frontiers in Physics

ORIGINAL RESEARCH
published: 24 December 2021
doi: 10.3389/fphy.2021.777691

Modeling of Gate Tunable Synaptic Device for Neuromorphic Applications

Yang Shen, He Tian, Yanming Liu, Fan Wu, Zhaoyi Yan, Thomas Hirtz, Xuefeng Wang and Tian-Ling Ren**

School of Integrated Circuits and Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China

The emerging memories are great candidates to establish neuromorphic computing challenging non-Von Neumann architecture. Emerging non-volatile resistive random-access memory (RRAM) attracted abundant attention recently for its low power consumption and high storage density. Up to now, research regarding the tunability of the On/Off ratio and the switching window of RRAM devices remains scarce. In this work, the underlying mechanisms related to gate tunable RRAMs are investigated. The principle of such a device consists of controlling the filament evolution in the resistive layer using graphene and an electric field. A physics-based stochastic simulation was employed to reveal the mechanisms that link the filament size and the growth speed to the back-gate bias. The simulations demonstrate the influence of the negative gate voltage on the device current which in turn leads to better characteristics for neuromorphic computing applications. Moreover, a high accuracy (94.7%) neural network for handwritten character digit classification has been realized using the 1-transistor 1-memristor (1T1R) crossbar cell structure and our stochastic simulation method, which demonstrate the optimization of gate tunable synaptic device.

Keywords: resistive random-access memory, neuromorphic synaptic device, neural network, online learning, device optimization

OPEN ACCESS

Edited by:
Chunsheng Jiang,
China Academy of Engineering Physics, China

Reviewed by:
Huajing Fang,
Xi'an Jiaotong University, China
Zegao Wang,
Sichuan University, China

****Correspondence:***
He Tian
tianhe88@tsinghua.edu.cn
Tian-Ling Ren
RenTL@tsinghua.edu.cn

Specialty section:
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

Received: *15 September 2021*
Accepted: *19 November 2021*
Published: *24 December 2021*

Citation:
Shen Y, Tian H, Liu Y, Wu F, Yan Z, Hirtz T, Wang X and Ren T-L (2021) Modeling of Gate Tunable Synaptic Device for Neuromorphic Applications. Front. Phys. 9:777691. doi: 10.3389/fphy.2021.777691

INTRODUCTION

Artificial intelligence has surpassed human beings in some areas, especially can help process complex and duplicate tasks constantly [1,2]. Traditional Von Neumann architecture is a mainstream architecture adopted to realize large scale AI computing, which separate the storage and calculation unit. The data transmission bus between memory circuit and CPU result in large power consumption, make the Von Neumann not suitable to execute the large scale AI processing in low-power environment such as the Internet of Things (IoT). Inspired by the structure of human brain, neuromorphic computing chips have been proposes and fabricated in a large scale [3,4]. To further improve the performance and energy efficiency, in recent years, emerging nonvolatile memory (NVM) based hardware for neuromorphic computing, which can emulate synapse function in a single device, has attracted much attention [5,6]. The continuous scaling down even down to sub-2 nm of NVM device also indicate that NVM is a great candidate to fabricate large scale high-density neuromorphic computing chips [7]. As for the underlying hardware components, numerous new devices that can integrate both storage and calculation functions have been investigated to establish brain-like circuits to accelerate the running time and reduce power dissipation. Among the NVMs device, RRAM is usually a simple two-terminal structure with a metal oxide layer between the two electrodes, which has been regarded as a promising next mainstream memory technology, due to its fast switching speed, excellent scalability, and non-volatile characteristic [8–11].

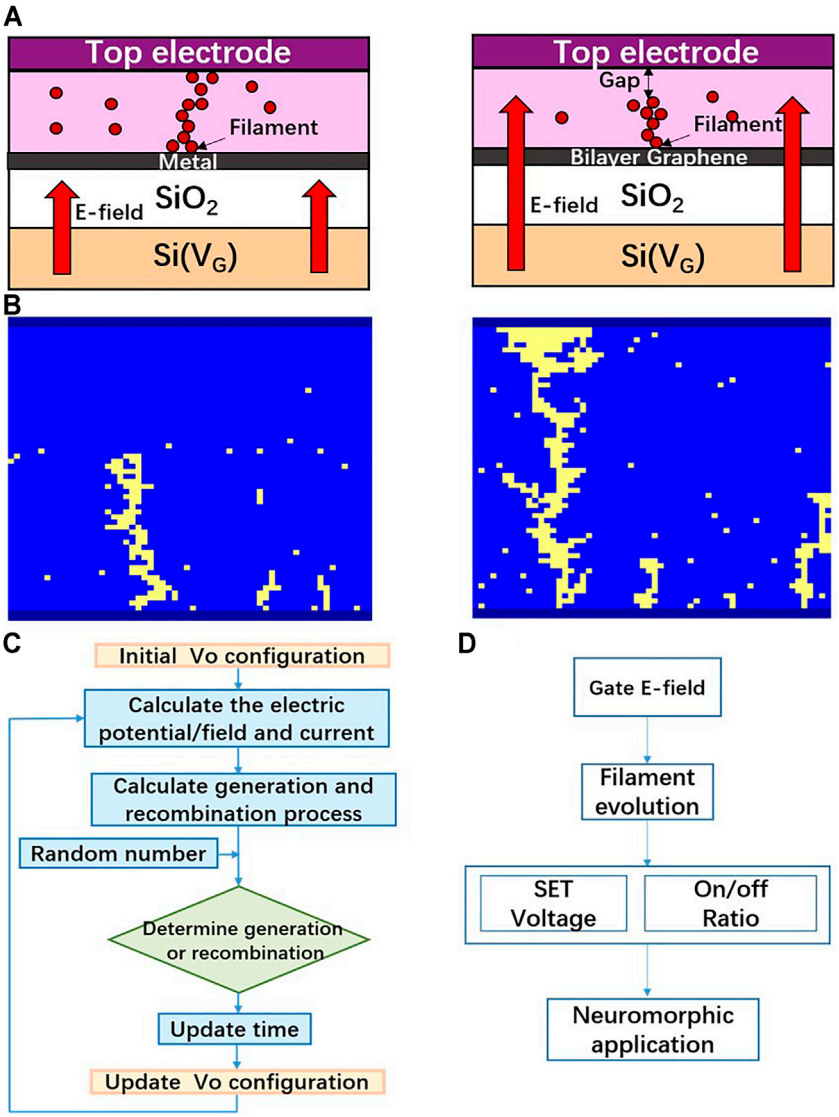

FIGURE 1 | Device structure, tuning mechanism and simulation flowchart. **(A)** Schematic structure of a gate tunable RRAM under a positive, and a negative gate bias. **(B)** The Vo distributions during the set process and after set. **(C)** Flow chart of the simulation process adopted in this work to study the switching process and the device performance. **(D)** The tuning mechanism of gate electric field toward device performance and neuromorphic application.

The resistance changes of a typical RRAM device are done by controlling the formation and the rupture of conductive filaments (CF). For most metal-oxide RRAM, the CFs are composed of oxygen vacancies [12]. Under the influence of an electric field, the metal-oxide is ionized into oxygen ions O^{2-} and oxygen vacancy. During the SET operation, O^{2-} are pulled out of the lattice and drift to the region near the top electrode/dielectric interface. The vacancies generated from the lack of oxygen are forming CFs which in turn are causing a dramatic increase in conductance. During the RESET operation, the reverse electric field repels the O^{2-} away from the top electrode/dielectric interface. The ions recombine with the vacancies, resulting in the CF rupture and therefore a decrease in conductance. Due to abrupt rupture and formation in resistive oxide layer, the RRAM device shows great advantages on switching speed, on/off ratio. On the other hand, the great scalability and recently proposed 3D integration technology of RRAM device.

In order to improve the characteristics of RRAM devices for neuromorphic computing applications, many studies have discussed different methods to enhance the control of the filament growth. One proposed solution is to use a gate tuning mechanism. This solution is effective to control the RRAM device set voltage and the On/Off ratio. This solution has another advantage as it is compatible with the traditional complementary-metal-oxide-semiconductor technology [13]. It is valuable to adopt a physics-based simulation combining the distribution of vacancies evolution with the device performances and analyze the physical mechanisms, so as to accurately predict

its characteristics. In this work, a physics-based stochastic simulation method has been conducted in order to achieve this goal. The particle evolution and the IV characteristics under various gate bias are simulated and analyzed. The simulation results can guide researchers to optimize the gate bias that needs to be applied to obtain an optimal device characteristic. Moreover, oxide RRAM based neuromorphic-computing simulation using a handwritten recognition neural network has been trained on the MNIST dataset to verify the optimization of the applied bias, which obtained a 94.7% recognition accuracy and provide a guideline to tune the online learning accuracy based on novel NVM device.

SIMULATION METHOD

The Structure of Gate Tunable RRAM

Figure 1E demonstrates the whole framework to simulate gate tunable RRAM devices. Following the stochastic method mentioned in the flowchart, it is able to judge the particle drift behaviors and simulate the evolution of a device. A cycle contains two processes, the SET, and the RESET operation. The traditional RRAM structure possesses three layers, a material that has properties that can be altered sandwiched between two layers that serve as electrodes. A gate tunable oxide-based RRAM on the other side possesses graphene and an oxide layer in addition to the three aforementioned layers. **Figure 1A** show the structure of a basic device of this type. The graphene layer is inserted above the bottom electrode between the resistive layer and a SiO_2 layer. In our simulation, the graphene is modeled as a single metal atom layer. The gate bias is applied to the back gate, from which the electric field penetrates through the oxide and the graphene layer and influences the generation and the recombination probability inside the resistive layer. **Figure 1B** shows the growth of set process.

The gate bias can help control the drift velocity of oxide ions and the generation barrier during the SET operation and the RESET operation. Based on the gate-tunable RRAM device, we establish a simulation method to simulate the behaviors of the device. **Figure 1C** shows the different steps that compose a simulation.

The Evolution Model of Oxide-based RRAM

During the SET operation, the generation of oxygen ions and vacancies are dominating. During the RESET operation, the recombination process is dominating. Both of the processes are related to the electric field present in the switching material. As the electric field and the vacancy distributions, the generation and recombination are updated every step. The generation and the recombination formulas are the followings

$$\mathrm{P}_g = f_0 \exp\left(-\frac{(E_a - \gamma q E)}{k_B T} \right) \tag{1}$$

$$\mathrm{P}_r = \beta \mathrm{C}_{ion} f_0 \exp\left(-\frac{E_a}{k_B T} \right) \tag{2}$$

where E_a is the migration barrier, γ is the enhancement factor to the local electric field, C_{ion} is the concentration of oxygen ions, and the E is the electric field which includes the gate electric field and the local electric field. During a DC sweep, P_g and P_r are calculated in all the mesh cells and at every step. They are then compared with a random number. If P_g is higher than the random number, a vacancy can be generated. Similarly, a vacancy can be recombined if P_r is higher than the random number generated. The vacancy distribution is updated this way for the next step. We can see from the P_g formula that the electric field can control the migration barrier.

The Model of Gate Tunable Oxide-based RRAM

The resistance of the device can be divided into two main parts, the resistive layer, and the gate oxide layer. The graphene layer and top/bottom electrode are assumed to be perfect conductors. Due to the uniformity of the gate bias in the dielectric layer, the gate electric field is added to the original electric field in this layer. The formulas that describe the electric fields in the devices are

$$\mathrm{E} = \mathrm{E}_{res-g} + \mathrm{E}_{res-d} \tag{3}$$

$$\mathrm{E}_{res-g} = \frac{V_g - I\rho_{ins}\dfrac{l}{\mathrm{S}}}{d_{res}} \tag{4}$$

$$\mathrm{E}_{res-d} = -\nabla \mathrm{V} \tag{5}$$

where E_{res-g} and E_{res-d} represents the electric field in resistive layer generated by the gate bias and the RRAM bias respectively, ρ_{ins} is the resistivity of the insulator in the gate oxide layer, d_{res} is the thickness of the resistive layer, V is the updated electric potential calculated using the Poisson equation, and V_g is the gate bias.

The technique used in this study consist in regarding the whole resistive material as a network of resistors, and using a Kirchhoff model to compute the electric field and the potential in all the cell of the aforementioned network. Techniques using Kirchhoff's Law fit well with conductive bridge random access memory experimental data [14]. In contrast to classical two terminals RRAM devices, our gate tunable oxide-based RRAM possesses three terminals. The leakage current from the back gate has to be considered. However, because of the high resistivity of the SiO_2, the leak current is far less than the outflow current from the bottom electrode, and can therefore be omitted [12].

SIMULATION RESULTS AND DISCUSSION

After applying voltage bias at the gate electrode, the positive or negative gate electric field have different influences on the I-V characteristics. The generation rate is balanced with the recombination rate under the original condition without bias. When a positive electric field penetrates the resistive layer, the set voltage is increased. In this case, the generation probability is reduced while the recombination rate remains

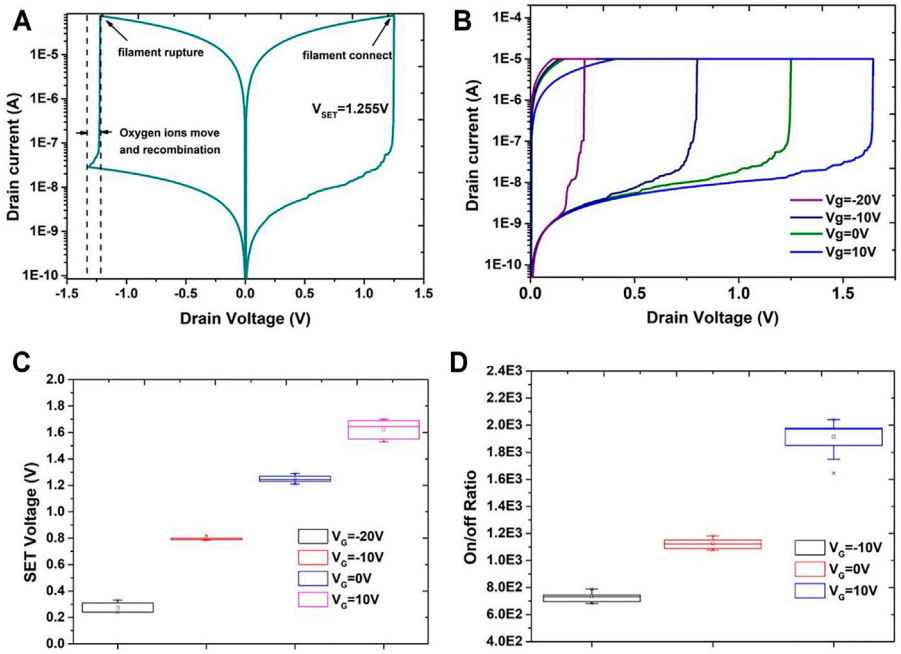

FIGURE 2 | Characteristics of gate tunable oxide-based RRAM. **(A)** The IV curve of a single device without gate bias **(B)** The IV curve of the SET process of different devices with different gate biases. **(C,D)** display the distribution of the SET voltages and the On/Off ratios when the gate bias is set at −20 , −10, 0, and 10 V.

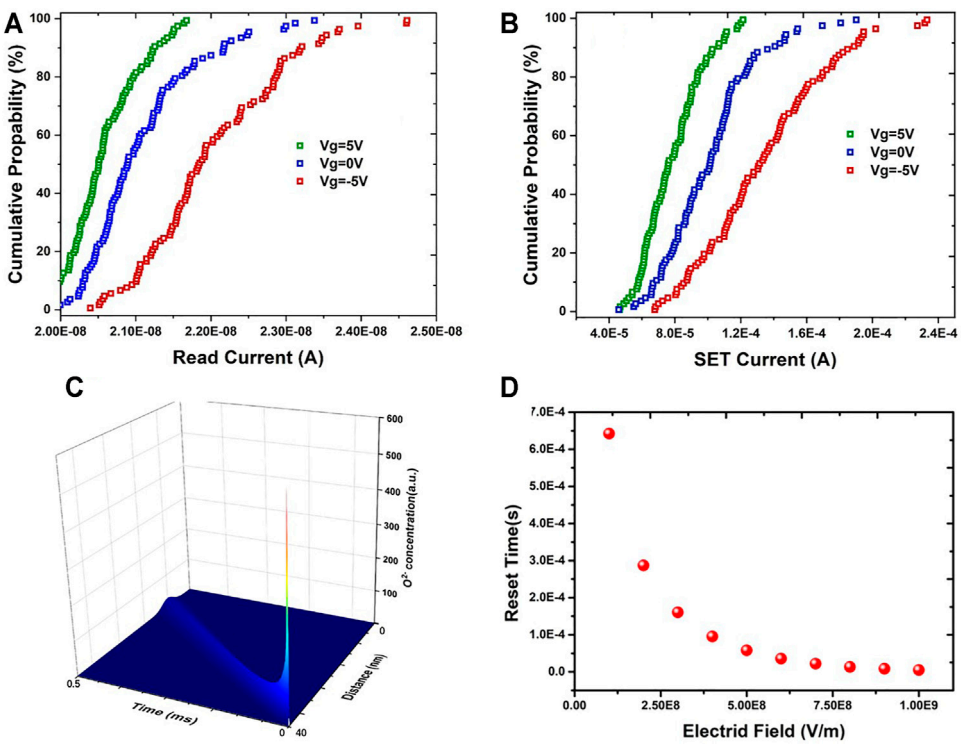

FIGURE 3 | Simulated current distribution and reset process. **(A)** Read current distribution under different gate bias. **(B)** Set current distribution under different gate bias. **(C)** Oxygen ions distribution at different times. **(D)** The relationship of reset time (the time that oxygen ions move to another interface) and E-field.

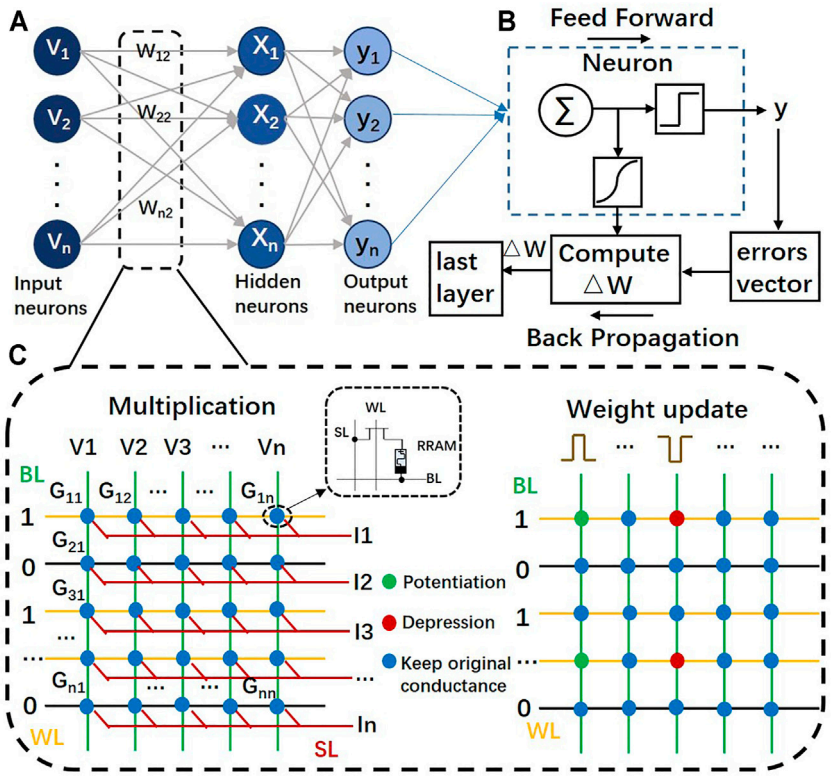

FIGURE 4 | The RRAM-based neuromorphic computing circuit. **(A)** The two layer MLP neural network including 400 neurons in input layer, 100 neurons in hidden layer and 10 neurons in output layer. **(B)** Neural structure consists of matrix-vector multiplication unit and activation function unit, while the sigmoid is used in back propagation and relu is used in feed forward. **(C)** RRAM-based 1T1R crossbar array for realizing synapse in neural network. With the same circuit, the first part illustrates the work mechanism of multiplication and summation, while the second part represents the weight update in BP.

the same, and vice versa when the electric field is negative. **Figure 2A** presents the simulated DC sweep results without gate bias. The set voltage is around 1.255 V. Based on **Eq. 1**, it can be deduced that the reset time is inversely proportional to the drift velocity of the oxygen ions. **Figure 2B** shows the I-V results during the SET process with different gate biases after setting a compliance current of 10^{-5} A. To present the variances of the device performance, a device-to-device simulation was conducted for 10 cycles and for every single gate bias. The results are shown in **Figures 2C,D**. Overall, both the set voltage and the On/Off ratio have an increasing trend as the applied gate bias is swept from negative to positive, which corresponds with the aforementioned results. Therefore, the increase of set voltage and On/Off ratio can be attributed to the filament evolution process. As **Figure 1B** shows, the increase of E-field in the resistive layer will result in the decrease of reset time. The exponential reduction trend mean that the reset time can be tuned by gate bias exponentially if we adopt the model in this work.

The read current and set current shown in **Figures 3A,B** have similar trend that negative or smaller positive gate bias cause the current shifting to larger magnitude of level and increase the distribution range. This's can be explained that generation probability is threatened and the stochastic property is enhanced under strong E-field. On the other hand, to simulate the reset process, oxygen ions transport equation was numerically solved. As the E-field value increase, it can be seen that reset time would be decreased exponentially.

GATE TUNABLE OXIDE-BASED RRAM FOR NEUROMORPHIC COMPUTING

A Ideal NVM device should be totally linear during potentiation and depression, and the conductance states would increase or decrease at a fixed step after applying positive pulse or negative pulse with identical magnitude. The other impact factors include On/Off ratio, cycle-to-cycle variations and device-to-device variations, all of which can deteriorate analog RRAM-based neuromorphic computing hardware towards application [15]. Moreover, the number of states can also determine the weight update precision. There are some efforts to improve the number of immediate states, and even 256 analogue states has been realized [16,17].

In this work, the gate tunable RRAM has been used to establish a neural network to investigate the gate tunable linearity and on/off ratio. In neuromorphic computing, the device characteristic highly determines the accuracy of the neural network. **Figure 4**

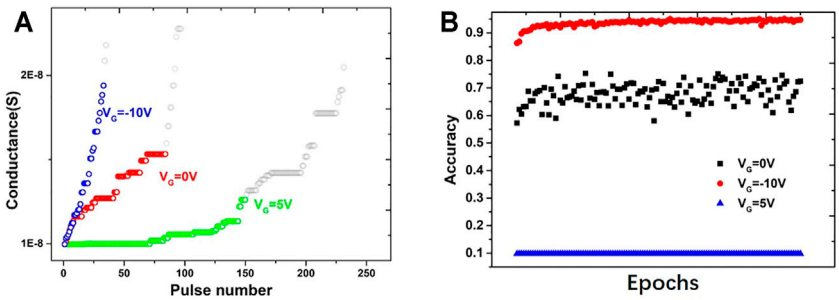

FIGURE 5 | Gate tunable RRAM-based neural network. **(A)** Gradual switching behavior using 1 V amplitudes and 10 ns width constant pulses. **(B)** The online learning accuracy based on the RRAM device and our stochastic simulation method under different gate bias after 125 epochs.

illustrated the whole circuit to implement two-layer perceptron (MLP) using RRAM crossbar arrays as core component. The adopted neural network consists of 400 neurons in input layer, 100 neurons in hidden layer and 10 neurons in output layer, as **Figure 4A** shows. The neurons are CMOS circuits to implement activation function. The training process use back propagation algorithm and stochastic gradient descent to update weight (**Figure 4B**). The adopted circuit system is based on a typical RRAM-based neural system [5].

In **Figure 4C**, the bit line (BL) is for applying input voltage or pulse voltage to calculate output in feed forward (FF) process or change the conductance state in back propagation (BP). The word line (WL) is used to set the gate voltage controlling on-off of transistor to determine the selected row which can implement weight update. According to the scheme, a simple and compact crossbar array structure is used as weight matrix to implement vector-matrix multiplication and weight update if pulse is applied.

Figure 5A shows the conductance versus pulse curve for different gate voltages. The highlighted data is the data that will be used for training the neural networks. For this experiment, we selected data that possess a linear coefficient larger than 0.9. The selected data has demonstrated that the additional gate in RRAM can further control the growth of the filament, which also improves the linearity and On/Off ratio of the linear segment of the conductance-pulse data. The device characteristics for neuromorphic computing gradually get better when the gate bias decrease from 5 to −10 V. Most notably, when −10 V is applied to the gate, the linearity of the devices can reach 0.9821. The Modified National Institute of Standards and Technology (MNIST) database has been used to train the network. The training result has been shown in **Figure 5B**, the array of devices possessing better linearity and On/Off ratio shows an improved recognition accuracy. With the 5 V back gate voltage, the devices have a lower On/Off ratio and lower linearity leading to poor results in accuracy. The optimized recognition accuracy can reach 94.8% under −10 V gate voltage. This result demonstrates that the optimized devices indeed improve the recognition accuracy when used in a neuromorphic application.

CONCLUSION

This work proposed an effective method to simulate gate tunable RRAM devices. We employed a physics-based stochastic model to investigate the internal mechanism of the devices. The controllability of the parameters has a stable trend and is promising in practical applications. Furthermore, devices under appropriate bias have been used to compose a neural network for handwritten digit classification. This application of the simulated device into a real-world example has been done to authenticate the effectiveness of our framework for modeling circuit and neuromorphic systems. This work reveals the tunable mechanism of the gate tunable oxide-based RRAM and promotes the application of the device for neuromorphic computing.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HT proposed the idea. YS wrote the paper and tested the main code. YL wrote the paper and tested the code. All the authors revised the manuscript.

FUNDING

This work was supported by the National Key R&D Program (2020YFA0709800) and the National Natural Science Foundation (62022047, 61874065, 51861145202, U20A20168) of China. This work is also supported by the Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001, and by the Fok Ying-Tong Education Foundation under Grant 171051. This work was also supported by the Beijing National Research Center for Information Science and the Technology Youth Innovation Fund (BNR2021RC01007).

REFERENCES

1. LeCun Y, Bengio Y, Hinton G Deep Learning. *nature* (2015) 521:436–44. doi:10.1038/nature14539
2. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the Game of Go without Human Knowledge. *nature* (2017) 550: 354–9. doi:10.1038/nature24270
3. Roy K, Jaiswal A, Panda P Towards Spike-Based Machine Intelligence with Neuromorphic Computing. *Nature* (2019) 575:607–17. doi:10.1038/s41586-019-1677-2
4. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E Memory Devices and Applications for In-Memory Computing. *Nat Nanotechnol* (2020) 15: 529–44. doi:10.1038/s41565-020-0655-z
5. Chen YS, Lee HY, Chen PS, Gu PY, Chen CW, Lin WP, et al. Highly Scalable Hafnium Oxide Memory with Improvements of Resistive Distribution and Read Disturb Immunity. In: Proceeding of the 2009 IEEE International Electron Devices Meeting (IEDM); 7-9 Dec. 2009; Baltimore, MD, USA. IEEE (2009). p. 1–4. doi:10.1109/IEDM.2009.5424411
6. Prezioso M, Mahmoodi MR, Bayat FM, Nili H, Kim H, Vincent A, et al. Spike-timing-dependent Plasticity Learning of Coincidence Detection with Passively Integrated Memristive Circuits. *Nat Commun* (2018) 9:5311–8. doi:10.1038/s41467-018-07757-y
7. Pi S, Li C, Jiang H, Xia W, Xin H, Yang JJ, et al. Memristor Crossbar Arrays with 6-nm Half-Pitch and 2-nm Critical Dimension. *Nat Nanotech* (2019) 14: 35–9. doi:10.1038/s41565-018-0302-0
8. Guan X, Yu S, Wong H-SP On the Switching Parameter Variation of Metal-Oxide RRAM-Part I: Physical Modeling and Simulation Methodology. *IEEE Trans Electron Devices* (2012) 59:1172–82. doi:10.1109/ted.2012.2184545
9. Chen W-H, Li K-X, Lin W-Y, Hsu K-H, Li P-Y, Yang C-H, et al. A 65nm 1Mb Nonvolatile Computing-In-Memory ReRAM Macro with Sub-16ns Multiply-And-Accumulate for Binary DNN AI Edge Processors. In: Proceeding of the 2018 IEEE International Solid - State Circuits Conference - (ISSCC); 11-15 Feb. 2018; San Francisco, CA, USA. IEEE (2018). p. 494–6. doi:10.1109/ISSCC.2018.8310400
10. Waser R, Aono M Nanoionics-based Resistive Switching Memories. *Nanoscience Technol A Collection Rev Nat Journals* (2010) 6:158–65.
11. Akinaga H, Shima H Resistive Random Access Memory (ReRAM) Based on Metal Oxides. *Proc IEEE* (2010) 98:2237–51. doi:10.1109/jproc.2010.2070830
12. Tian H, Wang X, Zhao H, Mi W, Yang Y, Chiu P-W, et al. A Graphene-Based Filament Transistor with Sub-10 mVdec−1 Subthreshold Swing. *Adv Electron Mater* (2018) 4:1700608. doi:10.1002/aelm.201700608
13. Tian H, Zhao H, Wang X-F, Xie Q-Y, Chen H-Y, Mohammad MA, et al. *In Situ* Tuning of Switching Window in a Gate-Controlled Bilayer Graphene-Electrode Resistive Memory Device. *Adv Mater* (2015) 27:7767–74. doi:10.1002/adma.201503125
14. Liu Y, Yang K, Wang X, Tian H, Ren T-L Lower Power, Better Uniformity, and Stability CBRAM Enabled by Graphene Nanohole Interface Engineering. *IEEE Trans Electron Devices* (2020) 67:984–8. doi:10.1109/ted.2020.2968731
15. Wu H, Yao P, Gao B, Wu W, Zhang Q, Zhang W, et al. Device and Circuit Optimization of RRAM for Neuromorphic Computing. In: Proceeding of the 2017 IEEE International Electron Devices Meeting (IEDM); 2-6 Dec. 2017; San Francisco, CA, USA. IEEE (2017). p. 11–5. doi:10.1109/IEDM.2017.8268372
16. Mochida R, Kouno K, Hayata Y, Nakayama M, Ono T, Suwa H, et al. A 4M Synapses Integrated Analog ReRAM Based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture. In: Proceeding of the 2018 IEEE Symposium on VLSI Technology; 18-22 June 2018; Honolulu, HI, USA. IEEE (2018). p. 175–6. doi:10.1109/VLSIT.2018.8510676
17. Wang Z, Yin M, Zhang T, Cai Y, Wang Y, Yang Y, et al. Engineering Incremental Resistive Switching in TaOxbased Memristors for Brain-Inspired Computing. *Nanoscale* (2016) 8:14015–22. doi:10.1039/c6nr00476h

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.