electronics

MDPI

*Article*

# A Noise-Resilient Neuromorphic Digit Classifier Based on NOR Flash Memories with Pulse–Width Modulation Scheme

**Gerardo Malavena *, Alessandro Sottocornola Spinelli and Christian Monzio Compagnoni**

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy; alessandro.spinelli@polimi.it (A.S.S.); christian.monzio@polimi.it (C.M.C.)
* Correspondence: gerardo.malavena@polimi.it

**Abstract:** In this work, we investigate the implementation of a neuromorphic digit classifier based on NOR Flash memory arrays as artificial synaptic arrays and exploiting a pulse-width modulation (PWM) scheme. Its performance is compared in presence of various noise sources against what achieved when a classical pulse-amplitude modulation (PAM) scheme is employed. First, by modeling the cell threshold voltage ($V_T$) placement affected by program noise during a program-and-verify scheme based on incremental step pulse programming (ISPP), we show that the classifier truthfulness degradation due to the limited program accuracy achieved in the PWM case is considerably lower than that obtained with the PAM approach. Then, a similar analysis is carried out to investigate the classifier behavior after program in presence of cell $V_T$ instabilities due to random telegraph noise (RTN) and to temperature variations, leading again to results in favor of the PWM approach. In light of these results, the present work suggests a viable solution to overcome some of the more serious reliability issues of NOR Flash-based artificial neural networks, paving the way to the implementation of highly-reliable, noise-resilient neuromorphic systems.

**Keywords:** artificial neural networks; neuromorphic computing; NOR Flash memory arrays; program noise; random telegraph noise; pulse-width modulation

check for updates

**Citation:** Malavena, G.; Sottocornola Spinelli, A.; Monzio Compagnoni, C. A Noise-Resilient Neuromorphic Digit Classifier Based on NOR Flash Memories with Pulse–Width Modulation Scheme. *Electronics* **2021**, *10*, 2784. https://doi.org/10.3390/electronics10222784

Academic Editor: Changhwan Shin

Received: 27 September 2021
Accepted: 10 November 2021
Published: 13 November 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Artificial neural networks (ANNs) are computing systems that take inspiration from biological neural networks to address many problems involving unstructured data, such as image recognition and classification [1,2]. What makes ANNs different from classical CMOS systems based on the Von Neuman architecture is that they do not feature distinct computing and memory units that communicate with each other through a bus; rather, they implement an in-situ computational paradigm, which is based on the matrix-by-vector multiplication (MVM) operation [1]. For that reason, ANNs represent a promising solution to achieve a performance and efficiency improvement in those systems designed to perform data-intensive tasks, for which the memory bottleneck, arising from the continuous data exchange between memory and CPU, represents a limiting factor.

A convenient way to implement ANNs consists in exploiting non-volatile memory (NVM) arrays as artificial synaptic arrays connecting adjacent layers of artificial neurons. To that purpose, different memory solutions have been investigated for their adoption in neuromorphic systems and presented in literature. They include works based on crossbar arrays of resistive elements, mainly resistive switching random access memories (RRAM) [3–5] and phase change memories (PCM) [6,7], or based on memory arrays of charge storage devices, such as NAND and NOR Flash memory arrays [8–15].

Among those different solutions, the adoption of Flash memory technologies sounds appealing for many reasons, such as their reduced power consumption, the virtually analog tuning of the synaptic weights stored in the memory array, and their mature and reliable CMOS-compatible manufacturing process. In particular, even though some ANNs implementations based on NAND Flash arrays have been presented [15,16], the parallel architecture of NOR Flash memory arrays makes them the most straightforward solution to implement the MVM at the basis of the operation of neuromorphic systems [1,17].

For this reason, different implementations of neuromorphic systems based on NOR Flash memory arrays have been analyzed, including both supervised and unsupervised networks, which usually rely on some modification to the cell design [18–20], to the array design [2,8–10] or to cells program/erase voltage schemes [11–13] to make the memory array operation compliant with the desired application. Notably, in [2] a fully integrated three-layer ANN (with dimensions 784 × 64 × 10) was implemented and tested for handwritten digits recognition via the gradient-descent method based on the backpropagation algorithm [21] reaching a 94.7% classification fidelity with a single-pattern classification time and energy equal to 1 µs and less than 20 nJ, respectively.

In this work, we take inspiration from [16], where a working scheme based on PWM is adopted for the implementation of a neuromorphic image classifier based on NAND Flash memory arrays, and show that a similar PWM-based approach can be employed to operate a NOR Flash-based neuromorphic digit classifier, replacing the typically adopted PAM scheme. In particular, by means of a simulation-based analysis, we demonstrate, thanks to that PWM scheme, the possibility to achieve a tremendous reduction in classifier sensitivity to noise sources such as PN, RTN, and temperature variations. The results of this analysis present a way to strongly relieve those issues, thus enabling the development of noise-resilient artificial neural networks based on scaled NOR Flash memory arrays.

After a brief review on related works dealing with various techniques to reduce noise sensitivity in ANNs (Section 2), PAM and PWM encoding schemes are introduced in Section 3 and Section 4, respectively. After that, in Section 5, we will present the architecture of the investigated neuromorphic digit classifier and the simulation results when no noise sources are accounted for. Then, in Section 6 we will discuss the impact of PN, RTN, and temperature variations on the classifier performance. Finally, conclusions will be drawn in Section 7.

## 2. Noise-Sensitivity Reduction Techniques in ANNs

Despite the advantages coming with ANNs with respect to Von Neunman architecture-based systems, their analog computing paradigm is inherently affected by noise, regardless the type of NVM arrays adopted. In fact the synaptic weights stored in the memory cells are inevitably impacted by several non-idealities, either occurring during the program phase and leading to a limited program accuracy, or following it, undermining the synaptic weights stability over time. The deviation of the stored synaptic weights from their ideal value will lead to a degradation of the network performance that may ultimately compromise its functionality.

For this reason, a considerable research effort is being devoted to the conception of various techniques to design neural networks with low noise sensitivity. For example, conductance variability in PCMs due to $1/f$ noise, drift noise [22], program noise (PN, [23]), and device variability have been recently addressed. In [24], additive noise is injected during the learning phase to train networks that are more robust to cells conductances noise; in [25], instead, the combination of multiplicative noise injection and drift regularization is presented to reduce network performance degradation due to PN and drift noise of about a factor 10.

Other techniques to relieve network reliability issues due to stochastic variations of RRAM devices resistance are presented in [26,27], resorting again on noise injection during training; in particular in [27] non-idealities arising from the IR drops along the resistance network are addressed. In [28], instead, the basic idea of biasing the learning phase to encourage the network to learn large synaptic weights is proposed; even though that results in a reduction in noise because large synaptic weights are less sensitive to conductance instabilities, it comes with the drawback of an increase in the network power consumption.

In the case of NOR Flash memory arrays, PN and RTN are two major issues of concern for network performance. Their impact on the classification accuracy of a NOR Flash-based neuromorphic digit classifier was studied in [29]. In that work, it was shown that stringent requirements on the memory array programming scheme and on the memory cells scaling are needed to limit the network performance degradation within an acceptable interval. The investigation of techniques that allow to relieve the impact of PN and RTN on ANNs based on NOR Flash memory arrays is therefore crucial to overcome those limitations in future implementations and to improve their performance.

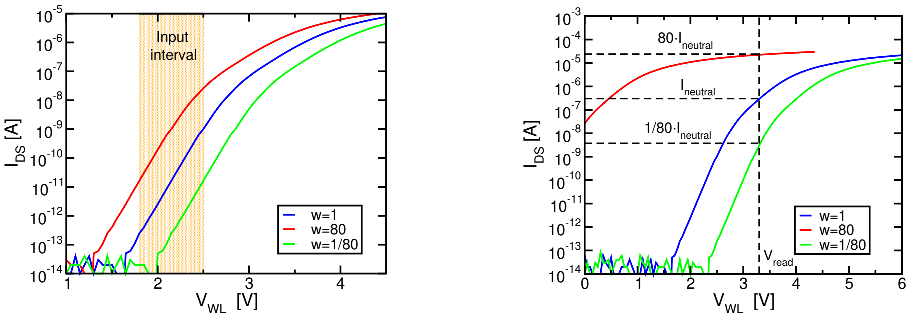
## 3. Pulse Amplitude Modulation

When employed in neuromorphic applications, NOR Flash cells are typically operated in subthreshold regime [30,31], where the drain-to-source current $I_{DS}$ displays an exponential dependence on the word-line (WL) voltage $V_{WL}$ according to:

$$\underbrace{I_{DS}}_{\text{output}} = \underbrace{I_0 \cdot \exp\left[\frac{q\alpha_G\left(V_{WL}-V_T^{ref}\right)}{mkT}\right]}_{\text{input}} \cdot \underbrace{\exp\left[-q\frac{\alpha_G \Delta V_T}{mkT}\right]}_{\text{weight}} \tag{1}$$

In the previous equation, $I_0$ is the current prefactor, $q$ is the elementary charge, $\alpha_G$ is the control-gate–to-floating-gate capacitive coupling ratio, $m$ is the subthreshold slope ideality factor, $kT$ is thermal energy, $V_T$ is the cell threshold voltage, $V_T^{ref}$ an arbitrary chosen reference cell $V_T$, and $\Delta V_T$ is the cell $V_T$ shift from $V_T^{ref}$. According to this approach, the input weight synaptic multiplication is implemented by considering $w = \exp[-q\alpha_G \Delta V_T / mkT]$ as the weight of the artificial synapse, while the remaining factor, which is a function of $V_{WL}$ but not of the cell $V_T$, plays the role of the input presynaptic signal. Since the input signal is modulated by the amplitude of $V_{WL}$, this approach can be referred to as pulse-amplitude modulation (PAM).

As shown in Figure 1a, by changing cell $V_T$ it is possible to modulate $I_{DS}$ at a given $V_{WL}$ and, therefore, the cell synaptic weight, with positive and negative $\Delta V_T$ leading to $w < 1$ and $w > 1$, respectively. Even though the synaptic weights could in principle assume analog values, their tuning resolution is ultimately limited by charge granularity, becoming relevant especially in very scaled devices [32]. In addition to this, an upper limitation for them is needed to keep the cell from exiting the subthreshold regime. In fact, according to the PAM mode the input weight synaptic multiplication is implemented thanks to the exponential law of Equation (1), and the polynomial $I_{DS} - V_{WL}$ characteristic of cells operating in the on-state regime would not allow such a behavior. For the same reason, once a maximum synaptic weight $w^{max}$ is chosen, also an upper bound for $V_{WL}$ must be selected. It follows that only a small portion of the $I_{DS} - V_{WL}$ curve can be exploited, resulting in a limited acceptable $\Delta V_T$ range.



(**a**) (**b**)

**Figure 1.** (**a**) $I_D - V_{WL}$ curve measured on a 40 nm embedded technology [33] and used to simulate the implementation of a NOR Flash-based digit classifier adopting a PAM approach. $I_D - V_{WL}$ curves corresponding to different values of the synaptic weights and the corresponding limited input range are shown. (**b**) Same as (**a**) but when PWM is employed; no input limitation is required in the latter case.

In addition to that, another drawback peculiar of the PAM approach is the strong sensitivity of $w$ to $\Delta V_T$ due to the exponential law that relates these two quantities. This represents an issue not to be overlooked, especially for scaled technologies, as several non-idealities could determine fluctuations in the value of $\Delta V_T$, resulting in even stronger variations of $w$. Indeed, in [29] the impact of program accuracy and of $V_T$ instabilities due to RTN on the performance of a NOR Flash-based neuromorphic digit classifier operated according to the PAM approach was investigated. The study was conducted by modeling the cell $V_T$ placement during a program-and-verify (P&V) scheme based on ISPP [34,35], and by performing a parametric analysis of the main dependences that affect the program phase, i.e., the $V_T$ discretization step ($V_s$) and the control-gate–to–floating-gate capacitance ($C_{pp}$). In addition to that, the role of cell $V_T$ instabilities due to RTN [36] was addressed for different values of the single-trap fluctuation amplitude ($\lambda$). Results revealed that to keep the degradation of the network performance within an acceptable range, a stringent upper bound to $V_s$ and $\lambda$ and a lower one to $C_{pp}$ are mandatory. This poses a severe limitation on the array programming time, mainly ruled by $V_s$ [34,37], and on the possibility to employ deeply scaled NOR Flash technologies, that would lead to lower $C_{pp}$ and larger $\lambda$ values [38–40]. In addition, for similar reasons, also unwanted temperature variations are expected to be detrimental for the classifier truthfulness when the PAM scheme is employed.

## 4. Pulse Width Modulation

The PWM approach is based on the idea of exploiting a different encoding scheme for the presynaptic signals. According to it, these are not encoded in the amplitude of the $V_{WL}$ pulse (such as the PAM approach), but rather in its duration with a constant amplitude. This means that $V_{WL}$ is kept constant to a value $V_{read}$, and its time duration changes according to the input signal: large signals correspond to long pulses and the other way round. To keep the proportionality between input and output signals, the overall charge that flows through the cell (and not the current) during the input pulse is taken as the output signal, and therefore Equation (1) is modified in:

$$\underbrace{\text{Charge}}_{\text{output}} = \underbrace{\text{Time}}_{\text{input}} \times \underbrace{\text{Current}}_{\text{weight}}. \tag{2}$$

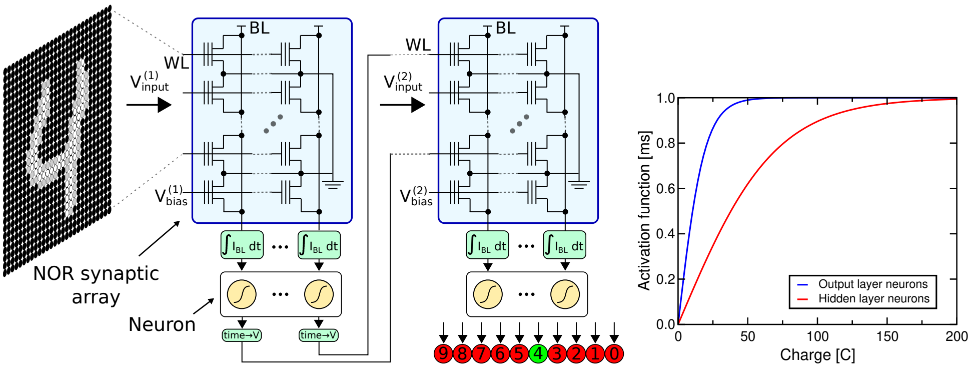
Since no limitation of the cell working regime is required in the PWM scheme, even though the cell weight is still related to the $\Delta V_T$ value, the link between them cannot be expressed using a single mathematical form anymore. As shown in Figure 1b, $I_{DS}$ for $\Delta V_T = 0$ and $V_{WL} = V_{read}$ corresponds to $w = 1$, and is referred to as $I_{neutral}$; then, when the cell is programmed with $\Delta V_T \neq 0$, $I_{DS}$ will be larger or lower than $I_{neutral}$, leading to an operative definition of the synaptic weight as $w = I_{DS}/I_{neutral}$.

What really makes such a PWM approach appealing is that one cell working point can span both the subthreshold and the on-state regime, allowing the $\Delta V_T$ range to be much wider than that of the PAM scheme. In addition to this, those memory cells that operate in the on-state regime present a polynomial relation between $I_{DS}$ and $V_T$, with a reduced sensitivity of their $w$ to $\Delta V_T$. For those reasons, the PWM approach looks more promising when dealing with the previously mentioned noise sources and surely deserves some attention.

The obvious downside of PWM, on the other hand, is that the implementation of large cell weights requires memory cells to be biased with higher $I_{DS}$ if compared to the PAM case, leading to a reduction in the classifier energy efficiency. Even though this, in principle, may be a limiting factor, from a practical standpoint the maximum current never exceeds 30 µA in our work (see the next section), which is comparable with typical operating current in PCM-based ANNs [25].

## 5. Neuromorphic Digit Classifier Based on PWM

In order to prove the benefits coming during adoption of the PWM scheme, we considered the NOR Flash-based neuromorphic digit classifier investigated in [29] and operated according to the PAM approach (see Figure 2a). It consists of a three-layer fully-connected feed-forward ANN trained to recognize the hand-written digits belonging to the MNIST database [41]. Since the MNIST digits are represented as a 28 × 28 greyscale images and range from "0" to "10", the input layer is made of 784 neurons, each one providing an analog input for a pixel, and the output layer features 10 neurons, equal to the number of digits to be identified; the number of neurons in the hidden layer, instead, was set to 40 to keep the classifier dimension as small as possible. According to such architecture, the NOR Flash memory arrays employed to connect each couple of adjacent layers of neurons must have dimensions equal to (784 + 1) × 40 and (40 + 1) × 10, respectively, where one more WL is needed to implement the bias $b$ of each neuron. It is worth mentioning that, according to the presented design choice with a single memory cell storing a synaptic weight, only positive weights can be reproduced. Following [29], the impact of this limitation on the classifier truthfulness was strongly mitigated by employing a complementary encoding of network outputs, which consists in identifying the digit that is shown at the input of the classifier not by looking at the neuron in the output layer with the highest output signal, but at the one with the lowest. Note that, even though a differential implementation of cell synaptic weights (see [1,2]) would easily lead to a even larger recognition truthfulness due to the possibility to reproduce negative weights too (at the expense, however, of a double array area occupancy), we preferred to keep a single-cell synaptic weight implementation, to focus on the impact of various noise sources on the classifier recognition accuracy, rather than on the maximization of its absolute value.

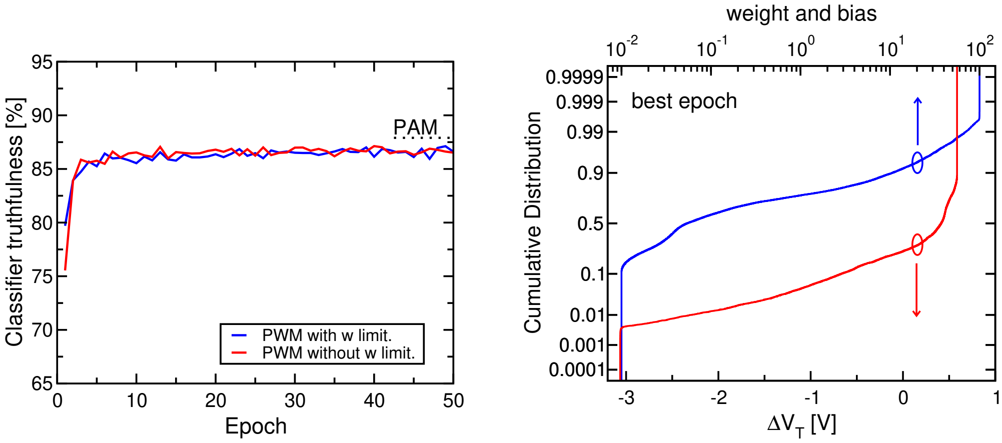


(**a**) (**b**)

**Figure 2.** (**a**) Schematic of the NOR Flash-based neuromorphic digit classifier investigated in this work when PWM is employed (adapted from [29], © 2019 IEEE). (**b**) Activation functions used for the neurons in the hidden and output layers; different scaling parameters were chosen to match the different dimensions of the respective NOR Flash arrays.

As explained in Section 4, for the classifier to be operated according to the PWM scheme, the input signals ($x_i$) must be encoded in the time duration of the voltage pulses applied to the WLs of NOR Flash arrays ($t_i$). Differently from the PAM case, in which the amplitude of input pulses must be confined in a well-defined interval to keep the memory cells in the subthreshold regime, no strict limitation affects this design choice. Therefore we picked $T = 0$ ms and 1 ms as the minimum and maximum pulse durations, respectively, leading to $t_i = x_i \cdot T^{max}$. Then, the working point of the memory cells was defined by considering the trans-characteristic shown in Figure 1b, measured on a NOR Flash cell developed with a 40 nm embedded technology [33] at a drain-to-source voltage $V_{DS} = 200$ mV. In particular, the reference I–V curve was taken for $V_T^{ref} = 3.1$ V, measured with a constant current criterion at $I_{DS} = 100$ nA; $V_{read}$ was chosen to be equal to 3.3 V, leading to $I_{neutral} = 300$ nA. Finally, the tanh–neuron activation function shown in Figure 2b was adopted; it was properly chosen to be consistent with the PWM scheme, therefore receiving a charge signal at its input and delivering a time signal at its output in the range [0 ms, 1 ms].

Figure 3a shows the recognition truthfulness of the PWM-based classifier during its training with the standard stochastic gradient-descent method based on the backpropagation algorithm [21] (blue curve), compared to the final value resulting from the classically adopted PAM approach. In both cases, a cross-entropy error function, a mini-batch size equal to 10 and a learning rate equal to 0.01 were adopted [21]. When no noise source are taken into account, the PWM and PAM schemes are practically equivalent from the performance standpoint at the end of the training phase, confirming the validity of the former. The red curve in Figure 3a, on the other hand, refers to the simulation results achieved with the PWM approach when the synaptic weights are forced in the interval $[w^{min}, w^{max}] = [0.01, 100]$; this means that if $w < w^{min}$ ($w > w^{max}$) results from the network training, its value is clamped to $w^{min}$ ($w^{max}$). Since no relevant variations in the classifier truthfulness are observed, this limitation will be safely applied to all the results presented in the following, allowing to limit $I_{DS}$ for each cell in the interval $[3\,\text{nA}, 30\,\mu\text{A}]$.



(**a**) (**b**)

**Figure 3.** (**a**) Simulation results of the PWM-based classifier truthfulness during training with the stochastic gradient-descent method; results with and without weights limitation are shown and compared to the maximum accuracy achieved in the PAM case. (**b**) Cumulative distribution of $w$ and $\Delta V_T$ in NOR Flash memory arrays for the training epoch corresponding to the maximum classifier truthfulness; the steep increase in the curves at their extremes is due the weights limitation.

The cumulative distribution of $w$ and $\Delta V_T$ for the epoch corresponding to the maximum classifier truthfulness is reported in Figure 3b. The accumulation of both quantities at the extremes occurs at quite low and high probabilities, confirming that only a small number of the NOR Flash cells are affected by the limitation of the synaptic weights. In addition to this, it is worth noting that the $\Delta V_T$ values resulting from training are distributed over quite a large range (>3.5 V); this represents an encouraging result suggesting the possibility to keep good classifier performance even in presence of different noise sources.

## 6. Noise-Sensitivity Analysis of the Classifier Performance

Even though the PAM and PWM schemes ideally lead to comparable values of classifier truthfulness, for the reasons explained in the previous sections significant differences are expected when a more realistic analysis of the network is carried out. In this section, the impact of PN, RTN, and temperature variations on the network performance is addressed.

### *6.1. Impact of Program Noise*

The program operation in NOR Flash memory arrays is based on a P&V scheme relying on ISPP. According to it, each memory cell is set to an erase state first, (i.e., with a very low-$V_T$). Then, a program pulse with WL and BL voltages equal to ($V_{WL}^{P,1}$, $V_{BL}^{P}$) is applied, triggering channel hot-electron injection into cell floating-gate that makes cell $V_T$ increase. Right after the program pulse, cell $V_T$ is read (this is referred to as verify phase) and compared with a target level $V_{PV}$. If $V_T$ is lower than $V_{PV}$, another program pulse is applied to the cell, with the WL voltage increased by a quantity $V_s$, that is $V_{WL}^{P,2} = V_{WL}^{P,1} + V_s$. This procedure is repeated until the condition $V_T > V_{PV}$ is verified, with $V_{WL}^{P,i} = V_{WL}^{P,i-1} + (i-1) \cdot V_s$ ($V_{WL}^{P,i}$ represents $V_{WL}$ at the i-th step and $i > 1$), when the program phase stops. It can be shown that, after a sufficiently large number of pulses, the average $V_T$ variation for each cell due to each programming pulse is exactly equal to $V_s$ and the final cell $V_T$ is expected to be in the $[V_{PV}, V_{PV} + V_s]$ range [34,35]. This is shown schematically in Figure 4a, where each vertical rectangle corresponds to a programming pulse with $V_{WL}$ of increasing amplitude and the white squares represents the increasing cell $V_T$.

However, due to the stochastic nature of the physics ruling the injection of electrons into the cell FG, memory cells sharing the same $V_{PV}$ are affected by PN (see Figure 4b), which manifests itself as a dispersion of the values of the final $V_T$s of those cells [37]. For example, Figure 4b shows the simulated $V_T$ evolution of four different memory cells with the same $V_{PV}$. Due to PN, each cell ends up with a different final $V_T$ value at the end of the program phase.

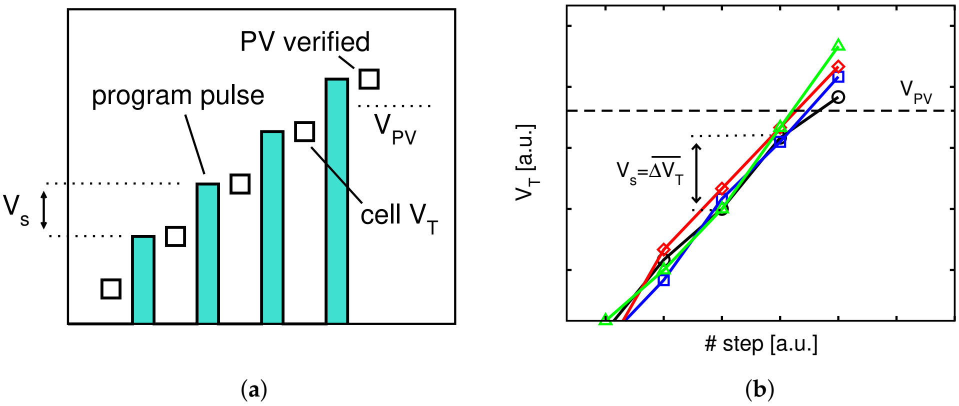
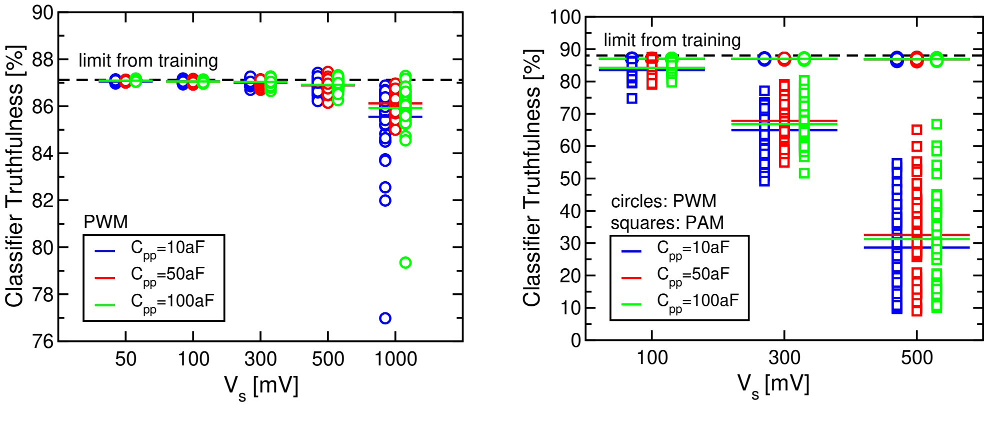


**Figure 4.** (**a**) Schematic of the ISPP phase and (**b**) simulated evolution during ISPP of the $V_T$ of four different cells with the same $V_{PV}$; due to PN each cell has a different $V_T$ at the end of program.

To evaluate the impact of PN on the performance of our NOR Flash-based digit classifier, the $V_{PV}$ levels were discretized with step equal to $V_s$, and each memory cell was associated with the closest $V_{PV}$ value lower than its target $V_T$. Then, the program operation was simulated, accounting for the randomness of the number of electron injected into cell FG during each ISPP pulse in a Monte Carlo fashion, as described in [32]. In particular, the analysis was repeated for different values of $V_s$ and $C_{pp}$, as PN is shown to be stronger in presence of large $V_s$ and small $C_{pp}$ [32]; for each $(V_s, C_{pp})$ couple, the Monte Carlo simulation was repeated 50 consecutive times, testing the classifier truthfulness after each repetition.

Simulation results are reported in Figure 5a, showing the classifier truthfulness after program as a function of $V_s$ and $C_{pp}$ when the network is operated according to the PWM scheme. Results reveal that just a weak reduction in the average classifier performance occurs in all cases, being slightly larger than 1% only for $V_s$ approaching the value of 1000 mV. In that case, even though cell weights sensitivity to noise sources affecting $V_T$ is weak thanks to the polynomial relation between $I_{DS}$ and $V_T$ itself, still such large $V_s$ values result in an extremely coarse P&V levels discretization and a strong program noise that lead, in turn, to a non-negligible reduction in the classifier truthfulness. In addition, also $C_{pp}$ has a certain impact on the average results, mainly for the $V_s = 1000$ mV case, since lower values of $C_{pp}$ tend to further enhance PN, thus leading to a noticeable performance degradation. Finally, even though a stronger statistical spread is found as $V_s$ becomes larger, just a few data points out of the 50 experiment repetitions results in a performance reduction lower than 2%, confirming again the robustness of the PWM scheme with respect to PN. Results appear even more striking if they are compared with those calculated when the digit classifier is operated according to the PAM approach, as shown in Figure 5b. In fact, a much stronger performance degradation is displayed in the latter case, not only in terms of average accuracy but also in terms of the statistical dispersion of the data. It is important to stress that $V_s > 300\,\text{mV}$ should never be used in the PAM case, as cells $\Delta V_T$ discretization would be so coarse to make some of them exit the subthreshold working region, therefore severely compromising the network performance. In this sense, another advantage coming with the adoption of PWM is the possibility to speed up the program phase by exploiting larger $V_s$, without degrading the classifier accuracy too much.



(**a**) (**b**)

**Figure 5.** (**a**) Simulated classifier truthfulness in presence of to PN for different values of $V_s$ and $C_{pp}$ when PWM is employed; for each condition, 50 consecutive simulations were performed (average values are indicated by horizontal lines). (**b**) Same as (**a**) but with the data points resulting from the adoption of the PAM scheme.

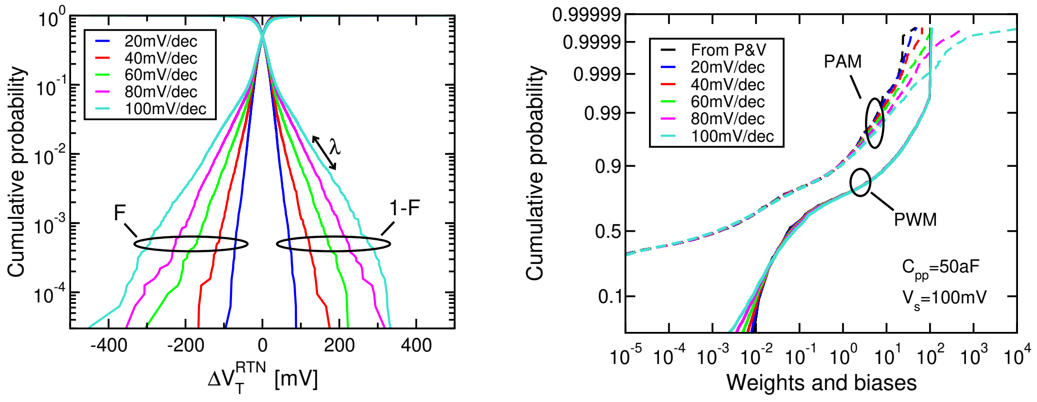
### *6.2. Impact of RTN*

Even though ISPP allows a cell $V_T$ tuning precise enough to keep network truthfulness degradation very limited, any source of $V_T$ instabilities may still compromise the classifier performance at later times. In NOR Flash memory arrays a major source of $V_T$ instabilities is represented by RTN, which arises from the capture and emission of electrons in tunnel-oxide traps. Both the amplitude and the timing of RTN-induced $V_T$ fluctuations are non-deterministic, therefore RTN is usually addressed by looking at the statistical distribution of the $V_T$ difference between two consecutive read operations ($\Delta V_T^{RTN}$) [39,42,43]. As shown in Figure 6a, a clear signature of the $\Delta V_T^{RTN}$ cumulative distribution is its exponential tails; its slope $\lambda$, which is taken as the most representative RTN parameter, can be shown to approximate the single-trap fluctuation amplitude [38,42,44] and increases when single-cell dimensions are shrunk down. The remaining parameters describing RTN are the average number of traps per cell $\langle N_t \rangle$, which determines the height of the RTN tails, and the capture and emission trap time constants, which are typically uniformly distributed over the logarithmic time axis.

In order to evaluate the impact of RTN on the classifier performance, RTN-induced fluctuations were simulated following the Monte Carlo approach presented in [45], with $N_t$ calibrated to reproduce the $\Delta V_T^{RTN}$ statistical distributions in Figure 6a and for different values of $\lambda$ spanning from 20 mV/dec to 100 mV/dec. After that, a different RTN waveform was added to the $V_T$ resulting from ISPP of each memory cell, and the network classification truthfulness was monitored periodically over time. The impact of such RTN-induced $V_T$ oscillations on the statistical distributions of the memory cells weights is shown in Figure 6b, for the case in which the program phase was simulated with $C_{pp} = 50\,\text{aF}$ and

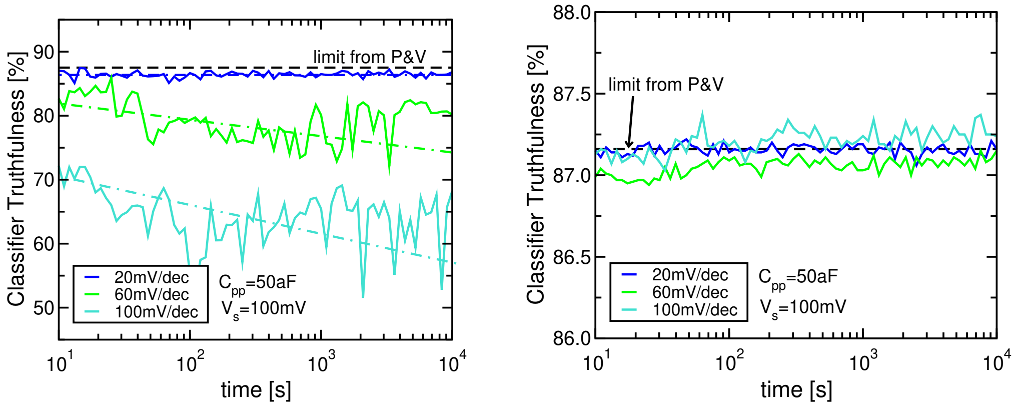
$V_s = 100\,\text{mV}$. Note that after 1000 s the enlargement in the weights distribution calculated in the PWM case is practically negligible, pointing towards a strong immunity to RTN for that scheme.

Simulation results reported in Figure 7a,b shows the RTN-induced instabilities in the classifier truthfulness for the cases in which the network is operated according to the PAM and PWM approaches, respectively. As already pointed out in [29], in the former case the classifier accuracy experiences strong instabilities, accompanied by an average degradation of the performance over time due to the increasing number of RTN traps coming into play as time goes by. This poses a limitation to the possibility to employ scaled NOR Flash memory arrays for neuromorphic applications, since the network performance degradation due to RTN becomes too severe as $\lambda$ becomes too large. On the other hand, when PWM is employed, RTN results in oscillations of the classifier truthfulness of the order of 0.1% only, therefore not significantly contributing to degrade the classifier performance. It is worth stressing that the resilience of PWM to RTN demonstrated by our analysis allows to overcome the limitations of the PAM approach, paving the way to the development of high-density neuromorphic systems based on scaled NOR Flash memory array.



(**a**) (**b**)

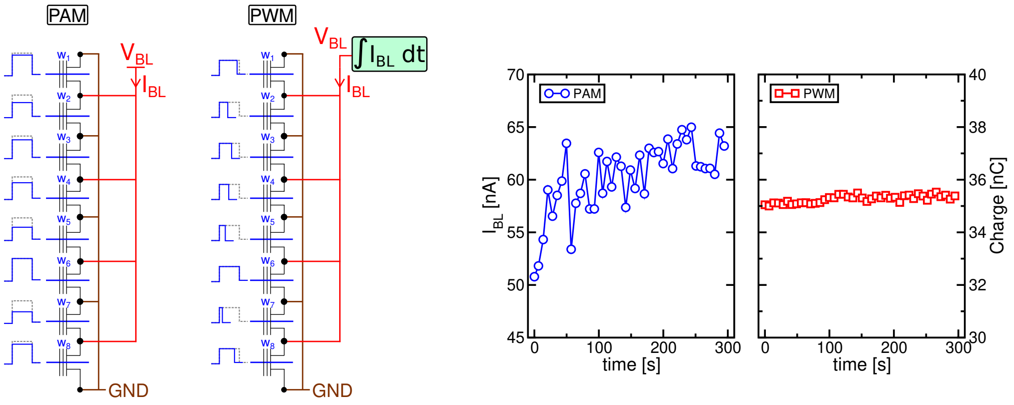
**Figure 6.** (**a**) Simulated cumulative statistical distribution (F) of $\Delta V_T^{RTN}$ and its complementary (1-F) one for different values of $\lambda$. (**b**) Cumulative statistical distribution of the memory cells weights resulting from RTN Monte Carlo simulations after 1000 s from the end of the program phase for increasing values of $\lambda$; the distribution calculated at the end of the P&V phase is shown too.



(**a**) (**b**)

**Figure 7.** Calculated classifier truthfulness in presence of RTN with increasing values of $\lambda$ for (**a**) PAM (dashed-dotted lines highlight the average reduction in performance over time) and (**b**) PWM. $C_{pp} = 50\,\text{aF}$ and $V_s = 100\,\text{mV}$ were assumed for the program phase.

To further confirm our results, we performed a wafer-level experimental test involving a NOR Flash array test structure with 8 WLs and 1 BL (see Figure 8a). The 8 memory cells along the BL were programmed by ISPP with $V_s = 300\,\text{mV}$ to $V_T$ values corresponding to weights uniformly distributed in the [1/80,80] range; each memory cell was then associated with an input signal randomly drawn in the [0,1] range. With all the input signals simultaneously applied to the WLs, the output signal was monitored for 300 s following the program phase. Results are reported in Figure 8b, for the two cases in which the experiment is carried out employing the PAM and PWM encoding schemes. As expected, PWM assures superior performance in terms of output signal stability, which is reflected in a very low RTN-sensitivity in those large-scale ANNs based on larger NOR Flash memory arrays.



(**a**) (**b**)

**Figure 8.** (**a**) Experimental test devised to reproduce the main features of a NOR Flash-based ANN in a 8WLs NOR Flash string. Each memory cell is programmed to have a weight in the [1/80, 80] range and is associated with an input signal between 0 and 1. The experiment is repeated twice, implemented first according to the PAM (left) and then according to the PWM (right) scheme. (**b**) Evolution of the output signals measured during the experiment shown in (**a**). Much larger instabilities of the output signal are measured in the former case with respect to the latter, confirming experimentally the strong immunity of PWM to noise sources affecting cells $V_T$s, such as RTN.

### *6.3. Impact of Temperature Variations*

Another source of truthfulness instabilities is represented by temperature variations that may occur after the program phase, affecting the synaptic weights and, therefore, the classifier operation. To investigate this point, we measured the single-cell $I_{DS} - V_{WL}$ transcharacteristic already shown in Figure 1 not only at 300 K, but also at temperatures as high as 420 K, as shown in Figure 9a. Then, assuming the training phase in our network to take place at 300 K, the impact of temperature variations on the classifier behavior was accounted for by transforming the $I_{DS} - V_{WL}$ curve of each NOR Flash memory cell in agreement with Figure 9a, and testing the network accuracy at the remaining temperatures. Note that the impact of PN on each cell $V_T$ resulting from program was considered negligible to focus just on the role played by temperature variations.

Figure 9b shows the results of the previous analysis, for the PAM and the PWM cases. As a consequence of the stronger temperature dependence of the curves of Figure 9a in their subthreshold region, temperature variations have a detrimental effect on the classification accuracy when the PAM scheme is employed. In the PWM case, instead, even though some memory cells operate in subthreshold regime, those with the largest weights are the ones in the on-state, displaying a much weaker temperature dependence. This results in a stronger immunity for PWM, allowing to keep the performance degradation within 10% for temperature variations close to 100 K.

A complete comparison between PAM and PWM approaches is finally reported in Table 1.

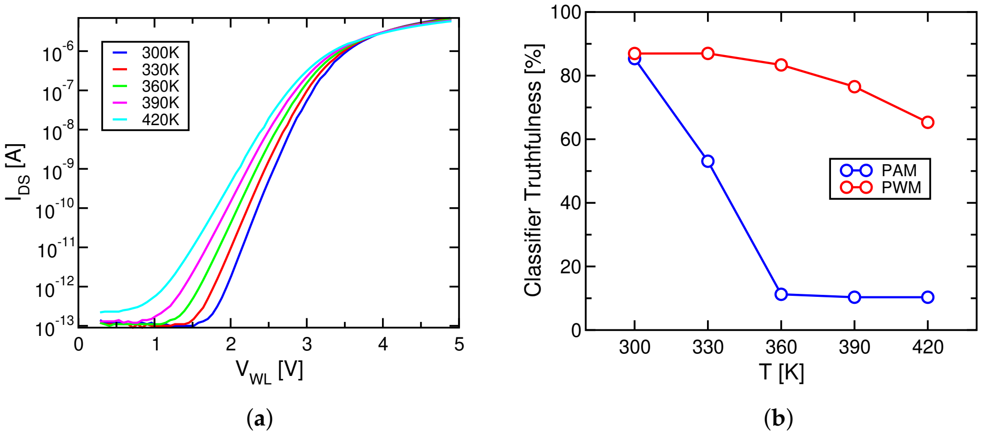


**Figure 9.** (**a**) $I_{DS} - V_{WL}$ characteristics measured on a 40 nm embedded technology at increasing values of temperature. (**b**) Classifier truthfulness evaluated following an ideal program operation (i.e., without $V_{PV}$ discretization and PN) in the PAM and PWM cases.

**Table 1.** Comparison between PAM and PWM approaches in terms of input–output encoding schemes, noise sensitivity, and energy efficiency.

| | Input | Output | PN Sensitivity | RTN Sensitivity | Temperature Sensitivity | Energy Efficiency |
|---|---|---|---|---|---|---|
| PAM | $V_{WL}$ | $I_{DS}$ | strong | strong | strong | high |
| PWM | $t_i$ | $\int_0^{t_i} I_{DS} dt$ | weak | weak | weak | moderate |

## 7. Conclusions

In this work, we have presented the implementation of a NOR Flash-based neuromorphic digit classifier based on PWM. By comparing our results with those previously reported for a similar classifier operated according to the classically adopted PAM scheme, we have shown that PWM and PAM are practically equivalent from the standpoint of the recognition accuracy when no noise sources are taken into account. Then, we have considered three distinct noise sources to affect the classifier performance, that is, PN, RTN and temperature variations, with the first limiting the cell $V_T$ tuning precision and the remaining ones impacting the cell $V_T$ stability over time after the program phase. When the impact of all those noise sources on the classifier performance is accounted for, PWM has been shown to lead to much a higher classification accuracy, representing a better choice with respect to PAM. In particular, we have shown that the superior noise immunity of PWM to PN and RTN enables the adoption of smaller $V_s$ during ISPP, thus speeding up the program phase, and of more scaled NOR Flash memory cells, leading to an increase in the network integration density. Finally, the main conclusions resulting from our simulation activities were also confirmed experimentally considering a 8 WLs test NOR Flash memory string programmed according to the PAM scheme first, and then to the PWM one. In both cases $I_{BL}$ was monitored at constant intervals over time, showing much more stable values in the latter case with respect to the former one. For all those reasons, results reported here represent an important step towards the development of large-scale high-density neuromorphic systems that employs NOR Flash memory arrays.

**Author Contributions:** Conceptualization, C.M.C.; software, G.M. and A.S.S.; investigation, C.M.C., G.M. and A.S.S.; original draft preparation, G.M.; review and editing G.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors would like to thank Simone Gangemi and Riccardo Teri for discussions and support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Burr, G.W.; Shelby, R.M.; Sebastian, A.; Kim, S.; Kim, S.; Sidler, S.; Virwani, K.; Ishii, M.; Narayanan, P.; Fumarola, A.; et al. Neuromorphic computing using non-volatile memory. *Adv. Phys. X* **2017**, *2*, 89–124. [CrossRef]
2. Merrikh-Bayat, F.; Guo, X.; Klachko, M.; Prezioso, M.; Likharev, K.K.; Strukov, D.B. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 4782–4790. [CrossRef] [PubMed]
3. Ielmini, D. Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling. *Semicond. Sci. Technol.* **2016**, *31*, 063002. [CrossRef]
4. Prezioso, M.; Merrikh-Bayat, F.; Hoskins, B.; Adam, G.C.; Likharev, K.K.; Strukov, D.B. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **2015**, *521*, 61–64. [CrossRef]
5. Yu, S.; Chen, P.Y.; Cao, Y.; Xia, L.; Wang, Y.; Wu, H. Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. In Proceedings of the 2015 IEEE International Electron Devices Meeting, Washington, DC, USA, 7–9 December 2015; pp. 451–454. [CrossRef]
6. Raoux, S.; Wełnic, W.; Ielmini, D. Phase change materials and their application to nonvolatile memories. *Chem. Rev.* **2010**, *110*, 240–267. [CrossRef] [PubMed]
7. Burr, G.W.; Shelby, R.M.; Sidler, S.; Di Nolfo, C.; Jang, J.; Boybat, I.; Shenoy, R.S.; Narayanan, P.; Virwani, K.; Giacometti, E.U.; et al. Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices* **2015**, *62*, 3498–3507. [CrossRef]
8. Merrikh-Bayat, F.; Guo, X.; Om'Mani, H.; Do, N.; Likharev, K.K.; Strukov, D.B. Redesigning commercial floating-gate memory for analog computing applications. In Proceedings of the 2015 IEEE International Symposium on Circuits and Systems, Lisbon, Portugal, 24–27 May 2015; pp. 1921–1924. [CrossRef]
9. Guo, X.; Merrikh-Bayat, F.; Bavandpour, M.; Klachko, M.; Mahmoodi, M.; Prezioso, M.; Likharev, K.; Strukov, D. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR Flash memory technology. In Proceedings of the 2017 IEEE International Electron Devices Meeting, San Francisco, CA, USA, 2–6 December 2017; pp. 151–154. [CrossRef]
10. Guo, X.; Bayat, F.M.; Prezioso, M.; Chen, Y.; Nguyen, B.; Do, N.; Strukov, D.B. Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR Flash memory cells. In Proceedings of the 2017 IEEE Custom Integrated Circuits Conference, Austin, TX, USA, 30 April–3 May 2017; pp. 1–4. [CrossRef]
11. Malavena, G.; Sottocornola Spinelli, A.; Monzio Compagnoni, C. Implementing spike-timing-dependent plasticity and unsupervised learning in a mainstream NOR Flash memory array. In Proceedings of the 2018 IEEE International Electron Devices Meeting, San Francisco, CA, USA, 1–5 December 2018; pp. 35–38. [CrossRef]
12. Malavena, G.; Filippi, M.; Sottocornola Spinelli, A.; Monzio Compagnoni, C. Unsupervised learning by spike-timing-dependent plasticity in a mainstream NOR Flash memory array—Part I: Cell operation. *IEEE Trans. Electron Devices* **2019**, *66*, 4727–4732. [CrossRef]
13. Malavena, G.; Filippi, M.; Sottocornola Spinelli, A.; Monzio Compagnoni, C. Unsupervised learning by spike-timing-dependent plasticity in a mainstream NOR Flash memory array—Part II: Array learning. *IEEE Trans. Electron Devices* **2019**, *66*, 4733–4738. [CrossRef]
14. Lee, S.T.; Lim, S.; Choi, N.; Bae, J.H.; Kim, C.H.; Lee, S.; Lee, D.H.; Lee, T.; Chung, S.; Park, B.G.; et al. Neuromorphic technology based on charge storage memory devices. In Proceedings of the 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 18–22 June 2018; pp. 169–170. [CrossRef]
15. Lee, S.T.; Lim, S.; Choi, N.Y.; Bae, J.H.; Kwon, D.; Park, B.G.; Lee, J.H. Operation scheme of multi-layer neural networks using NAND Flash memory as high-density synaptic devices. *IEEE J. Electron Devices Soc.* **2019**, *7*, 1085–1093. [CrossRef]
16. Lee, S.T.; Lee, J.H. Neuromorphic computing using NAND Flash memory architecture with pulse width modulation scheme. *Front. Neurosci.* **2020**, *14*, 945. [CrossRef]
17. Milo, V.; Malavena, G.; Monzio Compagnoni, C.; Ielmini, D. Memristive and CMOS devices for neuromorphic computing. *Materials* **2020**, *13*, 166. [CrossRef]
18. Kim, H.; Park, J.; Kwon, M.W.; Lee, J.H.; Park, B.G. Silicon-based floating-body synaptic transistor with frequency-dependent short-and long-term memories. *IEEE Electron Device Lett.* **2016**, *37*, 249–252. [CrossRef]
19. Kim, H.; Hwang, S.; Park, J.; Yun, S.; Lee, J.H.; Park, B.G. Spiking neural network using synaptic transistors and neuron circuits for pattern recognition with noisy images. *IEEE Electron Device Lett.* **2018**, *39*, 630–633. [CrossRef]
20. Kim, C.H.; Lee, S.; Woo, S.Y.; Kang, W.M.; Lim, S.; Bae, J.H.; Kim, J.; Lee, J.H. Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR Flash memory array. *IEEE Trans. Electron Devices* **2018**, *65*, 1774–1780. [CrossRef]
21. Nielsen, M.A. Neural Networks and Deep Learning. Determination Press. 2015. Available online: http://neuralnetworksanddeeplearning.com/ (accessed on 12 November 2021).
22. Ielmini, D.; Lacaita, A.L.; Mantegazza, D. Recovery and drift dynamics of resistance and threshold voltages in phase-change memories. *IEEE Trans. Electron Devices* **2007**, *54*, 308–315. [CrossRef]

23. Nandakumar, S.; Boybat, I.; Han, J.P.; Ambrogio, S.; Adusumilli, P.; Bruce, R.L.; BrightSky, M.; Rasch, M.; Le Gallo, M.; Sebastian, A. Precision of synaptic weights programmed in phase-change memory devices for deep learning inference. In Proceedings of the 2020 IEEE International Electron Devices Meeting, San Francisco, CA, USA, 12–18 December 2020; pp. 29.4.1–29.4.4. [CrossRef]
24. Joshi, V.; Le Gallo, M.; Haefeli, S.; Boybat, I.; Nandakumar, S.R.; Piveteau, C.; Dazzi, M.; Rajendran, B.; Sebastian, A.; Eleftheriou, E. Accurate deep neural network inference using computational phase-change memory. *Nat. Commun.* **2020**, *11*, 2473. [CrossRef]
25. Kariyappa, S.; Tsai, H.; Spoon, K.; Ambrogio, S.; Narayanan, P.; Mackin, C.; Chen, A.; Qureshi, M.; Burr, G.W. Noise-Resilient DNN: Tolerating Noise in PCM-Based AI Accelerators via Noise-Aware Training. *IEEE Trans. Electron Devices* **2021**, *68*, 4356–4362. [CrossRef]
26. Long, Y.; She, X.; Mukhopadhyay, S. Design of reliable DNN accelerator with un-reliable ReRAM. In Proceedings of the 2019 IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, 25–29 March 2019; pp. 1769–1774. [CrossRef]
27. He, Z.; Lin, J.; Ewetz, R.; Yuan, J.S.; Fan, D. Noise injection adaption: End-to-End ReRAM crossbar non-ideal effect adaption for neural network mapping. In Proceedings of the 56th Annual Design Automation Conference 2019, Las Vegas, NV, USA, 2–6 June 2019; pp. 1–6. [CrossRef]
28. Zheng, Q.; Kang, J.; Wang, Z.; Cai, Y.; Huang, R.; Li, B.; Chen, Y.; Li, H. Enhance the robustness to time dependent variability of ReRAM-based neuromorphic computing systems with regularization and 2R synapse. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems, Sapporo, Japan, 26–29 May 2019; pp. 1–5. [CrossRef]
29. Malavena, G.; Petrò, S.; Sottocornola Spinelli, A.; Monzio Compagnoni, C. Impact of program accuracy and random telegraph noise on the performance of a NOR Flash-based neuromorphic classifier. In Proceedings of the IEEE 2019 European Solid-State Device Research Conference, Cracow, Poland, 23–26 September 2019; pp. 122–125. [CrossRef]
30. Diorio, C.; Hasler, P.; Minch, A.; Mead, C.A. A single-transistor silicon synapse. *IEEE Trans. Electron Devices* **1996**, *43*, 1972–1980. [CrossRef]
31. Diorio, C.; Hasler, P.; Minch, B.A.; Mead, C.A. A floating-gate MOS learning array with locally computed weight updates. *IEEE Trans. Electron Devices* **1997**, *44*, 2281–2289. [CrossRef]
32. Monzio Compagnoni, C.; Sottocornola Spinelli, A.; Gusmeroli, R.; Beltrami, S.; Ghetti, A.; Visconti, A. Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics. *IEEE Trans. Electron Devices* **2008**, *55*, 2695–2702. [CrossRef]
33. Boccaccio, C. Embedded 1T Flash NOR: Still alive at 40 nm. And beyond? In Proceedings of the Leti Memory Workshop, Grenoble, France, 25–28 June 2013.
34. Calligaro, C.; Manstretta, A.; Modelli, A.; Torelli, G. Technological and design constraints for multilevel Flash memories. In Proceedings of the IEEE International Conference on Electronics, Circuits, and Systems, Rhodes, Greece, 16 October 1996; Volume 2, pp. 1005–1008. [CrossRef]
35. Monzio Compagnoni, C.; Chiavarone, L.; Calabrese, M.; Ghidotti, M.; Lacaita, A.L.; Spinelli, A.S.; Visconti, A. Fundamental limitations to the width of the programmed $V_T$ distribution of NOR Flash memories. *IEEE Trans. Electron Devices* **2010**, *57*, 1761–1767. [CrossRef]
36. Goda, A.; Miccoli, C.; Monzio Compagnoni, C. Time dependent threshold-voltage fluctuations in NAND Flash memories: From basic physics to impact on array operation. In Proceedings of the 2015 IEEE International Electron Devices Meeting, Washington, DC, USA, 7–9 December 2015; pp. 374–377. [CrossRef]
37. Monzio Compagnoni, C.; Gusmeroli, R.; Spinelli, A.S.; Visconti, A. Analytical model for the electron-injection statistics during programming of nanoscale NAND Flash memories. *IEEE Trans. Electron Devices* **2008**, *55*, 3192–3199. [CrossRef]
38. Ghetti, A.; Monzio Compagnoni, C.; Sottocornola Spinelli, A.; Visconti, A. Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer Flash memories. *IEEE Trans. Electron Devices* **2009**, *56*, 1746–1752. [CrossRef]
39. Sottocornola Spinelli, A.; Monzio Compagnoni, C.; Gusmeroli, R.; Ghidotti, M.; Visconti, A. Investigation of the random telegraph noise instability in scaled Flash memory arrays. *Jpn. J. Appl. Phys.* **2008**, *47*, 2598. [CrossRef]
40. Adamu-Lema, F.; Monzio Compagnoni, C.; Amoroso, S.M.; Castellani, N.; Gerrer, L.; Markov, S.; Spinelli, A.S.; Lacaita, A.L.; Asenov, A. Accuracy and issues of the spectroscopic analysis of RTN traps in nanoscale MOSFETs. *IEEE Trans. Electron Devices* **2012**, *60*, 833–839. [CrossRef]
41. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
42. Monzio Compagnoni, C.; Gusmeroli, R.; Spinelli, A.S.; Lacaita, A.L.; Bonanomi, M.; Visconti, A. Statistical model for random telegraph noise in Flash memories. *IEEE Trans. Electron Devices* **2007**, *55*, 388–395. [CrossRef]
43. Ghetti, A.; Amoroso, S.M.; Mauri, A.; Monzio Compagnoni, C. Impact of nonuniform doping on random telegraph noise in Flash memory devices. *IEEE Trans. Electron Devices* **2011**, *59*, 309–315. [CrossRef]
44. Amoroso, S.M.; Monzio Compagnoni, C.; Ghetti, A.; Gerrer, L.; Sottocornola Spinelli, A.S.; Lacaita, A.L.; Asenov, A. Investigation of the RTN distribution of nanoscale MOS devices from subthreshold to on-state. *IEEE Electron Device Lett.* **2013**, *34*, 683–685. [CrossRef]
45. Miccoli, C.; Paolucci, G.M.; Monzio Compagnoni, C.; Sottocornola Spinelli, A.S.; Goda, A. Cycling pattern and read/bake conditions dependence of random telegraph noise in decananometer NAND Flash arrays. In Proceedings of the 2015 IEEE International Reliability Physics Symposium, Monterey, CA, USA, 19–23 April 2015; pp. MY.9.1–MY.9.6. [CrossRef]