scientific reports

Check for updates

OPEN

Binarized neural network of diode array with high concordance to vector–matrix multiplication

Yunwoo Shin, Kyoungah Cho & Sangsig Kim

In this study, a binarized neural network (BNN) of silicon diode arrays achieved vector–matrix multiplication (VMM) between the binarized weights and inputs in these arrays. The diodes that operate in a positive-feedback loop in their p^+-n-p-n^+ device structure possess steep switching and bistable characteristics with an extremely low subthreshold swing (below 1 mV) and a high current ratio (approximately 10^8). Moreover, the arrays show a self-rectifying functionality and an outstanding linearity by an R-squared value of 0.99986, which allows to compose a synaptic cell with a single diode. A 2 × 2 diode array can perform matrix multiply-accumulate operations for various binarized weight matrix cases with some input vectors, which is in high concordance with the VMM, owing to the high reliability and uniformity of the diodes. Moreover, the disturbance-free, nondestructive readout, and semi-permanent holding characteristics of the diode arrays support the feasibility of implementing the BNN.

Keywords Binarized neural network, Multiply-accumulate, Vector–matrix multiplication, Gated p^+-n-p-n^+ diodes, Positive-feedback loop mechanism

With the explosive growth of data, brain-inspired (i.e. neuromorphic) computing systems have achieved significant improvements in parallel computing and efficient data processing[1–5]. In particular, binarized neural networks (BNNs) have recently demonstrated their capabilities in image recognition applications[6–10]. The accelerators in the BNNs perform a matrix "multiply accumulate" (MAC) operation (i.e. vector–matrix multiplication (VMM)) between the binarized weights and analog inputs[6,7]. The bitwise MAC operation enables extensive applicability to resource-constrained platforms, such as edge devices and mobile processors, promising a considerable reduction in memory (approximately 32 ×) and computation (approximately 2 ×) requirements compared with other neural networks[6–10]. Furthermore, it is still difficult for emerging synaptic devices to fully implement analog neural networks (analog input and analog weight) with nonlinear conductance changes and device variations[1–4,11–13]. However, digital synaptic devices are suitable for implementing BNNs because of their binarized weights[13–16].

Various memory devices, such as resistive random-access memories (RRAMs), magnetoresistive RAMs (MRAM), flash-based memory devices, static RAMs (SRAM), and dynamic RAMs (DRAMs), have been widely researched for BNN implementation[11–21]. Their definite bistable characteristics, with a high on/off ratio and input/output linearity, are required for the implementation[13–15]. However, despite their improved characteristics, these devices still do not satisfactorily meet this requirement[13–17]. Moreover, as the number of hidden layers in the multilayer perceptron structure of neural networks increases, the neuron circuits that interconnect the pre- and post-layers become more complicated, which degrades the area and power efficiencies of computing systems based on neural networks[21–24]. The activation function of neuronal circuits is simplified or merged into a memory array to alleviate degradation[25–27].

Recently, silicon diodes operating in a positive-feedback loop mechanism have demonstrated their feasibility as neuromorphic devices, owing to their superior electrical characteristics, high reliability, and reproducibility[28–32]. The positive-feedback loop mechanism enables diodes to exhibit steep switching and bistable characteristics with a high ON/OFF ratio[32]. Compared to other memory devices used as components of emerging BNNs[11–21], the p^+-n-p-n^+ diodes operating by the positive-feedback loop mechanisms have a significant advantage in terms of reliability and endurance since the relatively low operating voltage prevents the deterioration of devices under sustained operations. Furthermore, the BNN comprising the p^+-n-p-n^+ diodes can be more simplified than those designed by SRAMs and DRAMs[18,33]. In this study, we demonstrate BNNs with 4 × 1 and 2 × 2 arrays, consisting of p^+-n-p-n^+ diodes.

Department of Electrical Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea. email: sangsig@korea.ac.kr

Results

Optical images of the diode array and the p^+-n-p-n^+ diode in this array are shown in Fig. 1a. The bistable characteristic principle of the diode is illustrated in Fig. 1b with a diode schematic, energy-band diagrams in State 0 and State 1, and the red (blue) circuit symbol of State 0 (State 1). For the diode in the array, the gate voltage (V_{Gate}) controls the injection and accumulation of charge carriers by modulating the potential barrier in the n-doped region of the energy-band diagram. In State 0, excess charge carriers are absent in the potential wells in the n- and p-doped regions of the energy-band diagram, preventing the diode current (I_{Diode}) from flowing in the diode. In contrast, in State 1, excess charge carriers accumulate in the potential wells in the n- and p-doped

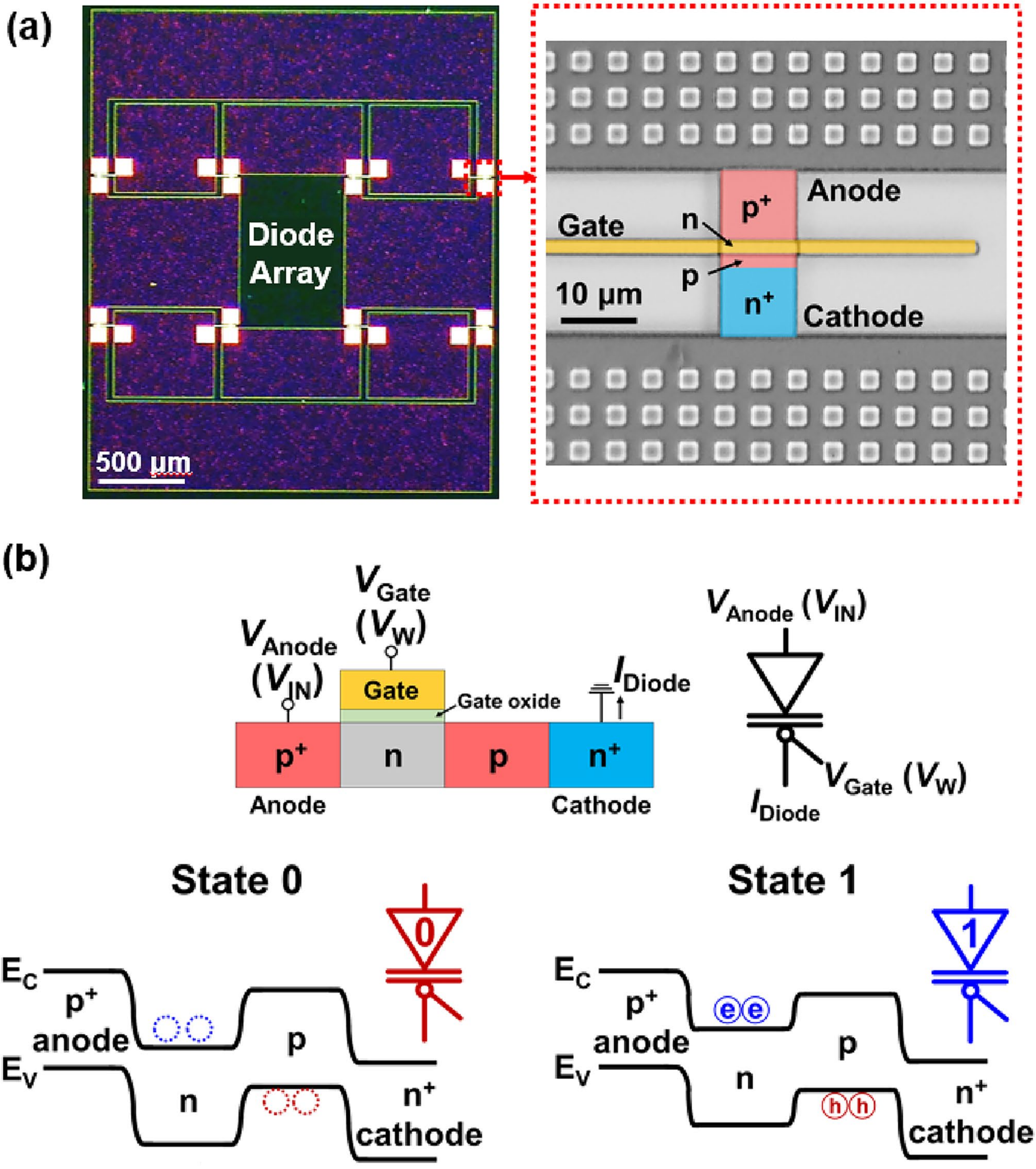

Figure 1. Optical images of (**a**) diode array with p^+-n-p-n^+ diode, and (**b**) diode schematic, energy-band diagrams in States 0 and 1, and red (blue) circuit symbol of State 0 (State 1). The n-doped region in the optical image is present beneath the yellow-colored gate electrode.

regions in the energy-band diagram, resulting in the I_{Diode} to flow in the diode. Modulation of the potential barrier allowed the diode to exhibit bistable characteristics. However, for the BNN operation, the conductance in State 0 (State 1) corresponds to Weight 0 (1), and the anode voltage (V_{Anode}) and V_{Gate} represent the input and weight voltages (V_{IN} and V_{W}), respectively.

Figure 2 shows the I_{Diode} versus V_{IN} curves for a diode at $V_{\mathrm{W}} = 0.0$ V and $V_{\mathrm{W}} = 1.0$ V. The BNN operating conditions are illustrated in the figure. As V_{IN} increases from − 3 to 3 V (black curves), I_{Diode} first remains at 0 mA, after which it abruptly increases at a V_{IN} of 1.5 V for $V_{\mathrm{W}} = 0.0$ V and a V_{IN} of 2.5 V for $V_{\mathrm{W}} = 1.0$ V. The generation of the positive-feedback loop leads to an abrupt increase in the I_{Diode}. As V_{IN} decreased from 3 to − 3 V (blue curves), I_{Diode} linearly decreased and reached 0 mA near $V_{\mathrm{IN}} = 1$ V; the positive-feedback loop was eliminated near $V_{\mathrm{IN}} = 1$ V. The bistable characteristics are presented in the I_{Diode} versus V_{IN} curves, and the high ratio of the current magnitudes of States 1 and 0 is approximately 10^8 at a V_{W} of 1.0 V for $V_{\mathrm{IN}} = 2.0$ V (see Supporting Information, Fig. S1). The unipolar switching (i.e. rectifying) characteristics inherit the electrical properties of the p–n diode. Regardless of V_{W}, the I_{Diode} versus V_{IN} curve shape in State 1 resembles that of the rectified linear unit function used for the activation function of neural networks, including BNNs[34–36].

In Fig. 2a, the diode state becomes State 1 (State 0) by the potentiation (depression) at a V_{IN} of 2.0 V (− 2.0 V), and V_{W} of 0.0 V. The potentiated (depressed) diode weighed 1 (0). In Fig. 2b, the potentiated or depressed diode is in a standby state at $V_{\mathrm{W}} = 1.0$ V and $V_{\mathrm{IN}} = 0.0$ V. For a V_{IN} range of 1.1 V to 2.0 V, the I_{Diode} of the potentiated diode is linearly proportional to V_{IN}, the I_{Diode} of the depressed diode remains at a low level, and the difference in the I_{Diode} of the potentiated and depressed diodes is used for the MAC operation.

To describe the BNN operation, the diode-array architecture, schematics, and circuit symbols of the p^+-n-p-n^+ diodes in the weight update, standby, and multiplication operations are shown in Fig. 3a–c. In the diode array, p^+, n^+, and gate electrodes were connected to the input lines (ILs), output lines (OLs), and weight lines (WLs), respectively. WLs and OLs were parallel to each other and perpendicular to ILs[37–39]. The multiplication of V_{IN} and the conductance (i.e. weight) of a diode using Ohm's law produced I_{Diode}, and the summation of I_{Diode} at each common OL (i.e. the same row in the diode array) using Kirchhoff's law yielded the output current (I_{OUT}). Thus, the diode-array architecture can perform parallel MAC operations. For the BNN operations, the V_{IN} for IL, V_{W} for WL, and I_{OUT} for OL signals represent the input, weight update selection, and output, respectively.

In the weight update operation in Fig. 3a, the binarized weight (**W**) matrix is updated by both V_{IN} (2.0 V or − 2.0 V) and V_{W} (0.0 V). The diodes selected in the array with $V_{\mathrm{W}} = 0.0$ V were potentiated (depressed) at V_{IN} of 2.0 V (− 2.0 V). In the standby operation shown in Fig. 3b, all V_{W} and V_{IN} are 1.0 V and 0.0 V, respectively. The diodes maintained their memory states with extremely low I_{Diode}. In the multiplication operation shown in Fig. 3c, the V_{W} (1.0 V) and V_{IN} (0.0 V or from 1.1 V to 2.0 V) are applied to the diode array. I_{Diode} represents the output of the product of V_{IN} and the binarized weight, and I_{OUT} is obtained from the summation of the I_{Diode}.

Figures 4 shows the BNN operation and I_{Diode} versus V_{IN} characteristics of the selected diode in an array. In Fig. 4a, for potentiation (depression), voltage pulses with $V_{\mathrm{IN}} = 2.0$ V (− 2.0 V) and $V_{\mathrm{W}} = 0.0$ V were applied to the diode. After the potentiation, the I_{Diode} is linearly proportional to the multipulse and staircase waveforms of V_{IN} ranging from 1.1 to 2.0 V in steps of 0.1 V. In contrast, after the depression, the I_{Diode} remained at a low level despite the same V_{IN} pulses as the potentiation. During the standby operation between the V_{IN} pulses, the diode memorized the weight with a power consumption of approximately 0 W. In Fig. 4b, the extracted I_{Diode} values of the potentiated diode for each V_{IN} pulse can be expressed as a linear function of V_{IN}, as follows:

$$I_{\mathrm{Diode}}(V_{\mathrm{IN}}) = 7.4 \times (V_{\mathrm{IN}} - 1). \quad (1)$$

Here, the coefficient 7.4 represents the conductance of the potentiated diode (i.e. Weight 1). For the depressed diode (i.e. Weight 0), the conductance is close to 0, and the I_{Diode} remains at a low level during multiplication after depression. The constant term − 1 indicates the lower limit of the V_{IN} range for diode multiplication. This expression implies that the single diode performs multiplication between V_{IN} and weight with outstanding linearity.

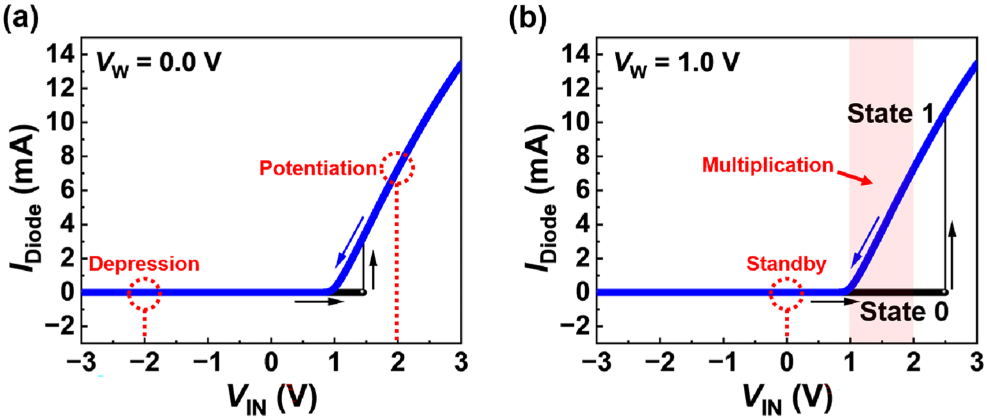

Figure 2. I_{Diode} vs. V_{IN} curves for a diode at (**a**) $V_{\mathrm{W}} = 0.0$ V and (**b**) $V_{\mathrm{W}} = 1.0$ V. The black and blue curves indicate the forward and reverse sweeps of V_{IN}, respectively. The BNN operating conditions are shown in the figures.

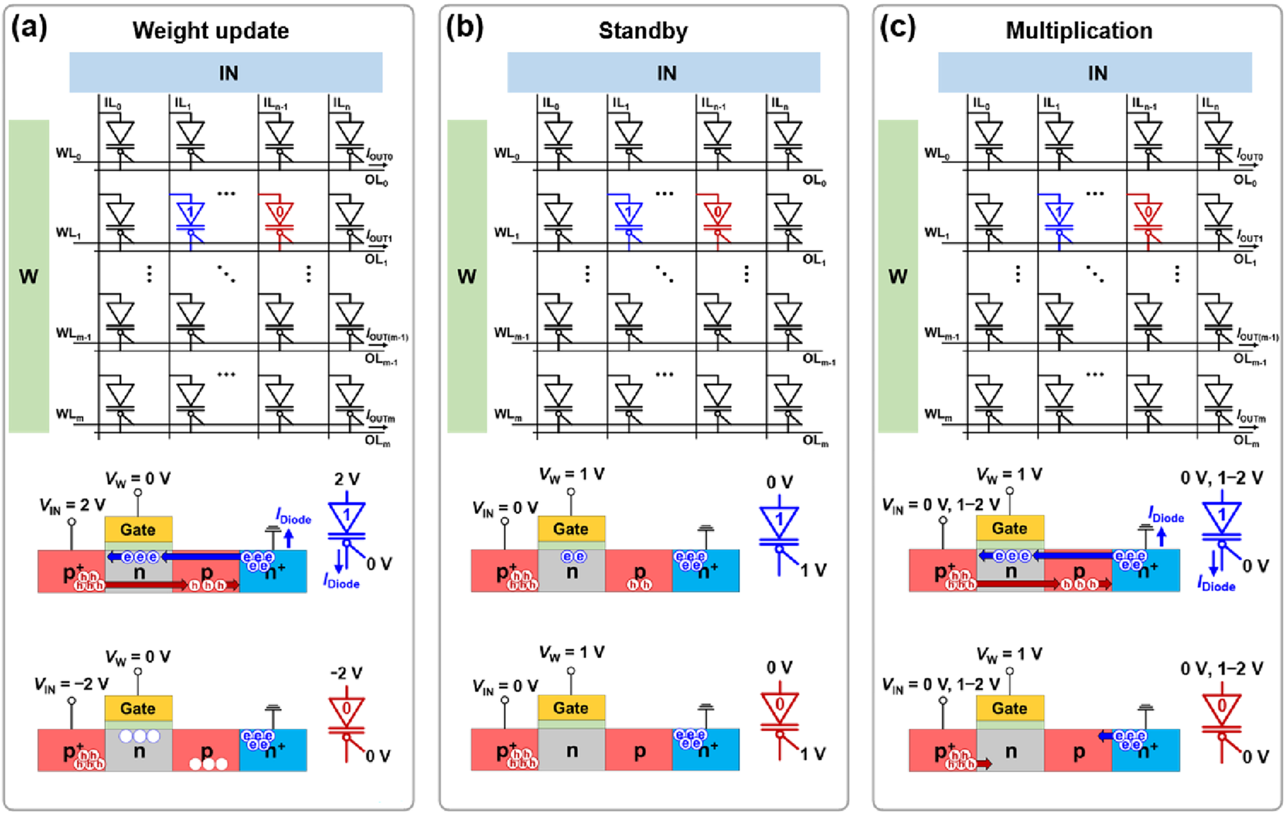

Figure 3. Diode-array architecture, schematics, and circuit symbols of diodes in (**a**) weight update, (**b**) standby, and (**c**) multiplication operations. Colored boxes with W and IN mean the peripheral parts of the weight and input lines (WLs and ILs), respectively.

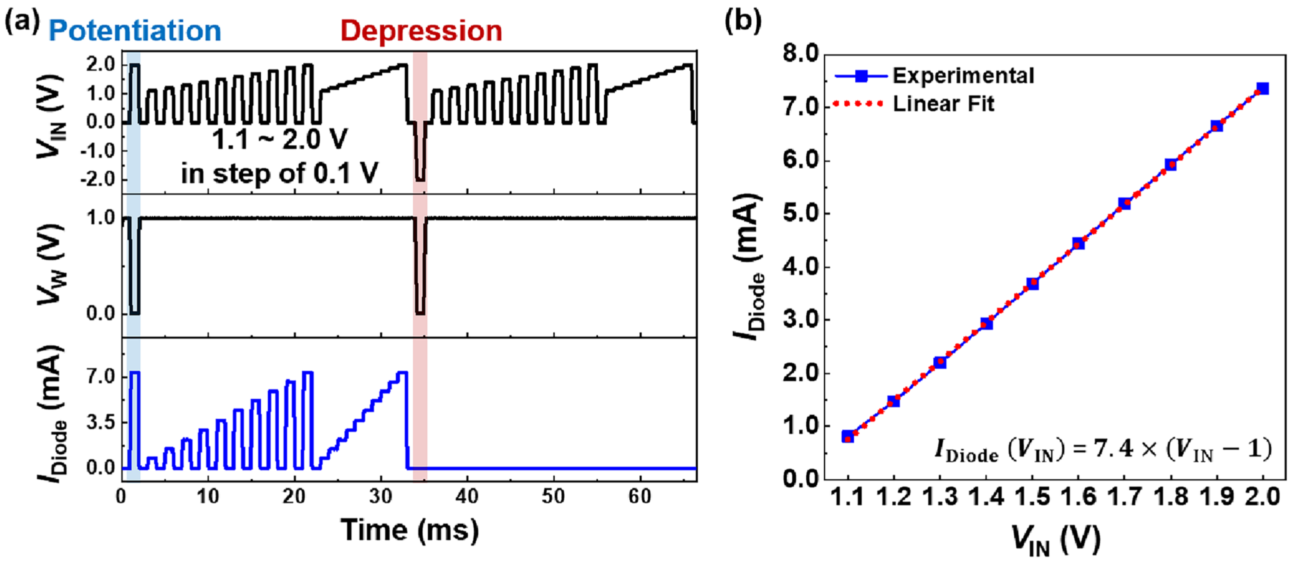

Figure 4. BNN operation of a diode. (**a**) I_{Diode} for potentiated and depressed diodes and (**b**) I_{Diode} vs. V_{IN} diode characteristics of potentiated diode. Here, the R-square value of the linear fit is 0.99986, which means that the diode features high linearity.

The bitwise MAC operation of the BNN in a 4 × 1 diode array is illustrated in Fig. 5. The MAC operation was conducted by multiplying the input V_{INn} and binarized weight W_n for each n^{th} diode. In a 4 × 1 diode array in Fig. 5a, I_{OUT} can be expressed as

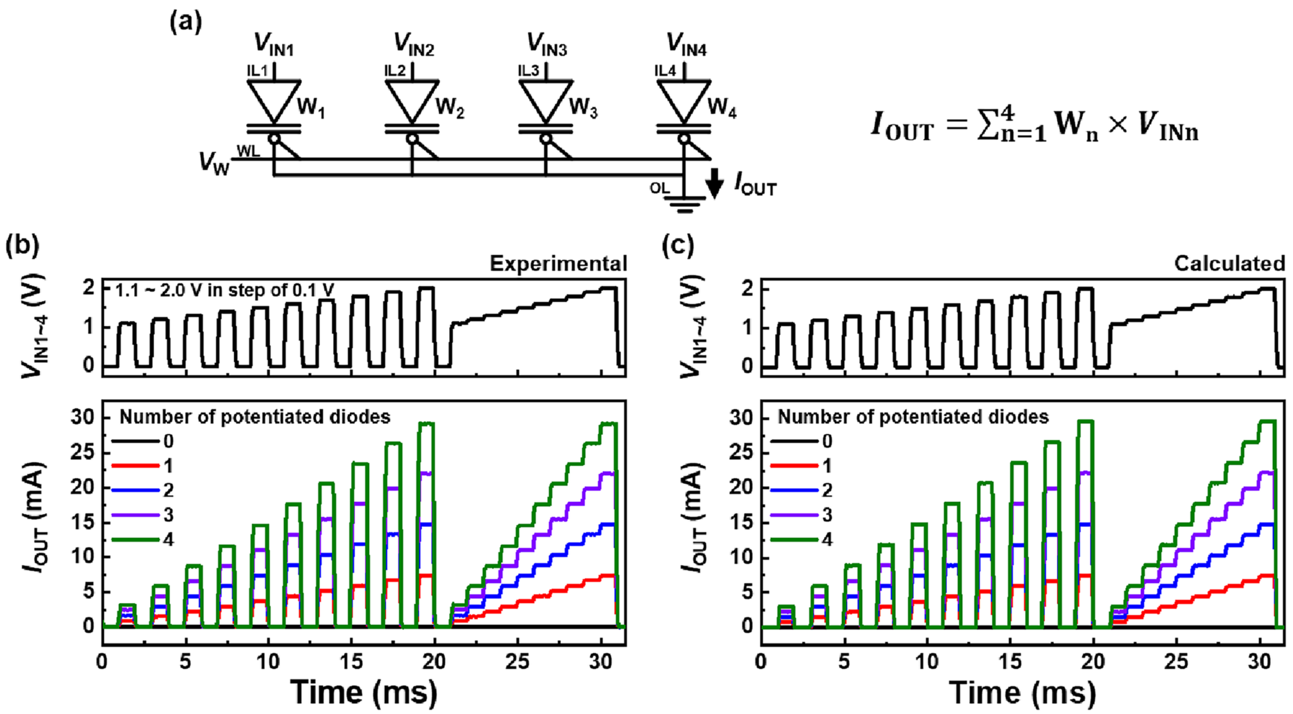

Figure 5. BNN operation of 4 × 1 diode array. (**a**) Circuit diagram of array, (**b**) experimental data of bitwise MAC operation, and (**c**) calculated data of bitwise MAC operation. Here, all V_{IN} (V_{IN1}–V_{IN4}) are identical.

$$I_{OUT} = \sum_{n=1}^{4} W_n \times V_{INn}. \tag{2}$$

The experimental data for the MAC operation are shown in Fig. 5b. All V_{IN} were applied in the same waveforms as those in Fig. 4a. W_n is updated before the MAC operation, and I_{OUT} remains linearly proportional to the multipulse and staircase waveforms of V_{IN}. Moreover, I_{OUT} is linearly proportional to the number of potentiated diodes for the same V_{IN} pulse, and there are few electrical variations between the component diodes in the array, owing to the wafer-scale full CMOS process. Here, the product of V_{Inn} and W_n can be replaced with the I_{Diode} in Eq. (1). Thus, Eq. (2) becomes

$$I_{OUT} = I_{Diode} \times n, \text{where } n = \#\text{of potentiated diodes}. \tag{3}$$

The MAC operation data (obtained from Eq. (3)) plotted in Fig. 5c were in good agreement with the experimental data in Fig. 5b. The uniformity of the component diodes in the array can minimize the BNN computational errors[39–4142]. Thus, a massive array architecture comprising gated p^+-n-p-n^+ diodes can be realized with a high ratio of the bistable current magnitude to uniformity.

A 2 × 2 diode array is used to perform the matrix MAC operation of the BNN, as shown in Fig. 6. The binarized weights of the diodes in Fig. 6a are the components of the **W** matrix $\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$, where I_{OUT1} ($I_{Diode} \times n = W_{11} \times V_{IN1} + W_{12} \times V_{IN2}$) and I_{OUT2} ($I_{Diode} \times m = W_{21} \times V_{IN1} + W_{22} \times V_{IN2}$) are obtained via the OLs, and n(m) is the number of potentiated diodes in the first (second) row in the array. In this array, the IN-vector component IN_1 or IN_2 is 0 (if V_{IN} is 0.0 V) or 1 (if V_{IN} is from 1.1 V to 2.0 V), and the OUT-vector component OUT_1 (OUT_2) is n (m). Thus, the matrix MAC operation can be expressed as $\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} IN_1 \\ IN_2 \end{bmatrix} = \begin{bmatrix} OUT_1 \\ OUT_2 \end{bmatrix}$, which represents the VMM between the weight matrix and input vector, $\mathbf{W} \cdot \mathbf{IN}^T = \mathbf{OUT}^T$.

The voltage pulses of V_{IN} (2.0 V or – 2.0 V) and V_W (0 V) are applied to the diode array to update the **W** matrix, as shown in Fig. 6a. In Fig. 6b, the 2 × 2 diode array performs the matrix MAC operation in the **W** = $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ case among the 16 different **W** cases (see Fig. 7 for other **W** cases). **W** is updated to the **W** = $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ case, after which V_{IN1} and V_{IN2} are applied to the same waveforms depicted in Fig. 4a for the matrix MAC operations. Thus, I_{OUT1} was twice as large as I_{OUT2}. Consequently, the diode array exhibits a simplified VMM of $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

In Fig. 7, we demonstrate the matrix MAC operations of the 2 × 2 diode array for six different **W** values. The IN-vector components IN_1 and IN_2 are both 1, where I_{OUT1} and I_{OUT2} correspond to the outputs OUT_1 and OUT_2 of the matrix MAC operation for each **W** case, respectively. Here, V_{IN1} and V_{IN2} are applied to the same

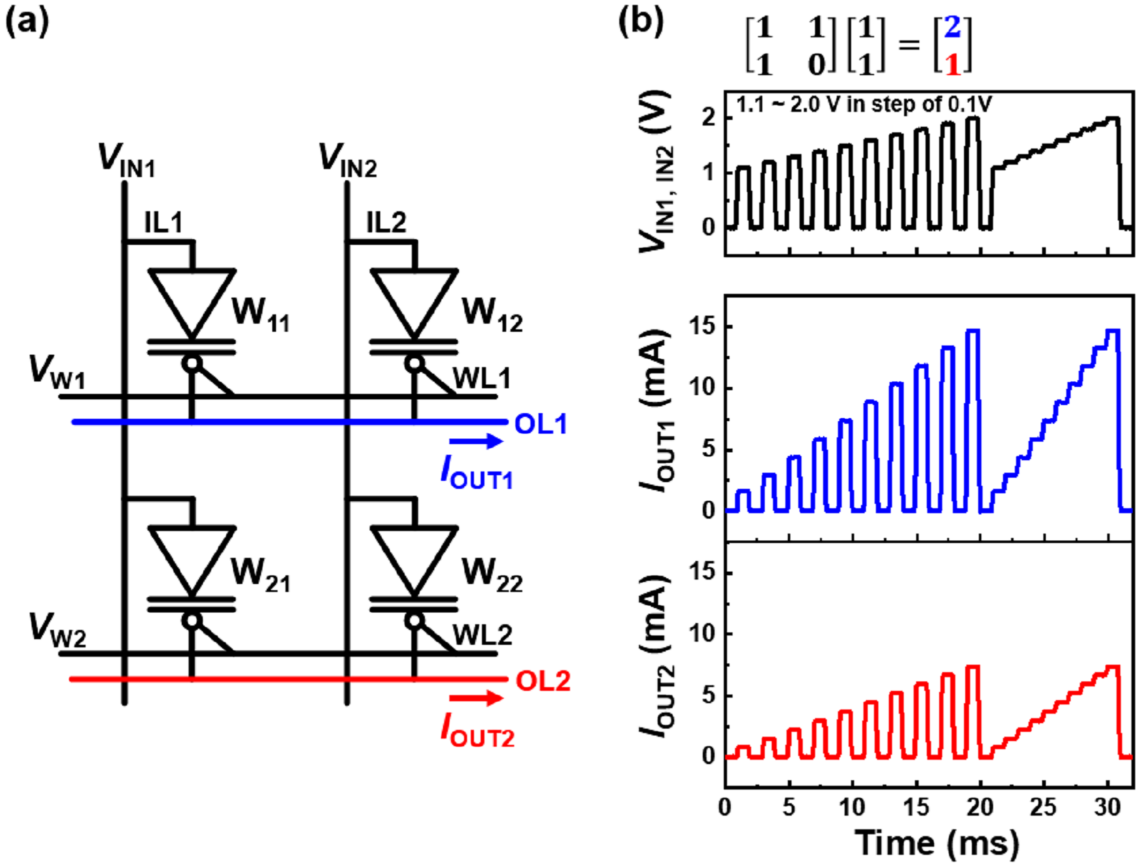

Figure 6. BNN operation of 2 × 2 diode array. (**a**) Circuit diagram of array and (**b**) matrix MAC operation with fixed **W** matrix and **IN** vector. Here, V_{IN1} and V_{IN2} are identical.

waveforms as those in Fig. 4a. I_{OUT1} and I_{OUT2} in Fig. 7e are identical to those in Fig. 6c. For each **W** case, the output of the matrix MAC operation has a high concordance with the simplified VMM equation in each figure owing to the high uniformity of the diodes.

For the **W** matrix $\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}$ in Fig. 8a, the OUT-vector components OUT_1 and OUT_2 vary with the IN-vector components IN_1 and IN_2, which are highly concordant with the simplified VMMs in Fig. 8b. I_{OUT1} and I_{OUT2} depend on the number of potentiated diodes in the array and the multipulse and staircase waveforms of V_{IN1} and V_{IN2}. After multiplication, the diode array maintains the binarized weight matrix owing to the nondestructive readout characteristics of the component diodes[43]. Instead, the component diodes refresh their weights during multiplication operations. They also exhibit superior electrical stability against bias stresses (continuous input pulses), unlike other memory devices[13,19–21,44,45]. The diode array holds the binarized weight matrix semi-permanently using refresh operations while performing multiplication operations. Moreover, the binarized weights are not updated unless V_{IN} and V_W are applied simultaneously. The binarized weight matrix did not change for multiplication. However, for the weight update operations, the matrix is determined by V_{W1} and V_{W2}.

Discussion

We demonstrated the BNN operation of an array composed of p^+-n-p-n^+ diodes with bistable characteristics. The component diode exhibited inherent unipolar switching characteristics, and the array was immune to sneak path problems, unlike previously proposed BNNs[11–13,16,43]. In addition to the outstanding bistable characteristics with a high current ratio (approximately 10^8), the rectifying characteristics can simplify peripheral neuron circuits. Moreover, the simplified peripheral neuron circuits help to reduce the burden of area and power consumption for BNN-based computing systems. The diode array performed matrix MAC operations with high concordance with the VMM between the binarized weight matrix and input vector. Furthermore, the diode array exhibited nondestructive readout and semi-permanent holding characteristics. Our diodes can realize the compact synaptic array in which the cell is composed of a single device with $6F^2$ (F = feature size) as well as operate at relatively low voltages (≤ 2 V). Consequently, p^+-n-p-n^+ diodes are the most suitable building blocks for area-/energy-efficient and reliable synaptic arrays in BNN.

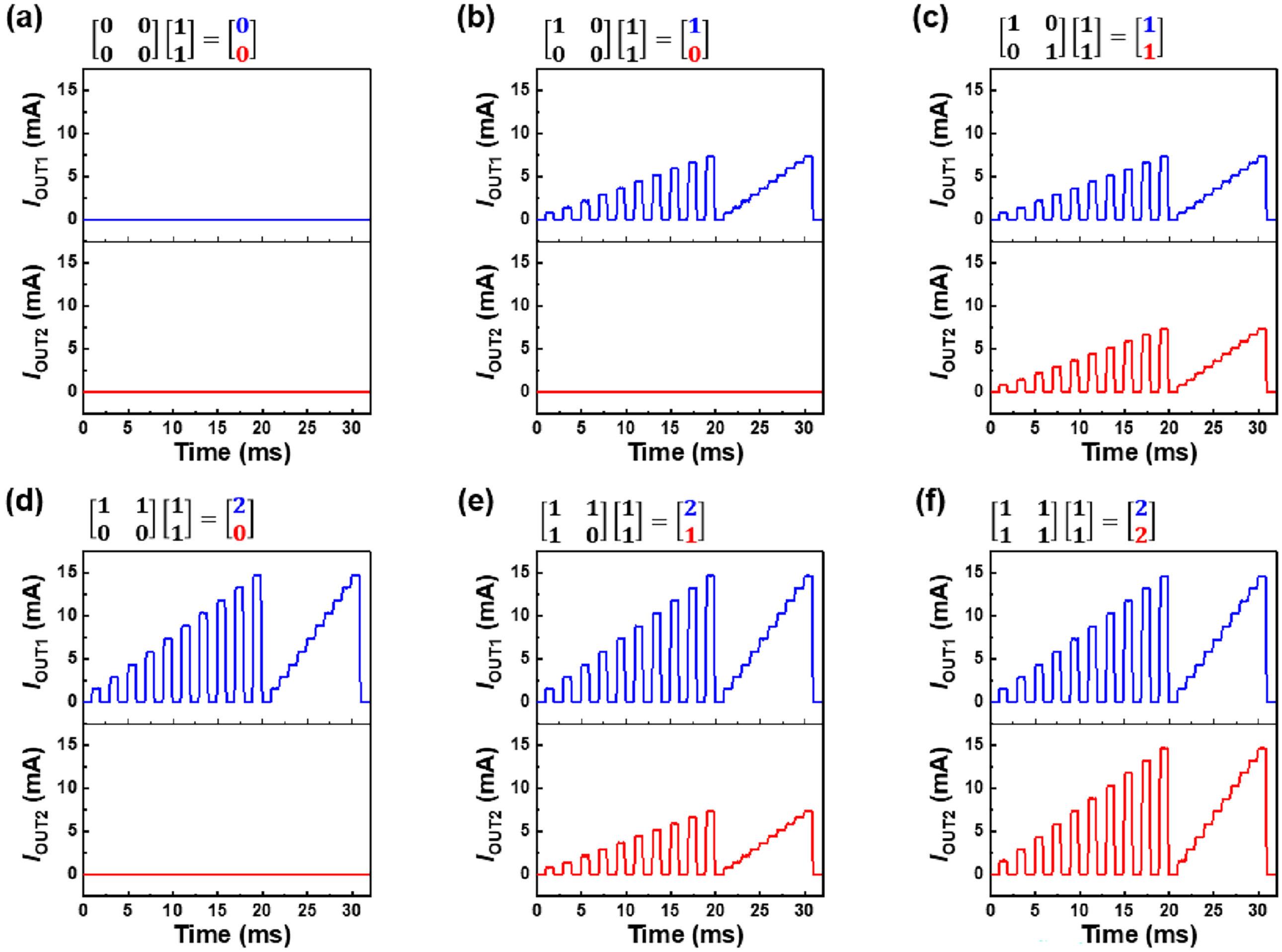

Figure 7. Matrix MAC operations of 2 × 2 diode array for (**a**–**f**) six different **W** cases with fixed **IN** vector. The other cases are excluded because their results of the matrix MAC operations are the same as these six cases.

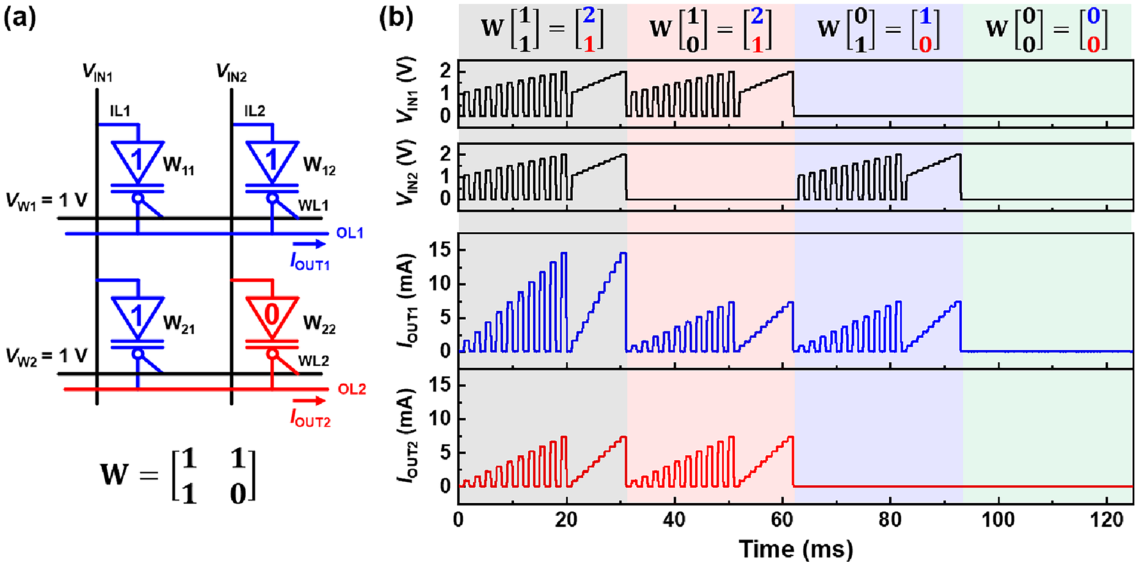

Figure 8. Matrix MAC operations of 2 × 2 diode array with fixed **W** matrix. (**a**) Circuit diagram of array and (**b**) matrix MAC operations for four different **IN** vectors. The **W** matrix is updated before the matrix MAC operations.

Methods

Device fabrication

The p^+-n-p-n^+ diodes were fabricated using full CMOS processes from a p-type (100)-oriented silicon-on-insulator wafer with a 340-nm thick top Si layer on a 2-μm thick buried oxide layer. First, the wafer was cleaned by standard clean-1 and 50:1 diluted HF to remove particles or impurities. A 10-nm thick SiO_2 layer was grown on the Si surface as a sacrificial layer to minimize channeling and damage to the channel during ion implantation by dry oxidation (800 °C, 200 min). The n-doped regions were formed using a conventional ion implantation process, in which P^+ ions were implanted at a dose of 3×10^{13} cm^{-2} with an ion energy of 60 keV. Subsequently, the wafer was annealed at 1100 °C for 30 min under ambient nitrogen. The 25-nm thick SiO_2 layers were thermally grown as the gate dielectric for the gated p^+-n-p-n^+ diodes by dry oxidation (850 °C, 270 min). A 400-nm thick and 1.5-μm wide poly-Si gate electrode was formed by low-pressure chemical vapor deposition (LPCVD), followed by photolithography, and a dry etching process was used to delineate the gate profile. The lightly doped drain extension regions were implanted with BF_2^+ ions at a dose of 1×10^{12} cm^{-2} and an ion energy of 10 keV. And then LPCVD-based tetraethyl orthosilicate (TEOS) was used to form the 200-nm thick gate sidewall spacers. For the p-doped regions, BF_2^+ ions at a dose of 3×10^{13} cm^{-2} were implanted at an ion energy of 40 keV. The n^+ cathode, heavily n-doped poly-Si gate, and p^+ anode regions were formed by masked ion implantation (P^+ at 3×10^{15} cm^{-2}, 50 keV and BF_2^+ at 3×10^{15} cm^{-2}, 30 keV). Thereafter, the wafer was annealed at 1000 °C for 30 min in ambient nitrogen, followed by rapid thermal annealing at 1050 °C for 30 s to activate the implanted dopants with uniform diffusion and eliminate defects. The dual-step annealing process was carried out to form the p^+-n-p-n^+ doping structure. Finally, a Ti/TiN/Al/TiN metal alloy was deposited in the drain, source, and gate contact regions after deposition of a 700-nm thick interlayer dielectric layer (for device protection) using the LPCVD-based TEOS.

Measurement

All electrical data were acquired at room temperature using a semiconductor parameter analyzer (HP4155C, Agilent Technologies) and Keithley 2636A and 2636 B source meters. The input-pulse width for the diode and MAC array operations was selected as 1 ms because of the limitations of the Keithley source meters. The experimental set-up to examine the BNN operations of the diode array is shown in Fig. S2 of supporting information.

Data availability

All data generated during this study are included in this published article (and its Supporting Information files).

Received: 18 November 2023; Accepted: 8 March 2024
Published online: 11 March 2024

References

1. Zhang, W. *et al.* Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
2. Hassanpour, M., Riera, M. & González, A. A Survey of near-data processing architectures for neural networks. *Mach. Learn. Knowl. Extr.* **4**, 66–102 (2022).
3. Wang, Y. *et al.* An in-memory computing architecture based on two-dimensional semiconductors for multiply-accumulate operations. *Nat. Commun.* **12**, 3347 (2021).
4. Sun, Z. *et al.* Solving matrix equations in one step with cross-point resistive arrays. *Proc. Natl. Acad. Sci.* **116**, 4123–4128 (2019).
5. Zidan, M. A. *et al.* A general memristor-based partial differential equation solver. *Nat. Electron.* **1**, 411–420 (2018).
6. Simons, T. & Lee, D.-J. A review of binarized neural networks. *Electronics* **8**, 661 (2019).
7. Yuan, C. & Agaian, S. S. A comprehensive review of binary neural network. *Artificial Intelligence Review*, 1–65 (2023).
8. Rastegari, M., Ordonez, V., Redmon, J. & Farhadi, A. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV*, 525–542 (Springer).
9. Qin, H. *et al.* Binary neural networks: A survey. *Pattern Recognit.* **105**, 107281 (2020).
10. Zhang, L., Tang, X., Hu, X., Zhou, T. & Peng, Y. FPGA-based BNN architecture in time domain with low storage and power consumption. *Electronics* **11**, 1421 (2022).
11. Zanotti, T., Puglisi, F. M. & Pavan, P. Energy-efficient non-von neumann computing architecture supporting multiple computing paradigms for logic and binarized neural networks. *J. Low Power Electron. Appl.* **11**, 29 (2021).
12. Yang, X., Taylor, B., Wu, A., Chen, Y. & Chua, L. O. Research progress on memristor: From synapses to computing systems. *IEEE Trans. Circuits Syst. I Regul. Papers* **69**, 1845–1857 (2022).
13. Kim, S. *et al.* Binarized neural network with Silicon nanosheet synaptic transistors for supervised pattern classification. *Sci. Rep.* **9**, 11705 (2019).
14. Ahn, D., Oh, H., Kim, H., Kim, Y. & Kim, J.-J. Maximizing parallel activation of word-lines in MRAM-based binary neural network accelerators. *IEEE Access* **9**, 141961–141969 (2021).
15. Jung, S. *et al.* A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022).
16. Zhao, Y., Wang, Y., Wang, R., Rong, Y. & Jiang, X. A highly robust binary neural network inference accelerator based on binary memristors. *Electronics* **10**, 2600 (2021).
17. Sun, X. et al. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 1423–1428 (IEEE).
18. Kim, Y., Li, S., Yadav, N. & Choi, K. K. A novel ultra-low power 8T SRAM-based compute-in-memory design for binary neural networks. *Electronics* **10**, 2181 (2021).
19. Kingra, S. K. *et al.* Dual-configuration in-memory computing bitcells using SiOx RRAM for binary neural networks. *Appl. Phys. Lett.* **120**, 034102 (2022).
20. Lee, S., Kim, H., Lee, S.-T., Park, B.-G. & Lee, J.-H. SiO2 fin-based flash synaptic cells in AND array architecture for binary neural networks. *IEEE Electron Device Lett.* **43**, 142–145 (2021).
21. Liu, R. et al. In *Proc. of the 55th Annual Design Automation Conference*, 1–6.
22. Ahn, J.-H. *et al.* On-chip adaptive matching learning with charge-trap synapse device and ReLU activation circuit. *Solid-State Electron.* **186**, 108177 (2021).
23. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).

24. Eshraghian, J. K., Wang, X. & Lu, W. D. Memristor-based binarized spiking neural networks: Challenges and applications. *IEEE Nanotechnol. Mag.* **16**, 14–23 (2022).
25. Oh, J. *et al.* Preventing vanishing gradient problem of hardware neuromorphic system by implementing Imidazole-based memristive ReLU activation neuron. *Adv. Mater.* **35**, 2300023 (2023).
26. Wang, Z. *et al.* Self-activation neural network based on self-selective memory device with rectified multilevel states. *IEEE Trans. Electron Devices* **67**, 4166–4171 (2020).
27. Huang, Y., Wang, W., Qiao, L., Hu, X. & Chu, T. Programmable low-threshold optical nonlinear activation functions for photonic neural networks. *Opt. Lett.* **47**, 1810–1813 (2022).
28. Woo, S., Cho, J., Lim, D., Cho, K. & Kim, S. Transposable 3T-SRAM synaptic array using independent double-gate feedback field-effect transistors. *IEEE Trans. Electron Devices* **66**, 4753–4758 (2019).
29. Park, Y.-S., Woo, S., Lim, D., Cho, K. & Kim, S. Integrate-and-fire neuron circuit without external bias voltages. *Front. Neurosci.* **15**, 644604 (2021).
30. Lim, D., Cho, K. & Kim, S. Single silicon neuron device enabling neuronal oscillation and stochastic dynamics. *IEEE Electron Device Lett.* **42**, 649–652 (2021).
31. Woo, S. & Kim, S. Neural oscillation of single silicon nanowire neuron device with no external bias voltage. *Sci. Rep.* **12**, 3516 (2022).
32. Lim, D., Cho, K. & Kim, S. Reconfigurable logic-in-memory using silicon transistors. *Adv. Mater. Technol.* **7**, 2101504 (2022).
33. Choi, H., Lee, Y., Kim, J.-J. & Yoo, S. *In 2020 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, 1–3.
34. Agarap, A. F. Deep learning using rectified linear units (relu). Preprint at https://arXiv.org/arXiv:1803.08375 (2018).
35. Eckle, K. & Schmidt-Hieber, J. A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Netw.* **110**, 232–242 (2019).
36. Banerjee, C., Mukherjee, T. & Pasiliao Jr, E. In *Proceedings of the 2019 ACM Southeast Conference*, 164–167.
37. Nishitani, Y., Kaneko, Y., Ueda, M., Morie, T. & Fujii, E. Three-terminal ferroelectric synapse device with concurrent learning function for artificial neural networks. *J. Appl. Phys.* **111**, 124108 (2012).
38. Kaneko, Y., Nishitani, Y., Ueda, M. & Tsujimura, A. In *2013 Symposium on VLSI Technology*, T238-T239 (IEEE).
39. Chen, K.-T. & Chen, J.-S. Perspective on oxide-based three-terminal artificial synapses in physical neural networks. *Appl. Phys. Lett.* **121**, 190501 (2022).
40. Wang, Y. *et al.* A Weight importance analysis technique for area-and power-efficient binary weight neural network processor design. *Cogn. Comput.* **13**, 179–188 (2021).
41. Kim, S., Kim, H.-D. & Choi, S.-J. Impact of synaptic device variations on classification accuracy in a binarized neural network. *Sci. Rep.* **9**, 15237 (2019).
42. Lee, S. .-T. . *et al. IEEE International Electron Devices Meeting (IEDM)* 38.34.31-38.34.34 (IEEE, 2019).
43. Jeon, J., Cho, K. & Kim, S. Disturbance characteristics of 1T DRAM arrays consisting of feedback field-effect transistors. *Micromachines* **14**, 1138 (2023).
44. Son, J., Cho, K. & Kim, S. Electrical stability of p-channel feedback field-effect transistors under bias stresses. *IEEE Access* **9**, 119402–119405 (2021).
45. Lim, D., Son, J., Cho, K. & Kim, S. Quasi-nonvolatile silicon memory device. *Adv. Mater. Technol.* **5**, 2000915 (2020).

Acknowledgements

This research was partly supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2020R1A2C3004538, 2022M3I7A3046571, RS-2023-00260876); the Brain Korea 21 Plus Project through the NRF funded by the Ministry of Science, ICT & Future Planning; and the Korea University Grant.

Author contributions

Y.S. and S.K. provided conceptualization and methodology. Y.S. and K.C. verified and investigated. Y.S., K.C. and S.K. analyzed the results and wrote the manuscript; S.K. supervised the research. All authors edited the manuscript and have given approval to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-56575-4.

Correspondence and requests for materials should be addressed to S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.