al., Device-aware test: a new test approach towards DPPB level, in: IEEE International Test Conference (ITC), 2019, pp. 1–10.
[33] M. Fieback, et al., Intermittent Undefined State Fault in RRAMs, IEEE European Test Symposium (2021) 1–6.
[34] Y. Joly, et al., Matching degradation of threshold voltage and gate voltage of NMOSFET after hot carrier injection stress, Microelectron. Reliab. 9 (51) (2017) 1561–1563.
[35] H. Aziza, et al., A capacitor-less CMOS neuron circuit for neuromemristive networks, in: IEEE International Conference on Electronics Circuits and Systems (NEWCAS), 2019.
[36] W.S. Ngueya, et al., An ultra-low power and high performance single ended sense amplifier for low voltage flash memories, Journal of Low Power Electronics 14 (1) (2018) 157–169.
[37] H. Aziza, et al., ReRAM ON/OFF resistance ratio degradation due to line resistance combined with device variability in 28nm FDSOI technology, in: Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), 2017, pp. 35–38.
[38] H. Aziza, A. Bosio, Embedded memory test, Silicon Systems For Wireless Lan 22 (2020) 387 (ISO 690).
[39] F. Su, C. Liu, H.-G. Stratigopoulos, Testability and dependability of AI hardware: survey, trends, challenges, and perspectives, IEEE Design & Test 40 (2) (2023) 8–58.