Time-Dependent Variability in RRAM-based Analog Neuromorphic System for Pattern Recognition

Jian Kang[1], Zhizhen Yu[1], Lindong Wu[1], Yichen Fang[1], Zongwei Wang[1], Yimao Cai[1,2*]
Zhigang Ji[1,3*], Jianfu Zhang[3], Runsheng Wang[1,2], Yuchao Yang[1,2] and Ru Huang[1,2*]
[1]Institute of Microelectronics, Peking University, 100871, Beijing, China (*Email: caiyimao@pku.edu.cn & ruhuang@pku.edu.cn)
[2]National Key Laboratory of Science and Technology on Micro/Nano Fabrication, 100871, Beijing, China
[3]Department of EEE, Liverpool John Moores University, Liverpool, L3 3AF, UK (*Email: z.ji@ljmu.ac.uk)

Abstract - For the first time, this work investigated the time-dependent variability (TDV) in RRAMs and its interaction with the RRAM-based analog neuromorphic circuits for pattern recognition. It is found that even the circuits are well trained, the TDV effect can introduce non-negligible recognition accuracy drop during the operating condition. The impact of TDV on the neuromorphic circuits increases when higher resistances are used for the circuit implementation, challenging for the future low power operation. In addition, the impact of TDV cannot be suppressed by either scaling up with more synapses or increasing the response time and thus threatens both real-time and general-purpose applications with high accuracy requirements. Further study on different circuit configurations, operating conditions and training algorithms, provides guidelines for the practical hardware implementation.

Introduction

RRAM-based neuromorphic circuit has attracted extensive attention **[1,2]**. The binary and multi-level RRAM synapses have been demonstrated (**Fig.1a**) **[3,4]**. To achieve the full potential, efforts have been made on the analog neuromorphic system using the plasticity property of RRAMs (**Fig.1b&c**) **[5-6]**. The variability inevitably becomes the critical concern. During training condition, many programming schemes, such as verification **[7]** have been developed to suppress resistance variability due to the stochastic filament growth. However, for the well-trained circuit, during operating condition, the time-dependent variability, TDV, induced by the noise, can dynamically change the RRAM resistance and thus the weight of the synapse, causing accuracy loss.

Time-dependent variability (TDV)

For low power operation, RRAMs with higher resistance are preferable for the synapses implementation. The range from tens of kΩ to several MΩ have been reported **[8-10]**. TDV increases gradually with the higher resistance (**Fig.2a-c**). Such trend can also be observed from literatures in recent years (**Fig.2d**). TDV originates from defects, therefore, the TDV-induced instability is expected to be an intrinsic issue for any defect-based RRAM technology **[11-12]** and thus needs to be addressed properly. *In this work, for the first time, the impact of TDV on the neuromorphic circuit under operating condition for pattern recognition is investigated. The emphasis is made on the interaction between TDV and the circuits with different implementations including the system scalability, the synapse configuration, the training algorithms and the circuit operating conditions.* RRAMs with TiN/Ta_2O_5/Pt structure and a 100nm sputtering-deposited TaOx oxide were used in the work. The fabrication process is detailed in ref. **13**.

Characterization and Model of TDV for RRAMs

TDV-induced resistance variation is monitored with the continuous current measurements in this work. As shown in **Fig.3a**, the measured current can be divided into two components: the time-invariant I_{stable} and the time-varying I_{fluc}. I_{fluc} introduces TDV, which varies with different RRAMs exhibiting additional cell-to-cell variability (**Fig.3a&b**). The current conduction by several defects (such as oxygen vacancy) has been proposed **[14]**. Each defect carries certain amount of current. With the statistical RTN analysis using HMM method **[15],** the current conducted by each defect follows an exponential distribution (**Fig.4**). By assuming the defect number is found to follow the Poisson distributed in different RRAMs **[16]**, the distribution of total current from multiple RRAMs can be formulated by summing the exponential distribution weighted by the Poisson probability, as shown in Eqn (3) **[17]**. By further derivation, the *average number of defect*, N, and *current conduction per defect*, ΔI, can be calculated with Eqns (4-5) based on the cell-to-cell variation of the current. Since RRAMs have higher cycle-to-cycle variation than the cell-to-cell variation **[18]**, the cell-to-cell variation can be analyzed using the cycle-to-cycle data measured within one RRAM.

Two examples under different compliance current, I_{cc}, are shown in **Fig.5**, in which I_{fluc} and I_{stable} are extracted from 300 cycles. The median values of I_{stable} is larger than I_{fluc} (**Fig.5a&b)**. However, the extracted ΔI are similar (**Fig.5c&d**), as further confirmed by comparing multiple RRAMs with various conditions (**Fig.6**). This suggests that defects for I_{fluc} have the same current conduction capability as defects for I_{stable}. **Fig.5e&f** shows the defect number corresponding to I_{fluc} is much less than I_{stable} (*cf.* $N_{fluc} < N_{stable}$). Both N_{fluc} and N_{stable} do not vary with voltage. Therefore, their values were taken from the currents measured at 0.1V hereafter.

N_{stable} shows strong dependence on I_{cc} and $V_{reset,}$ suggesting its sensitivity to the shape of the filament (**Fig.7a&b**): A shorter filament due to smaller V_{reset} and a larger cross-section formed with higher I_{cc} will contain more defects for current conduction **[19]**. In contrast, N_{fluc} keeps as a constant under different I_{cc} and V_{reset} (**Fig.7a&b**), supporting that they locate at the boundary of the filament **[20]**. I_{stable} also varies with I_{cc} and V_{reset}. However, its median and standard deviation exhibit a power law relationship (**Fig.8a**) **[18]**, based on which ΔI can be readily calculated.

During practical operation, the neurons will integrate the currents before firing which is expected to reduce TDV. In the device level, this can be investigated by using different measurement time, t_m (**Fig.9**): N_{fluc} reduces with longer t_m, while ΔI is unchanged. This is because at longer t_m, the trapping/de-trapping of some defects have been averaged out, leading to smaller N_{fluc}. Such averaging effect does not change the ability for the current conduction per defect and thus ΔI keeps constant. The temperature effect is another important factor. It introduces little effect on N_{fluc} (**Fig.10**) and the relationship between the median and standard deviation of I_{stable} (**Fig.8a**). However, the median value of I_{stable} exhibits temperature dependence with the activation energy of E_a (**Fig.8b**).

Given an ideal resistance of R_0 at room temperature, I_{stable} at any temperature can be firstly determined with Eqns (7-8). TDV is introduced (**Fig.11**): After calculating ΔI and N_{fluc} from Eqns (4-6), the defect number and the current conduction by each defect in one RRAM cell can be generated with their respective distributions defined in Eqns (1&2). By assuming these defects are uniformly distributed in space and energy **[21]**, the filling probability can be randomly generated. I_{fluc} is the summation of the currents conducted by the unoccupied defects which vary with time (Eqn (9)). The model is capable to produce the TDV behavior similar to the measurement in one cycle (**Fig.12a&b**). In addition, the model is validated by comparing the measured resistance distribution from multiple cycles (**Fig.12c**). Good agreement can be achieved.

Circuit/Device Interaction within Neuromorphic System

The neuromorphic system training with the winner-takes-all (WTA) algorithm was simulated for the MNIST handwritten digit recognition (**Fig.11c**) **[22]**. 60000 images were used in the training and finally the accuracy reached stabilization (**Fig.13**). The pre-trained weight metrics is implemented into the RRAM array with and without TDV (**Fig.11b**). One example is shown **Fig.14**. Five random-chosen inputs can be recognized by the system without TDV. One TDV-embedded *system instance* is generated and the same input is repeatedly used for pattern recognition. The weight metrics varies due to TDV and thus occasionally the system fails to recognize the input. In the following TDV analysis, the recognition accuracy is evaluated by repeating the above procedure on 1000 images and 30 *system instances* are used to extract the distribution of the accuracy. For clarity, the median value and the ±3σ level is shown in **Figs.15-23.** Unless specified, the integration time of 10ms and room temperature are used in the simulation.

a. TDV interaction with artificial synaptic configuration

For analog neuromorphic circuit with one RRAM as a synapse, the pre-trained weights need to be mapped into a range of resistances. The resistance adjustment relies on the potentiation and depression of RRAMs, which is closely related to the fabrication process. Therefore, it is critical to understand the TDV impact with different resistance ranges. The resistance range is firstly changed by gradually increasing the lower boundary R_{LB}, while keeping the constant span (e.g. $R_{UB} = 10*R_{LB}$). When RRAMs with higher R_{LB} are used, both the recognition accuracy and the power consumption reduces (**Fig.15a&b**). The resistance range can also be changed by keeping R_{LB} as the constant but gradually increasing R_{UB}. TDV-induced accuracy drop reduces and quickly reaches saturation when R_{UB}/R_{LB} is larger than one decade (**Fig.16a**). In terms of the power consumption, if R_{LB} is fixed, further increasing R_{UB} can reduce much less power compared with increasing R_{LB} with constant R_{UB}/R_{LB} (**Fig.15b&Fig.16b**). This shows that although a wider range is used in the system, the RRAMs with lower resistance still dominate the power and the accuracy of the system. They can carry larger current and contribute more under the WTA rule, while at the same time, dominate the power consumption.

For resistance adjustment, RRAM depression is usually more difficult to achieve compared with the potentiation and thus challenges the implementation with one RRAM as the synapse in the neuromorphic system. Therefore, two RRAMs with the opposite contribution to the neuron's integration has been suggested as one potential solution. Wherein, the RRAM depression can be converted to the potentiation of the second RRAM cell **[23]**. **Fig.17** shows that two-RRAM configuration exhibits slightly better immunity to TDV.

b. TDV interaction with the number of synapse

At the expense of higher power consumption, increasing the number of synapse has been considered as one effective way to achieve high accuracy (**Fig.18a&b**). However, TDV-induced accuracy drop does not reduce (**Fig.18a**). Therefore, the accuracy under practical operation is eventually limited by TDV, which must be minimized for high accuracy application.

c. TDV interaction with the response time

Different applications impose different requirements to the response time for pattern recognition. In the circuit level, WTA rule relies on the currents to be integrated within certain time before triggering the neuron to response. The longer response time are expected to be more effective for TDV suppression through averaging effect. However, TDV-induced accuracy drop is almost unchanged even under the response time of 1s in the neuromorphic circuit (**Fig.19a**). This is because when the circuit operates under low voltage, the characteristic time of many defects within TDV can be much longer than the response time, weakening averaging effect. Therefore, this suggests that TDV can be harmful to both real-time and general-purpose applications.

d. TDV interaction with the ambient temperature

3D architecture has been proposed for future circuit integration **[24]**. Due to thermal dissipation, circuits will be inevitably affected by the temperature. When temperature increases, the impact of TDV becomes smaller and the accuracy starts to increase (**Fig.20**). This can be understood from the nature of TDV: the higher temperature increases the frequency for the (de)trapping of defects and thus at the constant response time, the TDV can become smaller. Practically, the temperature is unlikely to rise within the entire circuit. With the temperature rises only in certain local areas, the TDV impact is found to be similar (**Fig.21**).

e. TDV interaction with the learning algorithms

Back-propagation algorithm (BP) is one of the most popular algorithm and has been investigated for hardware implementation **[2]**. Compared with WTA, BP exhibits much higher sensitivity to TDV (**Fig.22a**). To further understand this, the resistances implemented in the system trained with BP and WTA are compared (**Fig. 22b&c**). Very different distributions can be observed: For BP, most of the weights are narrowly distributed. A small variation in resistance will have high impact on the weight and thus induces high TDV sensitivity. Compared with WTA, BP in theory can achieve similar accuracy with lower power, however, the accuracy deteriorates after hardware implementation due to TDV (**Fig. 23**).

Conclusion

This work investigated the TDV in RRAMs and its interaction within the analog neuromorphic circuits for pattern recognition. TDV increases in RRAM cells with higher resistance because the corresponding number of defects do not scale accordingly. The circuit-level analysis revealed that TDV can deteriorate the pattern recognition accuracy when the neuromorphic circuit is implemented with high resistance for low power operation. The impact of TDV cannot be suppressed by either scaling up with more synapses or increasing the response time and thus challenges both real-time and general-purpose applications with high accuracy requirements. In addition, TDV exhibits strong interaction with training algorithms, which therefore must be properly chosen for practical hardware implementation.

Acknowledgement: This work was supported by National Natural Science Foundation of China (61574007, 61376087 and 61421005).

References

[1] G. Indiveri et al, Proc. IEEE, pp.1379. **[2]** S. Yu, et al, IEDM, 2015. **[3]** M. Suri et al, T-ED, 2013. **[4]** S. Yu, et al, IEDM, 2012. **[5]** Z. Chen, et al, IEDM, 2015. **[6]** G. Indiveri, et al, IEDM, 2015. **[7]** K. Kim, et al, Nano Lett, 2012. **[8]** S. Park, et al, IEDM, 2013. **[9]** D. Garbin, et al, IEDM, 2014. **[10]** S. Ambrogio, et al, T- ED, 2016 **[11]** A. Belmonte, et al, IMW, 2014. **[12]** B. Govoreanu, et al, VLSI, 2016. **[13]** Z. Wang, et al NVMTS, 2016. **[14]** R. Degraeve, et al, VLSI, 2012 **[15]** Z. Zhang, et al, IRPS, 2017. **[16]** S. Balatti, et al, IMW, 2013. **[17]** B. Kaczer, et al, IPRS, 2010. **[18]** A. Grossi, et al, IEDM, 2016. **[19]** D. Ielmini, T-ED, 2011. **[20]** D. Veksler, et al, IEDM, 2012. **[21]** Z. Chai, et al, IEDM, 2016. **[22]** Y. Lecun, et al, Proc. IEEE, 1998. **[23]** O. Bichler, et al, T-ED, 2012. **[24]** H. Li, et al, IEDM, 2016.

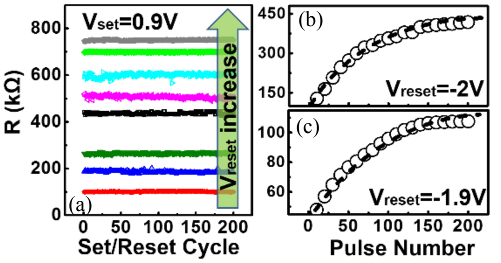


Fig.1 (a) Resistance with multiple levels measured with consecutive set/reset cycles under different V_{reset} varying from -1.5V to -2.0V. V_{set} = 0.9V. **(b&c)** Resistance change as a function of number of applied potentiation pulses with **(b)** -2V and **(c)** -1.9V. Pulse is with 10ns width. Forming I_{cc} = 300μA.

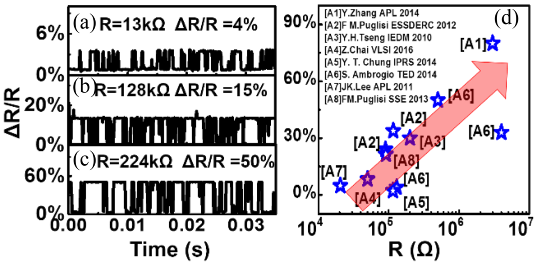


Fig.2 (a-c) Examples of TDV, measured on RRAMs with different resistances: 13kΩ, 128kΩ and 224kΩ. Resistance is measured at 0.1V. At higher resistance, higher variation in resistance can be observed. **(d)** TDV dependence of resistance from literature.

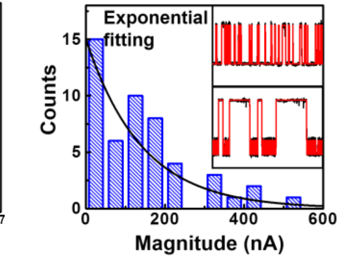


Fig.4 Histograms of RTN magnitudes from 100 cycles of 7 devices. I_{cc} = 30μA. V_{set}/V_{reset} = 0.9V/-1.2V. Inset: representative RTN and their HMM fitting.

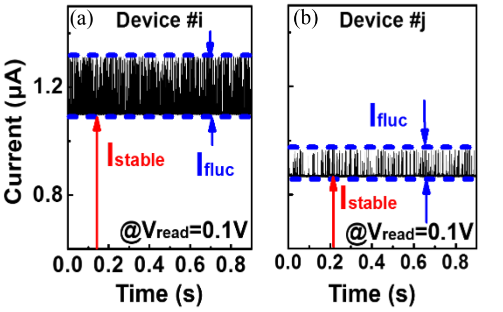


Fig.3 Current measurement under 0.1V for 1sec after RRAM reset to its HRS. test condition: I_{cc} = 300μA, V_{set}/V_{reset} = 0.9V/-1.5V. Two devices were shown in **(a)** and **(b)**. The minimum current is defined as I_{stable} and the peak-to-peak value as I_{fluc}.

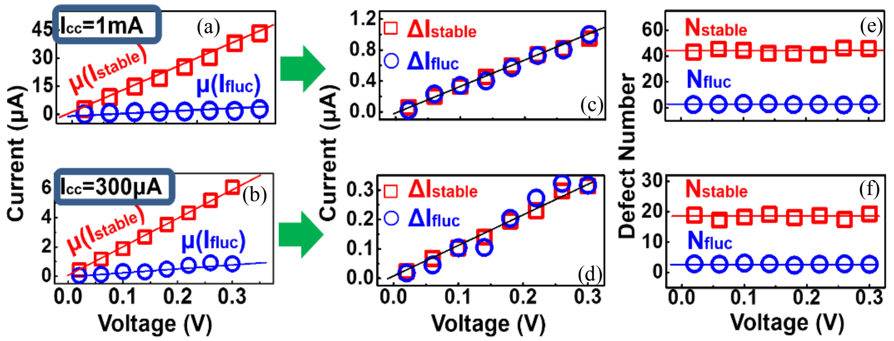


Fig. 5 I_{stable} and I_{fluc} after averaging from 300 cycles under I_{cc} = 1mA **(a)** or 300μA **(b)**. The read voltage is 0.1V and V_{set}/V_{reset} is 0.9V/-1.3V for both cases. **(c&d)** The current conduction per defect and **(e&f)** the effective number of defect for I_{stable} and I_{fluc} extracted from the measured data in **(a)** & **(b)** respectively using Eqn (4&5). The average current conduction per defect in the same for both I_{stable} and I_{fluc}.

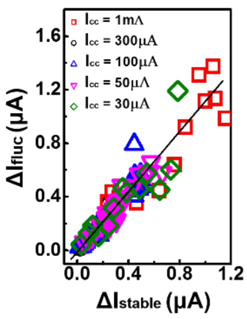


Fig. 6 Comparison of ΔI_{fluc} and ΔI_{stable} under different I_{cc} (30μA ~1mA), V_{read} (0.02V to 0.3V) and V_{reset} (-1.2V ~ -1.9V). V_{set} = +0.9V.

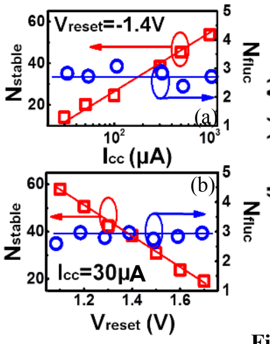


Fig. 7 Dependence of N_{stable} and N_{fluc} with **(a)** I_{cc} and **(b)** V_{reset}. V_{set} = +0.9V is used for all cases. Measurements are taken at room temperature.

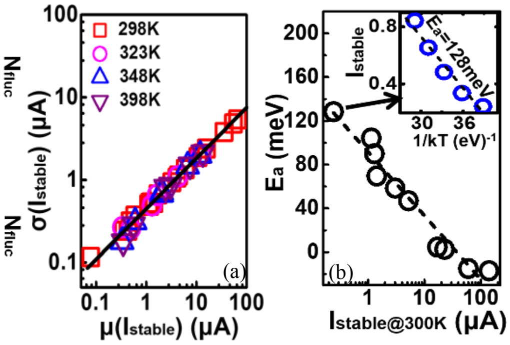


Fig.8 (a) Relationship between standard deviation and median value of I_{stable} measured at 0.1V under different temperatures. V_{reset} from -1.2V to -1.9V. V_{set} = +1.2V. I_{cc} = 300μA. **(b)** Dependence of activation energy, E_a, on the median resistance. Inset shows one examples for E_a extraction.

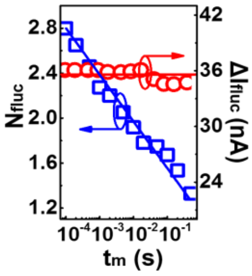


Fig.9 Dependence of the defect number N_{fluc} and current ΔI_{fluc} on the integration time, t_m.

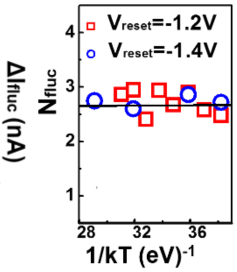
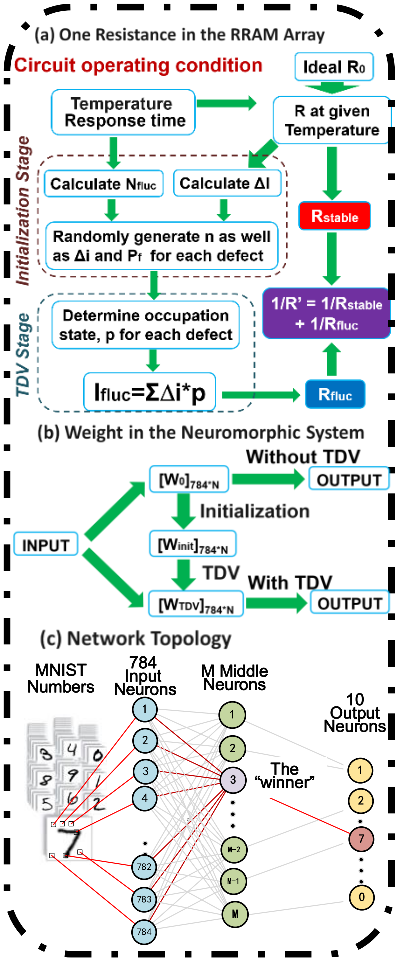


Fig.10 Dependence of the defect number, N_{fluc}, on the temperature. Two devices after reset under -1.2V and -1.4V are used..

Fig.11 (RHS) (a) Procedure for introducing TDV into the ideal resistance at room temperature. *Initialization stage*: I_{stable} at 0.1V under given temperature is first calculated. Then the defect number, and the current per defect is generated from their average value with Eqn (1&2). The filling probability of each defect is randomly generated from 0 to 1. *TDV stage*: the unoccupied defects are randomly chosen and the total current can be obtained by Eqn (9). Then R' can then be obtained. **(b)** Procedure for taking TDV into the network: The TDV model is applied on each resistance within the weight matrix and thus a new matrix is generated. **(c)** Network topology. The input layer contains 784 neurons. The hidden layer with M neurons which can vary in our investigation. The output layer has 10 neurons corresponding to 10 classes of digits.



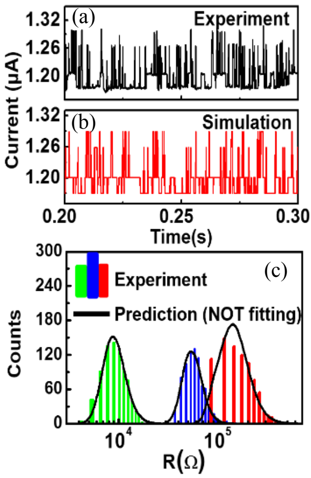


Fig.12 Similarity in TDV-induced current from **(a)** measurements and **(b)** simulation with the proposed model. **(c)** Comparison of the resistance distribution from the measurements and model prediction. The resistances were measured within 1s under 0.1V from 300 cycles. Three RRAMs were used with V_{reset} (-1.2V,-1.4V and -1.7V). I_{cc} = 300μA and V_{set} = 1.2V are used for all cases.

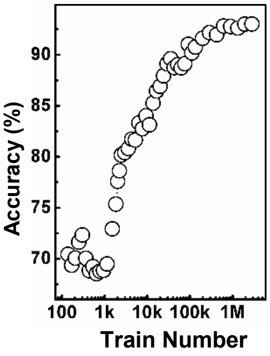


Fig.13 Iteration of the training. The accuracy increases when more samples are used and finally reach stabilization.

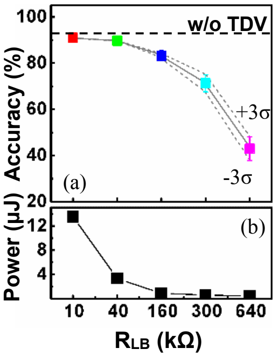


Fig.15 TDV impact with resistance range defined from R_{LB} to R_{UB}=10*R_{LB} on **(a)** recognition accuracy and **(b)** power consumption.

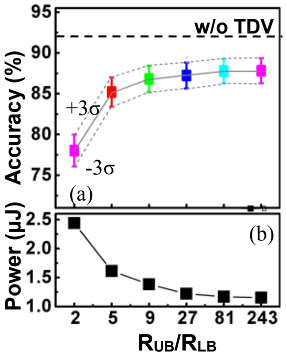


Fig.16 TDV impact with resistance range from R_{LB}=100kΩ to different R_{UB} for **(a)** recognition accuracy and **(b)** power consumption.

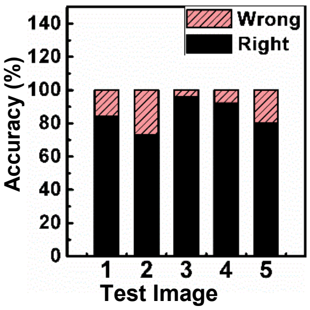


Fig.14 TDV impact on the recognition accuracy for 5 input patterns using one circuit instance, implemented with resistance from 100kΩ to 1MΩ.

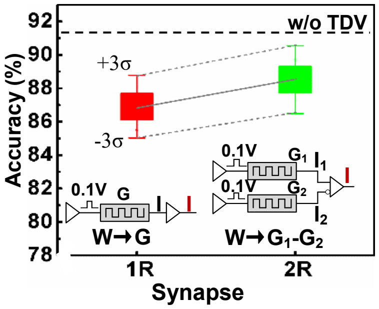


Fig.17 Comparison of TDV immunity for different types of synapses: 1R and 2R. Wherein, R of 100kΩ~1MΩ are used.

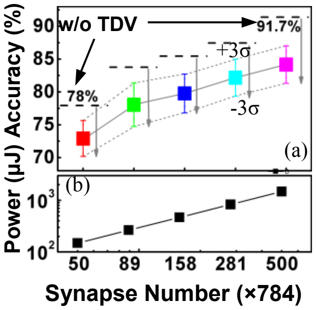


Fig.18 TDV impact comparison with the synapse number for **(a)** pattern recognition accuracy and **(b)** power consumption.

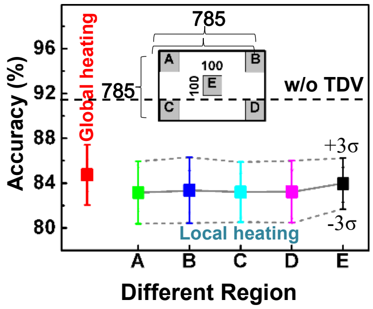


Fig.21 TDV impact comparison under different local temperature rise (defined in inset) on accuracy. R of 100kΩ~1MΩ are used.

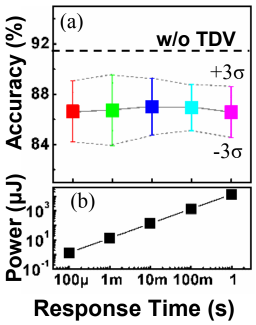


Fig.19 TDV impact comparison with integration time for **(a)** pattern recognition accuracy and **(b)** power consumption.

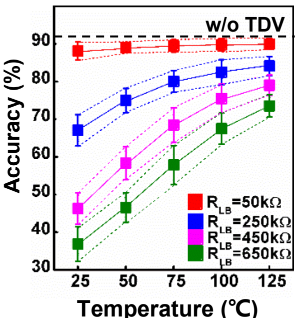


Fig.20 TDV impact comparison under different temperature on the accuracy implementing with resistance from R_{LB} to 10*R_{LB}.

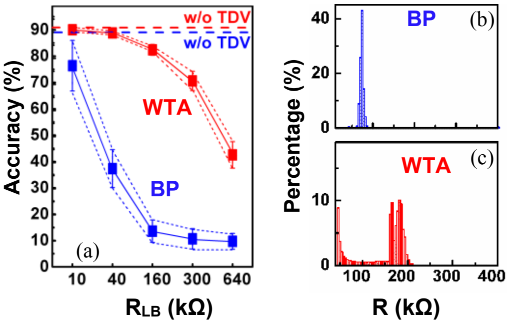


Fig.22 **(a)** TDV impact comparison with two different algorithms: WTA and BP. Different resistance range with R_{LB} to 10*R_{LB} are used. The resistance distribution for the well-trained system using **(b)** WTA and **(c)** BP.

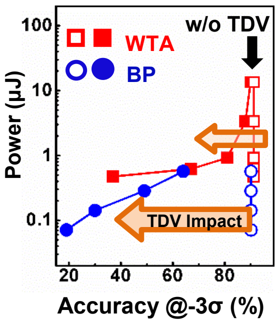


Fig.23 Power consumption vs. accuracy for WTA & BP. Power decreases when higher resistance ranges are used.

1. PDF of defect number and current per defect (single cell):

$$P_N(n)=\frac{e^{-N}\cdot N^n}{n!} \quad (1) \qquad f(\Delta i)=\frac{1}{\Delta I}\exp(-\frac{\Delta i}{\Delta I}) \quad (2)$$

n is the number of defect and Δi is the current conduction per defect.

2. CDF of current conduction from multiple cells:

$$F_N(I;\Delta I,N)=\sum_{n=1}^{+\infty}e^{-N}\cdot N^n/n!\cdot\left[1\text{-}\Gamma(n,I/\Delta I)/(n-1)!\right] \quad (3)$$

3. Average number of defect and current conduction (multiple cells):

$$N=2\cdot\left[\mu(I)/\sigma(I)\right]^2 \quad (4) \qquad \Delta I=\sigma(I)^2/(2\cdot\mu(I)) \quad (5)$$

N: average number of defect; ΔI: average current conduction per defect.

4. Response time dependence for N_{fluc}:

$$\mathrm{N}_{\mathrm{fluc}}=g_0\cdot t^{n1} \quad (6)$$

5. Temperature dependence for I_{stable}:

$$\mu(I_{stable},T)=\mu(I_{stable},RT)\cdot\exp(-E_a/kT) \quad (7)$$

$$E_a=a_1\cdot\log_{10}(\mu(I_{stable},RT))+a_2 \quad (8)$$

6. TDV on I_{fluc}: $I_{fluc}=\sum_{k=1}^{n}\Delta i_k\cdot p\,,\quad p=\begin{cases}0 & \mathrm{r}<\mathrm{p_{f,k}}\\ 1 & \mathrm{r}>\mathrm{p_{f,k}}\end{cases}$ (9)

$p_{f,k}$ is the filling probability for the kth defect. p_f is related to the spatial and energy location of the defect which is uniformly distributed between 0 to 1. r is a random number generated for each simulation run to determine the filling status of each defect.