nature portfolio

Peer Review File

Flexible Self-rectifying Synapse Array for Energy-efficient Edge Multiplication in Electrocardiogram Diagnosis

Corresponding Author: Professor Kyung Min Kim

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Comment to authors
This manuscript proposes to demonstrate real-time edge electrocardiogram diagnosis using a self-rectifying synapse array and corresponding peripheral circuits. Utilizing self-rectifying synapse arrays to accelerate vector-matrix multiplication in neural network training and inference has been proven to be highly efficient and is a promising solution. The main contributions of this manuscript are (i) the fabrication of a self-rectifying synapse array, (ii) a hardware-aware training approach to address nonlinear current issue, and (iii) experimental demonstrations of ECG diagnosis. However, after carefully reviewing the manuscript, I identified several issues. I consider that it cannot be published in Nature Communications until these issues can be thoroughly resolved. I hope these issues will help improve the manuscript future.
1. The Ti/Al2O3/NbOx/Ta2O5/Pt self-rectifying device and the 1kb array (except for replacing the substrate material) have already been published in the author's previous article in Advanced Science (Ref. 22). Plenty of device data is duplicated in Fig. 2 and Fig. 3 which seriously detracts from the novelty of this paper. If improvements exist, the author should elaborate on exactly what is new for optimizing key performance at the device-level.
2. The device-to-device distribution of high and low resistance states for 10 devices seems to be large in Fig. 3a. How about the D2D variation for more devices, even for the whole array? How about the yield of the 1kb array?
3. The on/off ratio in Fig. 3b is significantly larger than that in Fig. 3a and Fig. 2h. What is the reason?
4. The reviewer didn't quite understand the meaning of Fig. 3c. Is it that each color line represents five groups of data under various curvature radii?
5. The foundation of the hardware-aware training approach is device followed Schottky conduction instead of tunneling or Poole-Frenkel transport. The authors should provide precise proof.
6. What is the reason for introducing a constant k in equation (3)? How are the b and c in Supplementary Table 1 calculated?
7. The authors should indicate the source of the dataset in Supplementary Fig. 3. It should be ensured that the dataset is large enough to guarantee the accuracy of the calculation of the parameters.
8. Is the conversion between V and V' accomplished by setting up additional functional circuits? There is no corresponding description found in the manuscript.
9. Usually the training part consumes much more energy than inference. This manuscript performs the inference part, is it expected that the method will also be used in the training part to reduce energy consumption?
10. What are the full names of these five categories in Fig. 5? Is there a correspondence between the categories of this manuscript and Ref. 42?
11. Are there any innovations in weight update and deployment at the application level?
12. On page 10, are the 2kb and 4kb arrays already being fabricated? If not, I would suggest that the authors avoid assessing the metric based on a 2kb array that was not fabricated.
13. Any assessment of the metrics when the array linewidth is minimized to 100nm is not reasonable. In the case of read latency simulation, when the size is scaled down, the authors only considered the change in capacitance and not the variation in resistance. The same problem exists with the assessment of energy consumption, and the data may be incorrect on an order of magnitude. The authors should provide details of the intermediate process. This is not a core point of the manuscript and authors are advised not to highlight it inside the abstract and introduction.
Minor issue:

1. Incorrect citation on Page 4 Line 113. Not subscripted of NbOx in Page 4 Line 118.
2. Too many new abbreviations are introduced, such as MCA, CTM, HCS, LCS, etc. which interferes with the reading.
3. Does the endurance in Fig. 2h refer to pulse or DC operation? How are the parameters set?
4. High operating voltage leads to high energy consumption, have the authors considered any solutions?

Reviewer #2

(Remarks to the Author)
This paper reports an energy-efficient neuromorphic applications utilizing the f-MDPE and the device exhibits mechanical durability up to a 5-mm bending curvature, making it suitable for wearable. This work also introduces a hardware-aware training approach to address non-Ohmic behavior during VMM. I recommend a major revision of this paper. Some comments as below:

1. The retention of the device seems very poor especially at multiple states, only holding 10^3s at 85 degree. As time increases, the individual resistance states gradually overlap, which has an impact on the final inference accuracy. When will the accuracy degrade to an intolerable level?

2. This work proposes a hardware-aware training approach to solve the problem of I-V nonlinearity. In real-time edge processing, the pre-trained weights g' could be stored in the crossbar, but the input signal needs to be preprocessed to obtain V', which requires additional cost. Is it more fair to consider the overheads of this step when comparing with other works?

3. In many works, the power of the ADC accounts for a significant portion of the total power in VMM. But in Table S5 of this work, the energy of ADC is almost negligible. Please explain in detail how the data is calculated, and the information of the ADC used.

4. The demonstrated network models (32x5) and the ECG dataset are quite simple. To more convincingly demonstrate the effectiveness of the proposed method, more complex simulated or experimental network models and datasets, such as MNIST or CIFAR10, are suggested. Furthermore, it is only demonstrated that the proposed real-time ECG diagnostic system can detect normal signals "N" which are observed most frequently and it is unclear whether it can be used for the detection of other signals, i.e., "S", "V", "Q" and "F".

5. It is better to provide reliability study of proposed real-time ECG diagnostic system constructed with the f-MDPE array, because the medical scenario has a very high standard of reliability.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The author has addressed most of my issues, but there are still a few below.

1. I'm very curious as to the root reason for the different on/off ratios under DC I-V sweeps and pulses, which has occurred in many works.

2. For Schottky and P-F transport, the only difference in the equations for the functions of I and V is the temperature T, so variable temperature measurements may be required to verify exactly which relationship is met.

3. For Comment #6, I suggest that the author should add revision length to allow readers to understand it quickly.

4. For Comment #8, is the circuit of the controller integrated on the PCB board or is it realized through simulation? Besides, the authors should clarify exactly which are obtained by hardware testing and which are obtained by software simulation in the whole application process.

5. What do the authors consider to be the basic factors limiting the scale-up of arrays at both the experimental preparation and hardware testing levels?

6. For Comment #13, “On the other hand, the latency is more related to the wire capacitance than to the device itself.”, this expression is obviously problematic. When the device resistance is large, the line capacitance has rather less effect on the latency than the device itself.

7. The author has omitted Comment #14.

8. The endurance tests in Fig. 2h lack of more data points throughout the testing period and more devices comparison which can provide a clearer understanding of the device's endurance behavior. Adding more intermediate data points would help to better visualize and confirm the consistency and reliability of the switching performance over the entire testing period. For

the methodology, I recommend referring to “https://doi.org/10.1063/5.0227603”.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Review attached.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I think the authors have made proper revision with additional data and discussion to address previous concerns.
This study could provide a impressive demonstration of self-rectifying memristors with application in edge medical computing scenario.
I suggest potential publication as it is.

Reviewer #3

(Remarks to the Author)
The manuscript is well modified according to the comments. This reviewer suggests the acceptance of the work in its current form.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript proposes to demonstrate real-time edge electrocardiogram diagnosis using a self-rectifying synapse array and corresponding peripheral circuits. Utilizing self-rectifying synapse arrays to accelerate vector-matrix multiplication in neural network training and inference has been proven to be highly efficient and is a promising solution. The main contributions of this manuscript are (i) the fabrication of a self-rectifying synapse array, (ii) a hardware-aware training approach to address nonlinear current issue, and (iii) experimental demonstrations of ECG diagnosis. However, after carefully reviewing the manuscript, I identified several issues. I consider that it cannot be published in Nature Communications until these issues can be thoroughly resolved. I hope these issues will help improve the manuscript future.

[Author Response]

Thank you for your thoughtful comments. We have carefully reviewed the comments and prepared one-by-one responses. We hope that our responses will address the reviewer's suggestions appropriately.

Comment #1: *"The Ti/Al2O3/NbOx/Ta2O5/Pt self-rectifying device and the 1kb array (except for replacing the substrate material) have already been published in the author's previous article in Advanced Science (Ref. 22). Plenty of device data is duplicated in Fig. 2 and Fig. 3 which seriously detracts from the novelty of this paper. If improvements exist, the author should elaborate on exactly what is new for optimizing key performance at the device-level."*

[Author Response]

We appreciate your careful consideration of our work. As the reviewer indicated, our device structure is based on the structure we previously proposed. However, for this study, we made three major modifications to the device fabrication process.

The first is the modification related to process temperature constraints. The original process involved high-temperature processing above 250 °C, which could not be applied to flexible polyimide (PI) substrates sensitive to high temperatures. Therefore, we had to lower the overall process temperature to below 180 °C, requiring all high-temperature processes to be adjusted accordingly. This reduction in process temperature caused changes in thin-film composition

and deposition rates, which necessitated significant efforts to optimize. As a result, we were able to achieve performance comparable to the original structure at lower processing temperatures.

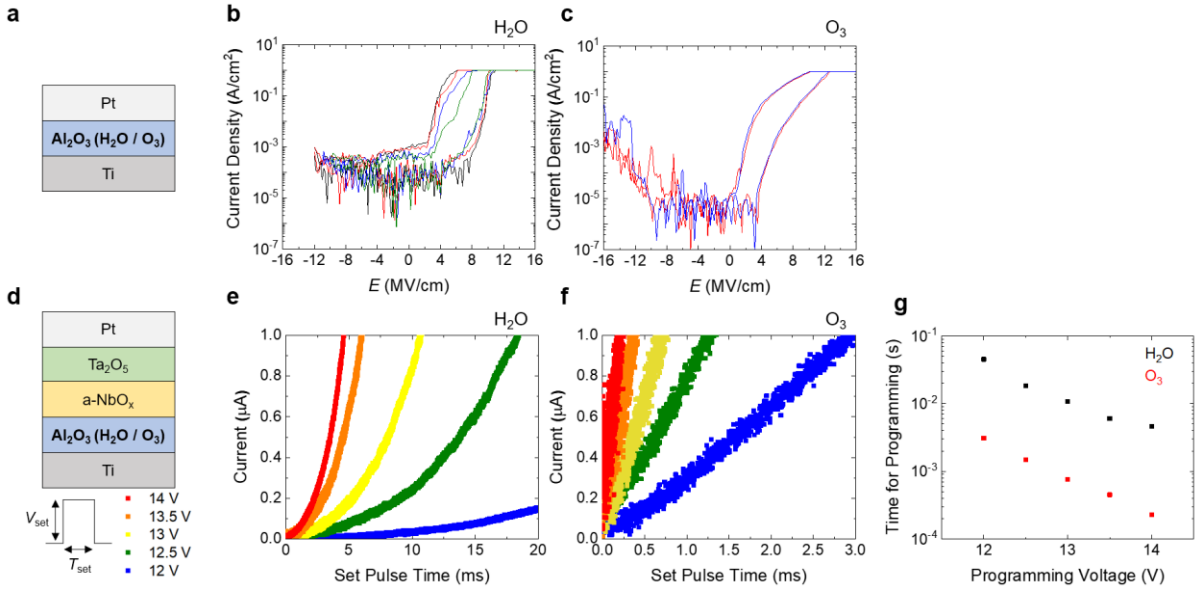
Second, we changed the reactant of the tunneling layer (Al_2O_3) to enhance programming speed. While the previous charge trap memristor could operate at low currents, it required relatively high voltage and took a long time to program. We made many attempts to improve this, and as a result, we modified the reactant from H_2O (as previously reported in Adv. Sci.) to O_3 (this work) and achieved significant improvements in programming speed. This improvement is attributed to the superior charge injection capability of the O_3-Al_2O_3 layer.

To support the improvement effects of the O_3-Al_2O_3 application, we conducted the following additional experiments. First, we fabricated Pt/H_2O-Al_2O_3/Ti and Pt/O_3-Al_2O_3/Ti capacitors (Supplementary Fig. 1a) and compared their *I-V* characteristics (Supplementary Figs. 1b and 1c). Interestingly, both devices exhibited charge trap memristor characteristics even though they included only a single Al_2O_3 layer, which is attributed to some charges being trapped within the Al_2O_3 layer. However, the two devices showed a significant difference in switching voltages, with the O_3-Al_2O_3 device switching at a lower voltage. This indicates that charge trapping occurs more easily in the O_3-Al_2O_3 device. We attribute this difference to the lower hydrogen content in the O_3-Al_2O_3 layer. The H_2O-Al_2O_3 inherently contains a higher amount of hydrogen, which passivates trap sites within the amorphous Al_2O_3, hindering the tunneling effect through it. In contrast, O_3-Al_2O_3 reduces the hydrogen content, thereby enhancing the tunneling effect [Y. Xiang et al., *Nanoscale Res. Lett.* **2015**, 10, 137], [K.-C. Ok et al., *IEEE Elec. Dev. Lett.* **2015**, 36, 9], [D. N. Goldstein et al., *J. Phys. Chem.* **2008**, 112, 19530].

Also, we measured the *I-t* graph of the full stack device under various voltages (Supplementary Fig. 1d). The results showed that the programming speed with H_2O-Al_2O_3 (Supplementary Fig. 1e) was enhanced after using O_3-Al_2O_3 (Supplementary Fig. 1f). In conclusion, this oxidant modification significantly improved the operational speed of the charge trap memristor array (Supplementary Fig. 1g) and reduced the energy required for learning.

We added the above explanations in Supplementary Fig. 1 as follows.

Revised Supplementary Information (Page 3):



Supplementary Figure 1 Modifying tunneling layer for enhanced programming speed. (a) A device architecture with a single Al_2O_3 layer and (b,c) *J-E* curves for (b) H_2O-Al_2O_3 and (c) O_3-Al_2O_3 device. Interestingly, both devices exhibited charge trap memristor characteristics even though they included only a single Al_2O_3 layer, which is attributed to some charges being trapped within the Al_2O_3 layer. However, the two devices showed a significant difference in switching voltages, with the O_3-Al_2O_3 device switching at a lower voltage. This indicates that charge trapping occurs more easily in the O_3-Al_2O_3 device. We attribute this difference to the lower hydrogen content in the O_3-Al_2O_3 layer. The H_2O-Al_2O_3 inherently contains a higher amount of hydrogen, which passivates trap sites within the amorphous Al_2O_3, hindering the tunneling effect through it. In contrast, O_3-Al_2O_3 reduces the hydrogen content, thereby enhancing the tunneling effect. (d) The charge trap memristor device stack, and (e,f) the programming transient current for (e) H_2O-Al_2O_3 and (f) O_3-Al_2O_3-based charge trap memristor under various V_{set}. (g) The programming time for each device to program the device to 1 μA by various V_{set}. This enhancement in programming speed will significantly improve the operational speed of the charge trap memristor array and reduce the energy required for learning.

The third modification was introducing a Ti/Pt/Ti structure for the bottom electrode to enhance the flexibility of the Ti layer. This was already mentioned in the original manuscript, but we have emphasized this point further with the two modifications above.

In summary, we added the following paragraph to the revised manuscript.

Revised Manuscript (Page 4):

We have previously proposed this structure, but it was on a SiO_2/Si substrate[22], so the device itself may not be novel. However, to implement this structure on a flexible substrate and overcome the slow operation speed of conventional devices, we introduced three key modifications in this study. First, since the PI substrate is vulnerable to high temperatures, we developed and applied low-temperature processes below 180°C for all layers. Second, we replaced the oxidant for the Al_2O_3 tunneling layer from H_2O to O_3 to reduce the H-passivated traps[30-33], thereby improving the programming speed (see Supplementary Fig. 1 in the supplementary information (SI) for a detailed discussion on the changes in the Al_2O_3 deposition process and their impact on the programming speed). Third, we inserted a 30-nm-thick Pt layer in the middle of the Ti layer (i.e., Ti/Pt/Ti) to enhance the flexibility of the bottom electrode (BE), as Ti alone is mechanically brittle.

Revised Reference (Page 16):

30. Xiang, Y., et al. Oxidation precursor dependence of atomic layer deposited Al_2O_3 films in a-Si:H(i)/Al_2O_3 surface passivation stacks. *Nanoscale Research Letters* **10**, 1 (2015).

31. Ok, K. C., et al. Effect of alumina buffers on the stability of top-gate amorphous InGaZnO thin-film transistors on flexible substrates. *IEEE Electron Device Letters* **36**, 917 (2015).

32. Goldstein, D. N., McCormick, J. A., & George, S. M. Al_2O_3 atomic layer deposition with trimethylaluminum and ozone studied by in situ transmission FTIR spectroscopy and quadrupole mass spectrometry. *The Journal of Physical Chemistry C* **112**, 19530 (2008).

33. Yeom, H. I., et al. High-performance oxide thin-film diode and its conduction mechanism based on ALD-assisted interface engineering. *Journal of Materials Chemistry C* **11**, 1336 (2023).

Comment #2: *"The device-to-device distribution of high and low resistance states for 10 devices seems to be large in Fig. 3a. How about the D2D variation for more devices, even for the whole array? How about the yield of the 1kb array?"*

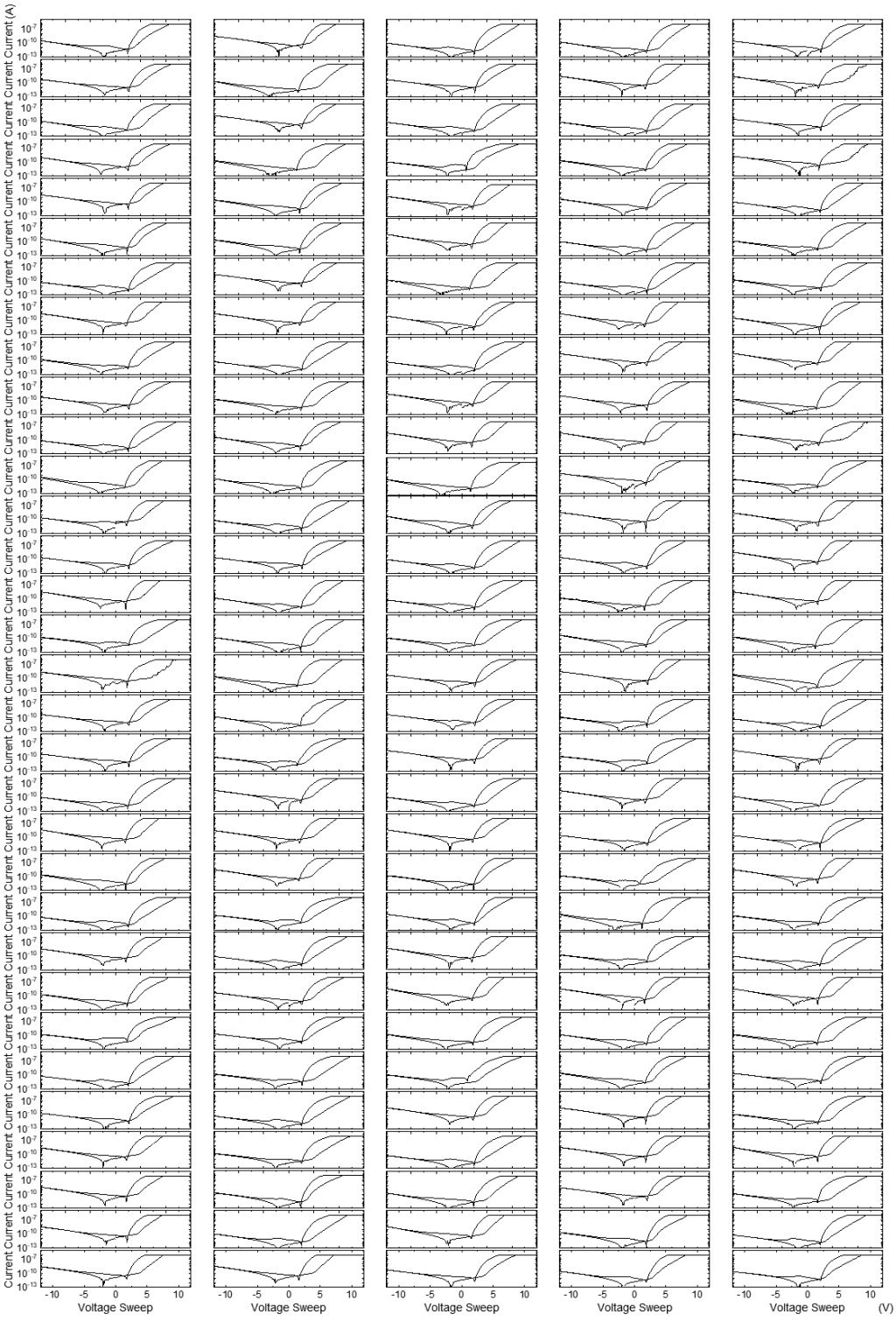
[Author Response]

We thank the reviewer for the insightful comment. We agree with the reviewer's suggestion and have accordingly added device-to-device (D2D) data from the array. In this study, we fabricated a flexible array capable of sufficiently running a 32×5 SLP and demonstrated its performance. Therefore, we conducted additional measurements of 160 *I-V* characteristics across the array corresponding to the size of the SLP. These measurements were performed using DC sweeps, with a compliance current set at 1 μA.

The results showed that all 160 devices worked, achieving a 100% yield. Since we utilized conductance states from 0.1 nS to 15 nS for the neural network demonstration, we would like to emphasize that the variation of the HRS below 100 pA does not affect the recognition rate.

We have added the measurement data of 160 devices, demonstrating D2D uniformity and yield, to Supplementary Fig. S3 as follows.

Revised Supplementary Information (Page 5):



Supplementary Figure 3 The *I-V* curves of 160 devices in the f-MDPE array. The voltage was first swept from 0 V to +12 V and back to 0 V, followed by a sweep from 0 V to −12 V and back to 0 V. A compliance current of 1 μA was used during the DC *I-V* measurements.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 4):

Fig. 2d shows *I-V* curves of 160 cells (32×5) in the array at an I_{CC} of 10^{-6} A, confirming a 100% yield and high array-level uniformity (see Supplementary Fig. 3 in the SI for the 160 raw *I-V* curves).

Revised Figure (Fig. 2d):

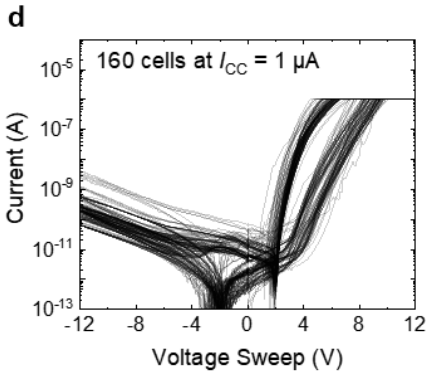


Fig. 2d *I-V* curves of 160 cells in the array under 1 μA of I_{CC}.

Comment #3: *"The on/off ratio in Fig. 3b is significantly larger than that in Fig. 3a and Fig. 2h. What is the reason?"*

[Author Response]

We are grateful for your thoughtful question. The on/off ratio in Fig. 3a and Fig. 3b are almost identical. They may look differently due to the greater data scatter in Fig. 3a, as it contains more data points.

Meanwhile, as the reviewer indicated, the on/off window in Fig. 2h (and also in Figs. 4, 5, and 6) is narrower than in Fig. 3. This difference arises from whether programming is performed using DC *I-V* sweeps or pulses. When using DC *I-V* sweeps, the full range of the on/off ratio can be obtained, resulting in a larger on/off ratio, as seen in Fig. 3. On the other hand, when programming with pulses, obtaining the full range of the on/off ratio would take too long time. Therefore, we considered the on/off range achievable with a 5 ms pulse as the practically available range, which falls between target conductance. In the experimental demonstration in Figs. 4-6, we also used this practical on/off range.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 5):

Additionally, Fig. 2h shows the endurance of the device for up to 10^5 cycles, using 5 ms of +12 V pulse for programming and 5 ms of −10 V pulse for erasing, respectively. Here, the on/off ratio was smaller than in Fig. 2g, where a DC sweep was used, due to the shorter pulse duration. Although this is not the full on/off ratio, we still achieve an on/off ratio of approximately 70, which we believe is sufficient for verifying the endurance characteristic.

Comment #4: *"The reviewer didn't quite understand the meaning of Fig. 3c. Is it that each color line represents five groups of data under various curvature radii?"*

[Author Response]

We are grateful to the reviewer for pointing out the clarity issue in our manuscript. Yes, we first programmed the device into specific conductance states and read the programmed states without bending at +3 V for 200 s. Then, we repeated the +3 V, 200 s measurements while gradually increasing the bending radius to 15 mm, 10 mm, 7.5 mm, and 5 mm. We collected this bending retention data in 8 conductance states by repeating this method.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 5):

Fig. 3c shows the high stability of multiple conductance states while changing curvatures from flat to r = 5 mm for 200 s at each curvature (total 10^3 s). For this measurement, we first programmed the device into specific conductance states and then measured the retention for 200 s at each bending radius while gradually decreasing the radius to 15 mm, 10 mm, 7.5 mm, and 5 mm. We repeated this measurement for 8 different conductance states.

Revised Figure (Fig. 3c):

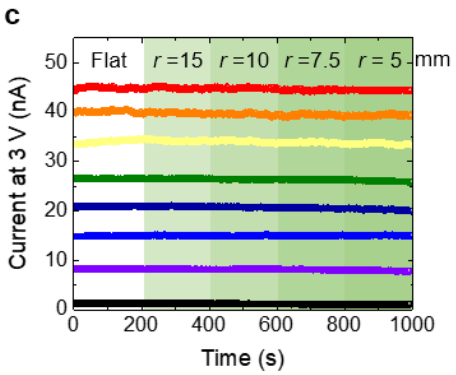


Fig. 3c Retention characteristics of 8 conductance states and **d** potentiation and depression under various curvature radii from flat to r = 5 mm.

Comment #5: *"The foundation of the hardware-aware training approach is device followed Schottky conduction instead of tunneling or Poole-Frenkel transport. The authors should provide precise proof."*

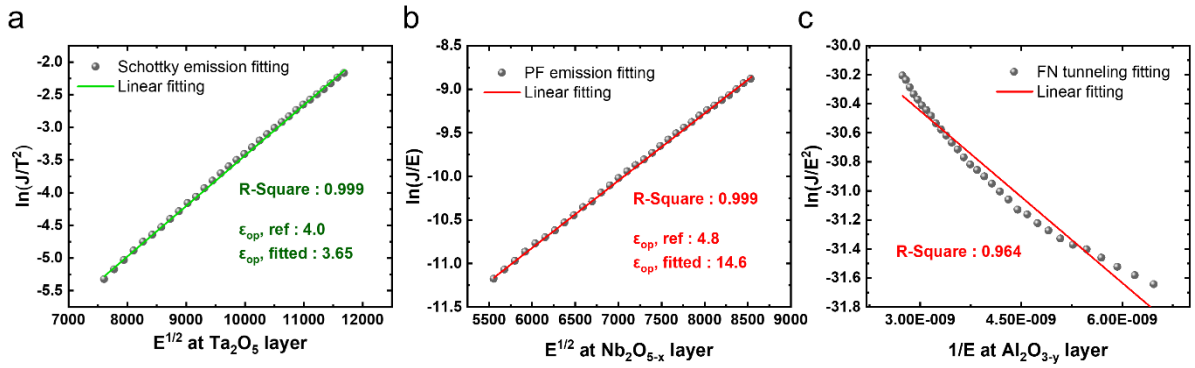
[Author Response]

We thank the reviewer for highlighting the need for precise proof regarding the conduction mechanism. We conducted additional conduction mechanism analysis to provide precise proof of Schottky conduction.

In this revision, to further clarify our Schottky conduction model, we extracted the optical dielectric constant (ε_{op}) based on the linear fitting results of possible conduction mechanisms in the dielectric layers. For Schottky conduction, as analyzed in our previous study [*Adv. Sci.,* 2023, 10, 2205654], we confirmed a linear fitting result with a reasonable ε_{op} value, indicating that the dominant conduction mechanism follows Schottky conduction at the Ta_2O_5 interface. Additionally, Poole-Frenkel emission through the NbO_x layer was also considered, and although a linear fitting result was observed, the extracted ε_{op} value was significantly higher than the reference value. Lastly, Fowler-Nordheim tunneling fitting was conducted but did not show a linear fitting result. These fitting results support the conclusion that the primary conduction mechanism in the device follows Schottky conduction.

We revised the conduction mechanism fitting results in Supplementary Fig. 5 as follows.

Revised Supplementary Information (Page 7):



Supplementary Figure 5 Fitting analysis of possible conduction mechanisms in the f-MDPE. Fitting results of a) Schottky emission ($\ln(J/T^2)$ vs. $E^{1/2}$) at the Ta_2O_5 layer, b) Poole-Frenkel emission ($\ln(J/E)$ vs. $E^{1/2}$) at the NbO_x layer, and c) Fowler-Nordheim tunneling ($\ln(J/E^2)$ vs. $1/E$) at the Al_2O_3 layer. For the fitting, the LRS *I-V* curve at room temperature was fitted in the voltage range of +3.5 V to +6.5 V. The local electric field in each layer was calculated using the thickness and dielectric constant of the oxide layers. The thicknesses were 6 nm for Ta_2O_5, 25 nm for NbO_x, and 7 nm for Al_2O_3. The dielectric constants were 24 for Ta_2O_5, 45 for NbO_x, and 9 for Al_2O_3[5-7]. The linear fitting results, along with the reasonable optical dielectric constant (ε_{op}) value, support that the primary conduction mechanism in this device is Schottky conduction at the Ta_2O_5 interface. In case of the Poole-Frenkel fitting, ε_{op} was significantly higher than the reference value. In case of the FN tunneling, the linear fit was not satisfactory.

Revised References in Supplementary Information:

5. H. C. Huang, T. E. Hsieh, *J. Appl. Polym. Sci*. 117, 1960 (2010).

6. N. Fuschillo, B. Lalevic, N. Annamalai, W. Slusark Jr., *J. Non-Cryst. Solids.* 22, 159 (1976)

7. H. Birey, *J. Appl. Phys*. 48, 5209 (1977)

Also, we elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 6-7):

The first step of this process is to determine the non-Ohmic function. The f-MDPE device followed Schottky conduction (see Supplementary Fig. 5 in the SI for the detailed conduction mechanism fitting results). Therefore, the I-V characteristics of the device can be expressed by the following Schottky equation.

Comment #6: *"What is the reason for introducing a constant k in equation (3)? How are the b and c in Supplementary Table 1 calculated?"*

[Author Response]

We appreciate the opportunity to provide further details about our approach. In our hardware-aware training that addresses the Schottky conduction mechanism in our charge trap memristor device, the current flow through the device can be expressed as $I(a,V) = \exp(a) \cdot \exp(b\sqrt{V-c})$. Here, a is a variable representing the device state relative to the height of the Schottky barrier, while b and c are constants. When applying this model in a neural network, the terms $\exp(a)$ and $\exp(b\sqrt{V-c})$ can be interpreted as the network's weight g' and input V', respectively.

					$k = 0$	$k = 7$
V'	**Input voltage (V)**		b	c	$V' = \exp(-k + b\sqrt{V-c})$	
	3.5		7.66	1.5	5.07E+4	4.62E+1
	⋮				⋮	⋮
	2.0				2.25E+2	2.05E-1
g'	**Analog states**	g_{ref} **(nS)**	a		$g' = \exp(a + k)$ **(× 10^9)**	
	16	15	-27.79064		8.52E-4	9.35E-1
	⋮	⋮	⋮		⋮	⋮
	1	0.1	-32.72548		6.52E-6	6.72E-3

Table R1. Comparison of g' and V' for compensation parameter $k = 0$ and $k = 7$ in hardware-aware training.

However, as shown in Table R1, directly using these expressions can lead to a significant scale mismatch between the weight $g' = \exp(a)$ and the input $V' = \exp(b\sqrt{V-c})$. Such a mismatch can cause several problems: significant differences in scale can result in unstable gradients during backpropagation, making it difficult to optimize network parameters effectively. If the scales are not aligned, the network might struggle to learn the underlying patterns due to imbalances in parameter magnitudes. Additionally, techniques like batch normalization, which rely on consistent scales for stability and efficiency, could become less effective. To address these issues, we introduced a compensation parameter k, adjusting the expressions to $g' = \exp(a + k)$ and $V' = \exp(-k + b\sqrt{V-c})$. This aligns the weight and input scales, preventing the problems mentioned above.

Overall, introducing the constant k is a crucial step in our hardware-aware training approach, addressing potential issues with scale consistency and ensuring that the neural network operates within a stable and efficient regime.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 7):

Here, we put a compensation constant k that scales the value of each exponential term accordingly, addressing potential issues with scale consistency and ensuring that the neural network operates within a stable and efficient regime.

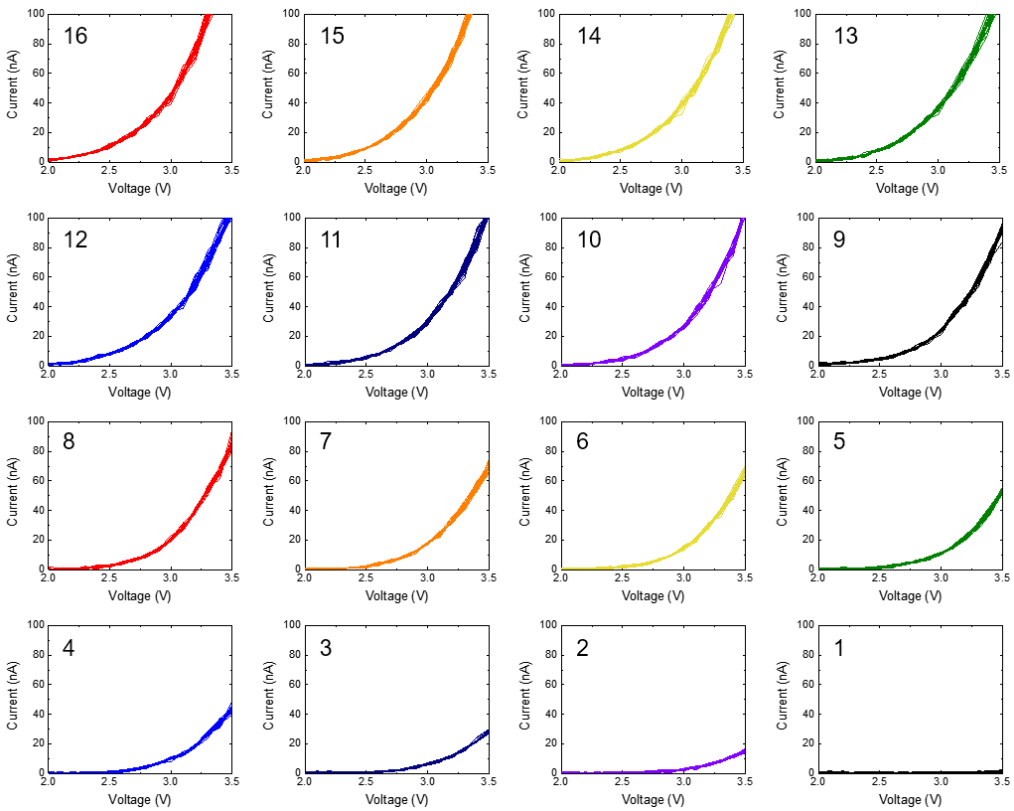
Comment #7: *"The authors should indicate the source of the dataset in Supplementary Fig. 3. It should be ensured that the dataset is large enough to guarantee the accuracy of the calculation of the parameters."*

[Author Response]

We thank the reviewer for this comment regarding clarifying the dataset details. We have modified Supplementary Fig. 3 with 20 *I-V* curves for each state that we used for extracting hardware parameters *a*, *b*, and *c* (a total of 320 curves). Each state was stably programmed, and we confirmed that Schottky fitting was well achieved. Further, we modified Supplementary Fig. 9 corresponding to the 20 *I-V* curves for linear fitting.

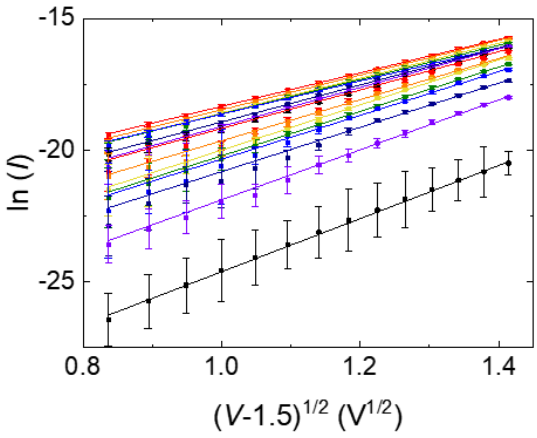
Therefore, we have revised Supplementary Information to reflect these changes.

Revised Supplementary Information (Page 11):



Supplementary Figure 8 *I-V* curves of 16 states at the read voltage region (+2 V ~ +3.5 V). Each curve is obtained from 20 devices, after programming to reach the target conductance at +3 V.

Revised Supplementary Information (Page 12):



Supplementary Figure 9 Ln(I) vs. $(V\text{-}1.5)^{1/2}$ plots for analog states to the Schottky emission equation corresponding to Supplementary Fig. 8.

We also elaborated on it in the manuscript as below.

Revised Manuscript (Page 8):

Fig. 4f shows the read currents for the 16 quantized states at various input voltages (square dots) alongside the Schottky fitting curves (solid lines) (see Supplementary Fig. 8 in the SI for the *I-V* curves of 16 states obtained by 20 devices demonstrating the stability of each state, and see Supplementary Fig. 9 and Supplementary Table 2 for the Schottky fitting results). For hardware-aware training, hardware parameters for each state are extracted from the fitting curves, and g' values are obtained from the hardware parameters.

Comment #8: *"Is the conversion between V and V' accomplished by setting up additional functional circuits? There is no corresponding description found in the manuscript."*

[Author Response]

We appreciate the reviewer for pointing out the need for further clarification. The input voltage V is actually generated by sensing an external signal, which is then converted into the input V through the controller's conversion functions. As such, for training, V' can be generated directly from the controller. The controller may include exponential converting circuits [H. Chamberlin, *Musical applications of microprocessors*, p. 194] for this conversion process, or it can also use software for the conversion function. In any case, they can be part of the controller and do not pose a significant burden on the overall circuit.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 7-8):

Here, V were transformed to V' using the hardware parameters, retaining the features of the original signal (see Supplementary Fig. 7 for an example of the transformation from V to V'). Note that both V and V' are directly generated by receiving external signals through a controller unit. Therefore, the conversion from V to V' does not impose any additional circuit burden.

Comment #9: *"Usually the training part consumes much more energy than inference. This manuscript performs the inference part, is it expected that the method will also be used in the training part to reduce energy consumption?"*

[Author Response]

We sincerely value your detailed review. We agree that using this device could indeed lead to energy savings during the training process. There are three major factors that influence energy consumption during training: the energy required for a weight update, the energy consumed per training epoch, and the total number of epochs.

First, the low inference energy of our device (Fig. 6f) is primarily due to its low operating current. Similarly, the programming energy of our device is much lower compared to other devices, which suggests that energy consumption for weight updates will also be reduced.

However, despite the low energy consumption of our device, the non-Ohmic characteristics induced additional energy consumption when calculating the weight gradient for gradient descent training. The differing parts, excluding the identical terms in the process of calculating the weight gradient, are summarized in the revised Supplementary Table 1 in below. A critical point here is the weight gradient value for updates during the training process. Unlike other models where the partial derivative of the output with respect to the weight is simply the input, non-Ohmic devices introduce an exponential term into the partial derivative, resulting in a more complex value. Since these complex calculations occur with every weight update, larger network sizes and more training epochs inevitably demand greater computational resources. At this point, if the computation of the output current can be transformed and expressed with a constant g' matrix, the partial derivatives needed for the weight update process can be simplified, solving the issues mentioned above. We believe this transformation technique will be very effective during the training process, and it is expected to provide a global solution for low-power operation using non-Ohmic memristors.

Further, it is necessary to examine whether hardware-aware training impacts the training process, such as requiring more training epochs. We evaluated the recognition accuracy as a function of training epochs, both with hardware-aware training (Schottky device) and without (Ohmic device), and found no significant difference between the two methods as described Figure R1 below. This indicates that hardware-aware training does not incur significant additional energy costs during the training process.

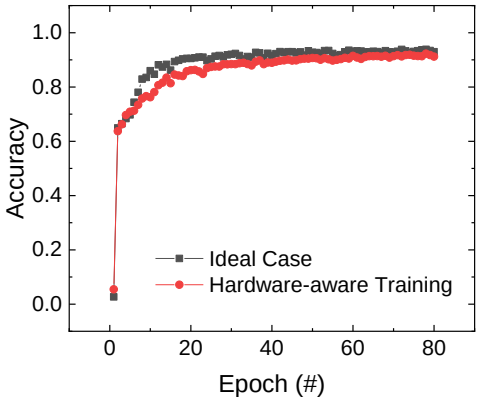
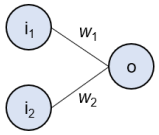
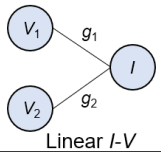
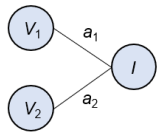
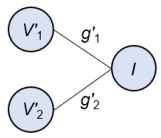


Figure R1 Accuracy (F1 macro score) comparison over training epochs between ideal case (Ohmic device) and hardware-aware training (Schottky device).

We have added the following supplementary table in Supplementary Information and revised the manuscript.

Revised Supplementary Information (Page 9):

	Model	Output (For inference)	Weight Gradient (For training)
Software (Ideal)		$o = w_1 i_1 + w_2 i_2$	$\frac{\partial o}{\partial w_1} = i_1$
Memristor (Ohmic)	 Linear *I-V*	$I = g_1 V_1 + g_2 V_2$	$\frac{\partial I}{\partial g_1} = V_1$
Our Device (Schottky)	 Non-linear *I-V*	$I_{sch}(a,V) = e^{a} e^{b\sqrt{V-c}}$ $I = I_{sch}(a_1,V_1) + I_{sch}(a_2,V_2)$	$\frac{\partial I}{\partial a_1} = e^{a_1} e^{b\sqrt{V_1-c}}$
	 Transformation	$I = g'_1 V'_1 + g'_2 V'_2$	$\frac{\partial I}{\partial g'_1} = V'_1$

Supplementary Table 1 Comparison of computing a neural network between software (ideal case), memristor (Ohmic conduction), and our device (Schottky conduction). By transformation, the partial derivatives are simplified enabling energy-efficient backpropagation.

Revised Manuscript (Page 7):

By transforming linear VMM, unnecessary additional partial derivatives of Schottky functions during backpropagation in the training process are avoided, enabling energy-efficient neural network learning. (See Supplementary Table 1 for the comparison of the network model, output form, and weight gradient function between three cases: ideal software case, hardware using Ohmic memristor case, and hardware using Schottky memristor case).

We have also added the following discussion to the conclusion to reflect this finding.

Revised Manuscript (Page 12):

In this study, the dataset was trained in software and subsequently mapped onto hardware. As a result, the inference energy was the primary focus, with the proposed f-MDPE demonstrating low inference energy due to its low-current operation. For the same reason, the f-MDPE exhibited low programming energy, and hardware-aware training does not slow down the overall training process. Therefore, the f-MDPE is expected to offer competitive energy consumption during on-chip training as well.

Comment #10: *"What are the full names of these five categories in Fig. 5? Is there a correspondence between the categories of this manuscript and Ref. 42?"*

[Author Response]

We appreciate the reviewer's inquiry regarding the details of the dataset used. The MIT-BIH arrhythmia database is a 30-minute electrocardiogram (ECG) signal obtained from 47 subjects studied at the BIH Arrhythmia Laboratory. This database classifies ECGs into five categories: Normal (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Fusion beat (F), and Unknown beat (Q). This classification corresponds to the categories used in reference 42, which is the categorized system of the MIT-BIH dataset.

We elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 8):

For training, we adopted the ECG dataset from the MIT-BIH arrhythmia database, which classifies the ECG signals into five categories[46]: Normal (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Unknown beat (Q), and Fusion beat (F).

Comment #11: *"Are there any innovations in weight update and deployment at the application level?"*

[Author Response]

Thank you for highlighting this point. When mapping the weights trained in software onto hardware, we used the standard Incremental Step Pulse Programming (ISPP)-type write and verification scheme. This method is not a new approach, and our previous research [W. H. Cheong et al., *Adv. Funct. Mater.* 2012, 32, 2200337] demonstrated that precise analog weight update can be achieved based on the current array measurement system.

We have added the following details to the manuscript accordingly.

Revised Manuscript (Page 9):

Next, we programmed the f-MDPE with the prepared weight matrix in Fig. 5b and examined its programmed accuracy. For this weight mapping, we used ISPP (Incremental Step Pulse Programming)-type programming and verification scheme to accurately program the cells[49]. Fig. 5d shows the measured currents of the programmed f-MDPE read at 3 V.

Revised Reference (Page 17):

49. Cheong, W. H., et al. Demonstration of Neuromodulation-inspired Stashing System for Energy-efficient Learning of Spiking Neural Network using a Self-Rectifying Memristor Array. *Advanced Functional Materials* **32**, 2200337 (2022).

Comment #12: *"On page 10, are the 2kb and 4kb arrays already being fabricated? If not, I would suggest that the authors avoid assessing the metric based on a 2kb array that was not fabricated."*

[Author Response]

We understand the reviewer's concern and have carefully considered the implications of evaluating the performance of devices with sizes that we have not directly fabricated.

While it is true that we cannot currently implement arrays of 2k or 4k size with our testing platform, it is feasible to achieve this by operating multiple 1k-sized f-MDPEs in parallel. Also, the approach of expanding the array size based on actual device results to predict large-area scalability has been adopted in previous studies, such as *Nat. Electron.*, 2023, 6, 45-51, and *Adv. Mater.*, 2024, 36, 2309314. Therefore, we believe that using this method is reasonable and not overly speculative.

Therefore, rather than removing the simulation section for 2k and 4k arrays entirely, we have added a more detailed explanation of its significance and limitations to clarify its context and relevance to our work.

Revised Manuscript (Page 11):

Although the 2k and 4k arrays were not physically fabricated, they were simulated assuming that the 1k-sized f-MDPE array was directly scaled up. This virtual scaling approach is widely applied in other studies as well[13,52], as it provides valuable insights into large-scale performance with high reliability in advance.

Comment #13: *"Any assessment of the metrics when the array linewidth is minimized to 100nm is not reasonable. In the case of read latency simulation, when the size is scaled down, the authors only considered the change in capacitance and not the variation in resistance. The same problem exists with the assessment of energy consumption, and the data may be incorrect on an order of magnitude. The authors should provide details of the intermediate process. This is not a core point of the manuscript and authors are advised not to highlight it inside the abstract and introduction."*

[Author Response]

We appreciate your careful consideration of our work. As the reviewer indicated, we assumed a linear change in capacitance according to the area but we did not take resistance changes into account. This is because we wanted to estimate the energy consumption under the harshest conditions. Regarding the resistance, when the size of the memristor is reduced, operational current can decrease. However, if the operating current becomes too low, sensing margin problems can arise. Therefore, the device itself must be improved to maintain an appropriate operating current as scaling progresses. Considering these factors, we assumed the operating current of the device is unchanged after scaling. Please note that the operating current of the device is very low, so wire resistance remains negligible even after scaling.

On the other hand, the latency is more related to the wire capacitance than to the device itself. Therefore, in energy calculations, the reduction in energy caused by wire capacitance can be directly improved through scaling.

In summary, to focus on calculating the energy solely based on the latency reduction effect, we excluded the resistance change, assuming the operating current is unchanged.

However, as the reviewer suggested, this energy consumption estimation is an approximation, and we agree that making definitive statements about it could be misleading. Therefore, we deleted the corresponding statements in the abstract and introduction as below.

Revised Abstract (Page 1):

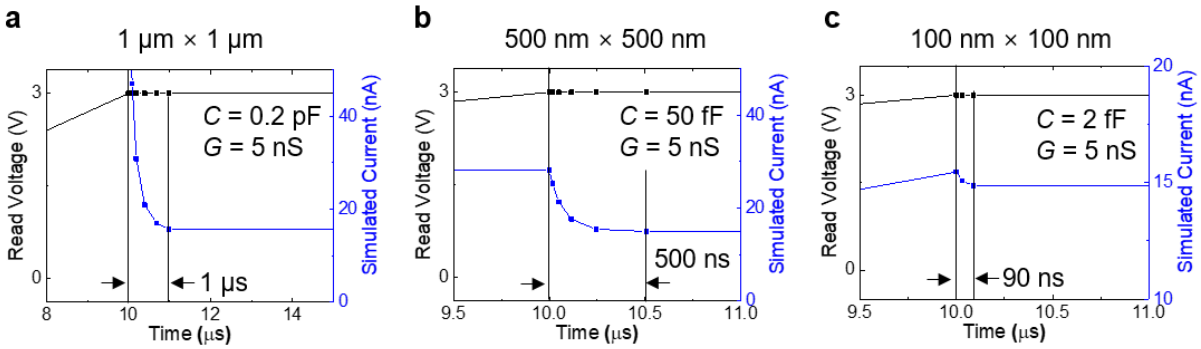
This approach achieved an ECG classification accuracy of 93.5%, while consuming only 0.3% of the energy compared to digital approaches~~, Furthermore, our simulations indicated that the energy reduction could be further reduced to 0.001% through device scaling to a 100-nm-line width,~~ highlighting the strong potential of this approach for emerging edge neuromorphic hardware.

Revised Manuscript (Page 3):

Then, we conducted laboratory-scale real-time ECG diagnosis using the f-MDPE, demonstrating its potential for in-situ ECG inference, as it consumes only 0.3% of the energy compared to digital approaches~~. Furthermore, our simulation proposed that scaling to 100 nm may achieve less than 0.001% energy consumption compared with digital approaches~~, with a classification accuracy of 93.5%, highlighting its potential for edge AI hardware.

Additionally, we have supplemented the explanation energy consumption calculations in our scaling analysis as follows.

Revised Supplementary Information (Page 22):



Supplementary Figure 15 The simulated read latency of the scaled-down MDPE was investigated for various device sizes; (a) 1 μm × 1 μm, (b) 500 nm × 500 nm, and (c) 100 nm × 100 nm. These simulations were also conducted utilizing LTSpice, incorporating parallel capacitors and considering the reduction in capacitance following the geometric scaling-down effect. Note that we assumed *G* remains unchanged by scaling. This evaluation only considers the effects of latency improvement. In reality, *G* would also decrease with scaling, reducing energy consumption. However, if *G* becomes smaller, it could affect the device's sensing margin and alter its operational conditions. Therefore, it is reasonable to assume that *G* may not change due to improvements in the device, even with scaling.

Minor issue:

Comment #M1: *"Incorrect citation on Page 4 Line 113. Not subscripted of NbOx in Page 4 Line 118."*

[Author Response]

We corrected the manuscript as follows.

Revised Manuscript (Page 4):

The NbO_x layer is non-stoichiometric and amorphous, which is a key factor that acts as a charge trap layer[22].

Comment #M2: *"Too many new abbreviations are introduced, such as MCA, CTM, HCS, LCS, etc. which interferes with the reading."*

[Author Response]

To enhance readability, we have minimized the use of abbreviations in the text. Additionally, we have standardized HCS and LCS to HRS (high resistance state) and LRS (low resistance state), as these terms are more commonly recognized in the field of memristors. All the revisions we made are highlighted below.

Revised Manuscript (Page 2):

One of the most notable technologies is the memristive dot product engine (MDPE), which performs large-scale vector-matrix multiplication (VMM) operations using a memristive crossbar array ~~(MCA)~~[6-13]. The MDPE is fundamentally very energy-efficient, as it performs VMM operations based on physics principles (i.e., Kirchhoff's law and Ohm's law). Additionally, flexible crossbar arrays have been actively investigated recently, and they have potential in diverse edge applications, including image recognition[14], classifying input data measured from sensors[15-17], and diagnosing bio-signals such as ECGs[18,19].

Revised Manuscript (Page 2):

In MDPE fabrication, selecting a memristor is crucial, and we suggest that a charge trap memristor ~~(CTM)~~ is the best choice based on its various proven advantages[20-26].

Revised Manuscript (Page 2):

In addition, the fabrication of charge trap memristor-based MDPE on flexible substrates should proceed at a low temperature because a high thermal budget destroys the substrate, which has not been demonstrated.

Revised Manuscript (Page 3):

In this study, we introduce the charge trap memristor-embodying flexible MDPE (f-MDPE) and propose its utilization in an edge AI system for ECG diagnosis. We fabricated a 32×32 crossbar array on a polyimide (PI) substrate containing a low-temperature (< 180°C) processed charge trap memristor as the f-MDPE.

Revised Manuscript (Page 4):

The layered structure of the charge trap memristor (Ti/Pt/Ti/Al_2O_3/NbO_x/Ta_2O_5/Pt, from bottom to top) was integrated on a flexible PI substrate. More details on the device fabrication process can be found in the Method section.

Revised Manuscript (Page 5):

Fig. 2g shows the retention characteristics of the high resistance state (HRS) and low resistance state (LRS) for 10^4 s at different temperatures (RT, 85°C, and 125°C). The Arrhenius plot suggested that both the HRS and LRS could be maintained for more than 3 years (see Supplementary Fig. 4 in the SI for the Arrhenius plot with the failure time at 20% conductance change).

Comment #M3: *"Does the endurance in Fig. 2h refer to pulse or DC operation? How are the parameters set?"*

[Author Response]

The endurance measurements were performed using pulse operation. We have elaborated on this point in the revised manuscript as follows.

Revised Manuscript (Page 5):

Additionally, Fig. 2h shows the endurance of the device for up to 10^5 cycles, using 5 ms of +12 V pulse for programming and 5 ms of −10 V pulse for erasing, respectively.

Comment #M4: *"High operating voltage leads to high energy consumption, have the authors considered any solutions?"*

[Author Response]

As the reviewer indicated, lowering the operating voltage in charge trap memristors would be a highly interesting development. We have recently observed that inserting nanodots into the tunneling layer can enhance the electric field for charge trapping, reduce the operating voltage, and increase the switching speed. Although these results are still unpublished, and we will not go into detail here, reducing operating energy is a critical challenge. To highlight the importance of this issue, we will add a discussion on it in the conclusion section as follows.

Revised Manuscript (Page 12):

Lastly, while the charge trap memristor used in this study has the advantage of low-current operation, it still has drawbacks, such as higher operating voltage and slower switching speed compared to other memristor technologies. If these aspects can be improved in the future, a more energy-efficient solution can be provided, expanding the applicability of charge trap memristor technology to a broader range of applications.

Reviewer #2 (Remarks to the Author):

This paper reports an energy-efficient neuromorphic applications utilizing the f-MDPE and the device exhibits mechanical durability up to a 5-mm bending curvature, making it suitable for wearable. This work also introduces a hardware-aware training approach to address non-Ohmic behavior during VMM. I recommend a major revision of this paper. Some comments as below:

[Author Response]

We appreciate the reviewer for thoughtful review and comments. We hope that our revision responds well to the comments.

Comment #1: *"The retention of the device seems very poor especially at multiple states, only holding 10^3s at 85 degree. As time increases, the individual resistance states gradually overlap, which has an impact on the final inference accuracy. When will the accuracy degrade to an intolerable level?"*

[Author Response]

We appreciate your detailed observations and valuable comments. To verify whether the multi-level state is well-maintained at high temperatures, we conducted additional measurements at 85 ℃ for up to 10^4 s, but still, we could not observe any degradation.

The retention degradation mainly occurred at the highest conductance state. Therefore, we conducted a more systematic analysis of the highest conductance state retention at 85 ℃ and 125 ℃. Based on a 20% conductance loss criterion, we could conclude that the device can maintain the programmed state well for up to 3 years at room temperature, as described in Figure 2g and revised Supplementary Figure 4.

We elaborated on it in the revised manuscript as follows.

Revised Figure (Fig. 2f):

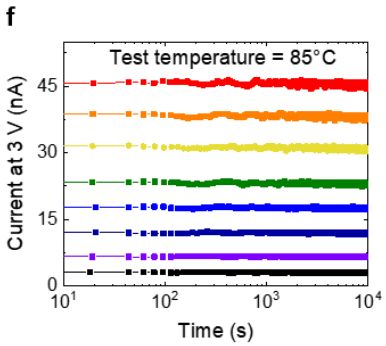


Fig. 2f Retention characteristics of 8 conductance states at 85°C.

Revised Manuscript (Page 5):

The retention characteristics of eight different conductance states were confirmed at 85°C, as shown in Fig. 2f. All conductance states were stable up to 10^4 s without significant changes. This gives rise to an operating conductance window from 0.1 nS to 15 nS at 3 V.

Comment #2: *"This work proposes a hardware-aware training approach to solve the problem of I-V nonlinearity. In real-time edge processing, the pre-trained weights g' could be stored in the crossbar, but the input signal needs to be preprocessed to obtain V', which requires additional cost. Is it more fair to consider the overheads of this step when comparing with other works?"*

[Author Response]

We appreciate your careful reading and the points you've raised. The use of V' was necessary during the training process, and we conducted *ex-situ* training using software. Therefore, this process did not involve any additional circuits and was performed using general-purpose software. During inference, V' was not utilized; instead, the actual ECG signal V was used. As a result, the in-situ inference presented in Figures 4 and 5 are not related to the use of V'.

Additionally, the preprocessed V' was needed for the training process, which required additional computing resources. Nevertheless, the transformation method was an effective approach in terms of training accuracy and energy consumption.

Devices with non- linear *I-V* characteristics do not have constant conductance, which means they cannot be directly matched to the weight values in conventional artificial neural networks. If we ignore these non-Ohmic characteristics, proper learning cannot take place as described Figure R1 below. To properly train a network using such non-Ohmic devices, we need to use a weight matrix composed of non-linear functions rather than a constant weight matrix.

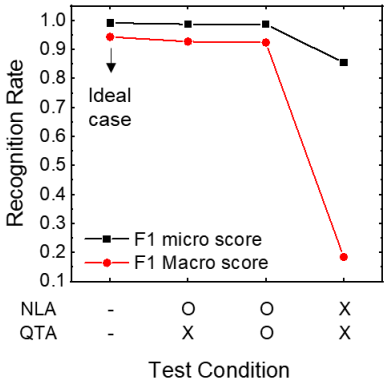


Figure R1 Accuracy (i.e., recognition rate) with/without hardware-aware training. Compared to the ideal case, accuracy drops significantly without hardware-aware training. When non-linearity aware (NLA) training was applied and both NLA and quantization aware (QTA)

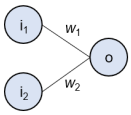
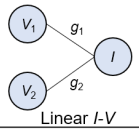
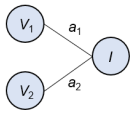
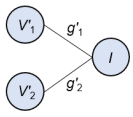
training were applied, the accuracy was as high as the ideal case, confirming the effectiveness of the hardware-aware training.

The important point here is the weight gradient value for weight updates during the training process. Unlike other models where the partial derivative of the output with respect to the weight is simply the input, in the case of non-Ohmic devices, the partial derivative includes an exponential term, resulting in a more complex value. The differing parts, excluding the identical terms in the process of calculating the weight gradient, are summarized in the revised Supplementary Table 1 in below. Since these complex calculations occur with every weight update, larger network sizes and more training epochs inevitably demand greater computational resources. Additionally, parameter scale is important when updating the neural network, and since the value corresponding to the weight is a part of the exponential term, it is challenging to adjust the scale appropriately.

At this point, if the computation of the output current can be transformed and expressed with a constant g' matrix, the partial derivatives needed for the weight update process can be simplified, solving the issues mentioned above. We believe this transformation technique will be very effective during the training process, and it is expected to provide a universal solution for low-power operation using non-Ohmic memristors.

We have added the following Supplementary Table in Supplementary Information and revised the manuscript.

Revised Supplementary Information (Page 9):

	Model	Output (For inference)	Weight Gradient (For training)
Software (Ideal)		$o = w_1 i_1 + w_2 i_2$	$\frac{\partial o}{\partial w_1} = i_1$
Memristor (Ohmic)	 Linear *I-V*	$I = g_1 V_1 + g_2 V_2$	$\frac{\partial I}{\partial g_1} = V_1$
Our Device (Schottky)	 Non-linear *I-V*	$I_{sch}(a,V) = e^{a} e^{b\sqrt{V-c}}$ $I = I_{sch}(a_1,V_1) + I_{sch}(a_2,V_2)$	$\frac{\partial I}{\partial a_1} = e^{a_1} e^{b\sqrt{V_1-c}}$
	 Transformation	$I = g'_1 V'_1 + g'_2 V'_2$	$\frac{\partial I}{\partial g'_1} = V'_1$

Supplementary Table 1 Comparison of computing a neural network between software (ideal case), memristor (Ohmic conduction), and our device (Schottky conduction). By transformation, the partial derivatives are simplified enabling energy-efficient backpropagation.

Revised Manuscript (Page 7):

By transforming linear VMM, unnecessary additional partial derivatives of Schottky functions during backpropagation in the training process are avoided, enabling energy-efficient neural network learning. (See Supplementary Table 1 for the comparison of the network model, output form, and weight gradient function between three cases: ideal software case, hardware using Ohmic memristor case, and hardware using Schottky memristor case).

Comment #3: *"In many works, the power of the ADC accounts for a significant portion of the total power in VMM. But in Table S5 of this work, the energy of ADC is almost negligible. Please explain in detail how the data is calculated, and the information for the ADC used."*

[Author Response]

We greatly appreciate the reviewer's comment. We have detailed the number of operations performed by each component when inferring a single ECG signal through the 1D CNN 2k model. We have conducted more accurate energy calculations.

We benchmarked the peripheral circuit assuming it is based on 28-nm CMOS process technology. Specifically, for the ADC, we referred to a 28-nm 8-bit ADC that consumes 2.01 fJ/conv. step of energy [T.-V. Cao *et al.,* NORCHIP 2012 IEEE, 1]. Additionally, when diagnosing the ECG using a 1D CNN with an 8-bit ADC, there is no accuracy drop [M. Saeed *et al.*, *IEEE Transactions on Biomedical Circuits and Systems*, 2021, 15, 1129-1139]. Therefore, the energy consumed by the 8-bit ADC per operation is 2.01×2^8 fJ. The energy consumption per operation is well matched with a reference using convolution layers [*Nat. Comm.*, 2024, 15, 1132]. The ADC is operated 4,273 times in our 2k model, and further details are provided in Revised Supplementary Table 5.

Based on the energy consumption and number of operations of each peripheral circuit, the energy required to run the 2k model with a 5-μm-width f-MDPE is 1.20×10^{-7} J, which is 0.29% of the energy consumed by a GPU. This has been updated in Revised Supplementary Table 6.

However, when assuming a 100-nm-width MPDE, the energy consumption in the peripheral circuit became dominant. Nevertheless, by reducing the energy consumption through the scale-down of the charge trap memristor device, we were able to significantly reduce the energy required for ECG diagnosis to 2.64×10^{-9} J, which is 0.006% of the energy consumed by a GPU and 2.19% of the energy consumed by the 5-μm-width f-MDPE. This suggests that further energy reduction can be achieved through the device scale-down strategy.

We elaborated on it in the revised Supplementary Information as follows.

Revised Supplementary Information (Page 20):

2k model		**Input register**	**MDPE**	**ADC**	**MUX**	**MaxPool**	**Output register**
	Array size	# of operation	# of multiplication	# of conversion	# of operation	# of operation	# of operation
Conv.1	21×4	300	25,200	1,200	1,200	0	300
Pool.1	9×1	300	2,700	300	300	0	300
Conv.2	23×4	300	27,600	1,200	1,200	0	300
Pool.2	6×1	288	1,728	288	288	0	288
Conv.3	25×2	288	14,400	576	576	0	288
Pool.3	6×1	192	1,152	192	192	0	192
Conv.4	27×2	192	10,368	384	384	0	192
Pool.4	4×1	128	512	128	128	0	128
FC layer	128×5	1	640	5	5	1	1
SUM		**1,989**	**84,300**	**4,273**	**4,273**	**1**	**1,989**

Supplementary Table 5 Operation counts for each component in the 2k model. In the 2k model, input register, analog-to-digital converter (ADC), multiplexer (MUX), max-pooling (MaxPool), and output register are considered as the parts of peripheral circuits. Note that max-pooling is used for final classification at the output neuron in the fully connected (FC) layer.

Revised Supplementary Information (Page 23):

Energy Consumption per Inference	**f-MDPE** (5 μm × 5 μm)		**Scaled-down MDPE** (100 nm × 100 nm)		**CPU**	**GPU**
	Energy (J)	Counts (#)	Energy (J)	Counts (#)		
Input register	1.55×10^{-14}	1,989	1.55×10^{-14}	1,989		
MDPE	1.40×10^{-12}	84,300	4.19×10^{-15}	84,300		
ADC	5.15×10^{-13}	4,273	5.15×10^{-13}	4,273		
MUX	5.71×10^{-15}	4,273	5.71×10^{-15}	4,273		
MaxPool	4.00×10^{-13}	1	4.00×10^{-13}	1		
Output register	1.55×10^{-14}	1,989	1.55×10^{-14}	1,989		
Total	1.20×10^{-7} J		2.64×10^{-9} J		1.74×10^{-4} J	4.24×10^{-5} J

Supplementary Table 6 Energy consumption of the 2k model. The detailed energy consumption of the 2k model was compared between digital approaches and MDPE. For MDPE, the average conductance state is 5.161 nS in 2k model, and we assume that the average input read voltage is 3 V. Then, we calculated the energy consumption of the f-MPDE per multiplication (5 μm × 5 μm) as $E = V^2 \times G \times t = 1.40 \times 10^{-12}$ J, where t is the read latency (t = 30 μs). While the read latency could be reduced under a scaled-down device for 300 ns, as described in Supplementary Fig. 15, the energy consumption per multiplication of 100 nm × 100 nm MDPE was estimated to be 4.19×10^{-15} J. The feature size of the scaled-down device was estimated by using a general E-beam lithography process, which is sufficient for fabricating a 100-nm array. The energy consumption of peripheral circuits was estimated based on 28-nm CMOS technology including I/O register[10], ADC[11], MUX[12], MaxPool[13]. Thus, the total energy consumption for inference was 1.20×10^{-7} J and 2.64×10^{-9} J, respectively. Furthermore, for the digital approaches benchmarked by Su-in Yi et al.[13] and Da Li et al.[14], a CPU was benchmarked based on 22-nm CMOS technology using the Intel Xeon E5-2650 v2, and a GPU was based on 28-nm CMOS technology using the NVIDIA K20. The benchmark CPU and GPU consumed 2,069 pJ and 503 pJ, respectively, per multiplication. The total energy consumption for an inference was 1.74×10^{-4} J and 4.24×10^{-5} J, respectively.

Revised References in Supplementary Information:

10. Yao, P., et al. Fully hardware-implemented memristor convolutional neural network. *Nature* 577, 641-646 (2020).

11. Cao, T. V., Aunet, S., & Ytterdal, T. A 9-bit 50MS/s asynchronous SAR ADC in 28-nm CMOS. *NORCHIP 2012 IEEE*, 1 (2012).

Also, we revised manuscript as follows.

Revised Manuscript (page 11):

However, the f-MDPE with a 5-μm-line width required only 120 nJ, far below that of conventional computing approaches. This energy consumption can be further reduced by scaling the array with a 100-nm-line width[54,55]. We anticipate that the energy consumption can reach 2.64 nJ, outperforming GPU-based inference by more than four orders of magnitude (the detailed energy consumption and latency are described in Supplementary Figs. 14-16 and Tables 5-6).

Comment #4-1: *"The demonstrated network models (32x5) and the ECG dataset are quite simple. To more convincingly demonstrate the effectiveness of the proposed method, more complex simulated or experimental network models and datasets, such as MNIST or CIFAR10, are suggested."*

[Author Response]

We thank you for providing the opportunity to confirm the general applicability of our hardware-aware approach. We demonstrated classification using a 32×5 single-layer perceptron (SLP) on the MIT-BIH ECG dataset, implemented with a memristor crossbar array. Unlike image datasets such as MNIST and CIFAR-10, ECG data captures temporal patterns, making it inherently different. While our 32x5 SLP showed the feasibility of the approach, it had limited performance (F1-macro: 0.774, F1-micro: 0.815), largely due to its inability to effectively handle temporal patterns.

To address this limitation, we simulated a 1D convolutional neural network (1DCNN) model based on f-MDPE, which significantly improved performance, achieving an F1-micro score of 0.99 and an F1-macro score of 0.94. However, we observed performance saturation as the model complexity increased, as illustrated in Figure 6e.

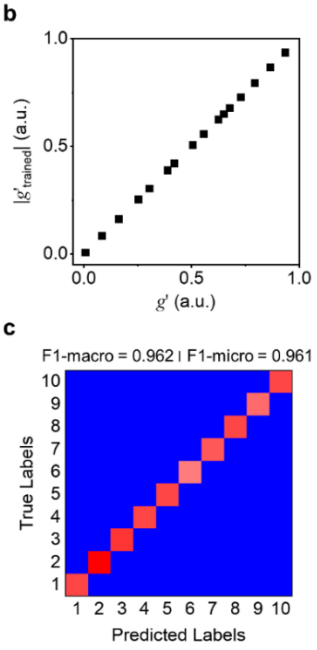
In response to the reviewer's comments on the necessity to validate our hardware-aware training approach across different datasets, we applied the same methodology to the MNIST dataset. Given that MNIST consists of two-dimensional images, we utilized a 2D convolutional neural network (2D CNN) for this purpose. We conducted Schottky conduction-based variable transformations (V' and g') for hardware-aware training, using 16 quantized states, consistent with the approach employed in the 1D CNN.

We have added the following Supplementary Figure in Supplementary Information.

Revised Supplementary Information (Page 19):

a

Input data	data size	28 x 28
Conv. 1	channels_in	1
	channels_out	4
	kenel size	3 x 3
	stride	1
	padding	1
MaxPool. 1	kernel size	2 x 2
	stride	2
Conv. 2	channels_in	4
	channels_out	16
	kenel size	3 x 3
	stride	1
	padding	1
MaxPool. 2	kernel size	2 x 2
	stride	2
FC	in	784
	out	10



Supplementary Figure 13 a Hardware-aware trained CNN for MNIST classification information. **b** Alignment between trained weights ($g'_{trained}$) and quantized weights (g'). **c** Confusion matrix of trained MNIST classifier.

In the simulation results illustrated in Revised Supplementary Figure 13, our hardware-aware trained 2D CNN using the f-MDPE approach demonstrated strong alignment between the 16 quantized weights (g') from the charge trap memristor device and the trained weights ($g'_{trained}$), as shown in Revised Supplementary Figure 13b. The confusion matrix in Revised Supplementary Figure 13c reveals that the f-MDPE 2D CNN model achieved high numerical accuracy across all digits in the MNIST dataset, with an F1-macro score of 0.962 and an F1-micro score of 0.961 from 10,000 data points.

These results indicate that our hardware-aware training approach is not only applicable to ECG datasets but can also be effectively generalized to more complex datasets with a greater number of labels. We also modified the manuscript as follows.

Revised Manuscript (Page 11):

The hardware-aware training method is also applicable to 2D CNNs, indicating its broad applicability across various applications (See Supplementary Fig 13 for more details of hardware-aware training models).

Comment #4-2: *"Furthermore, it is only demonstrated that the proposed real-time ECG diagnostic system can detect normal signals "N" which are observed most frequently and it is unclear whether it can be used for the detection of other signals, i.e., "S", "V", "Q" and "F"."*

[Author Response]

Thank you for your insightful comment on the detection capabilities of our system. In real-time demonstration, we only observed signal N, but through simulation, we were able to distinguish all the signals, which can be found in Figure S12 in the SI.

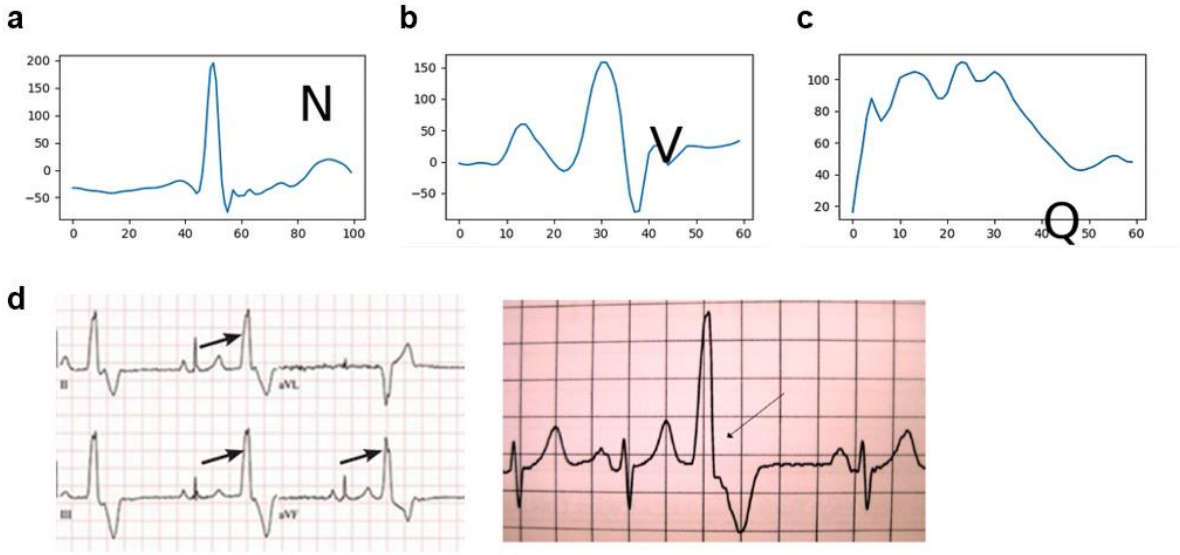


Figure R2 a N pulse, **b** V pulse, and **c** Q pulse detection & classification by the real-time ECG classification system. **d** V pulse (premature ventricular contraction) references. Arrows indicate V pulse.

In real-time ECG classification demonstration, since the subject was normal, mostly normal pulses were generated. Please note that this study is not a clinical trial involving patients, which would require highly complex procedures. However, we can deliberately destabilize the sensor system (by intentionally touching the contact point of the sensor) and generate abnormal signals on purpose. Figure R2 shows examples of such intentional signal manipulation, where the signals are recognized as V (b) and Q (c), compared to the normal signal (a). The V pulse is known to occur in cases of premature ventricular contraction (PVC). Figure R2 (d) and (e) show the actual V pulse references from [Akdemir *et al.*, *Cleveland Clinic journal of medicine*, 2016, 83, 524-530] (left) and [Hughey, Cardiac Rhythm Interpretation, Brookside Associates,

2023] (right). The Q pulse corresponds to an unknown type of pulse.

Comment #5: *"It is better to provide reliability study of proposed real-time ECG diagnostic system constructed with the f-MDPE array, because the medical scenario has a very high standard of reliability."*

[Author Response]

We agree with the reviewer's concerns about the reliability study. For real-time ECG diagnosis, the process involves the steps outlined in the flow chart below.

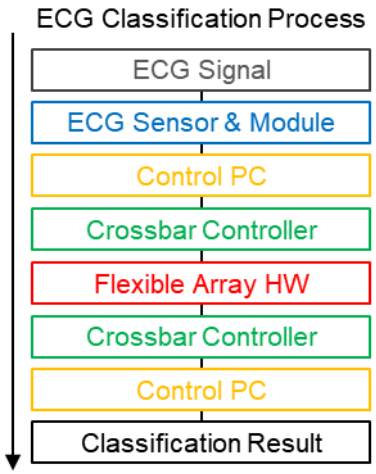
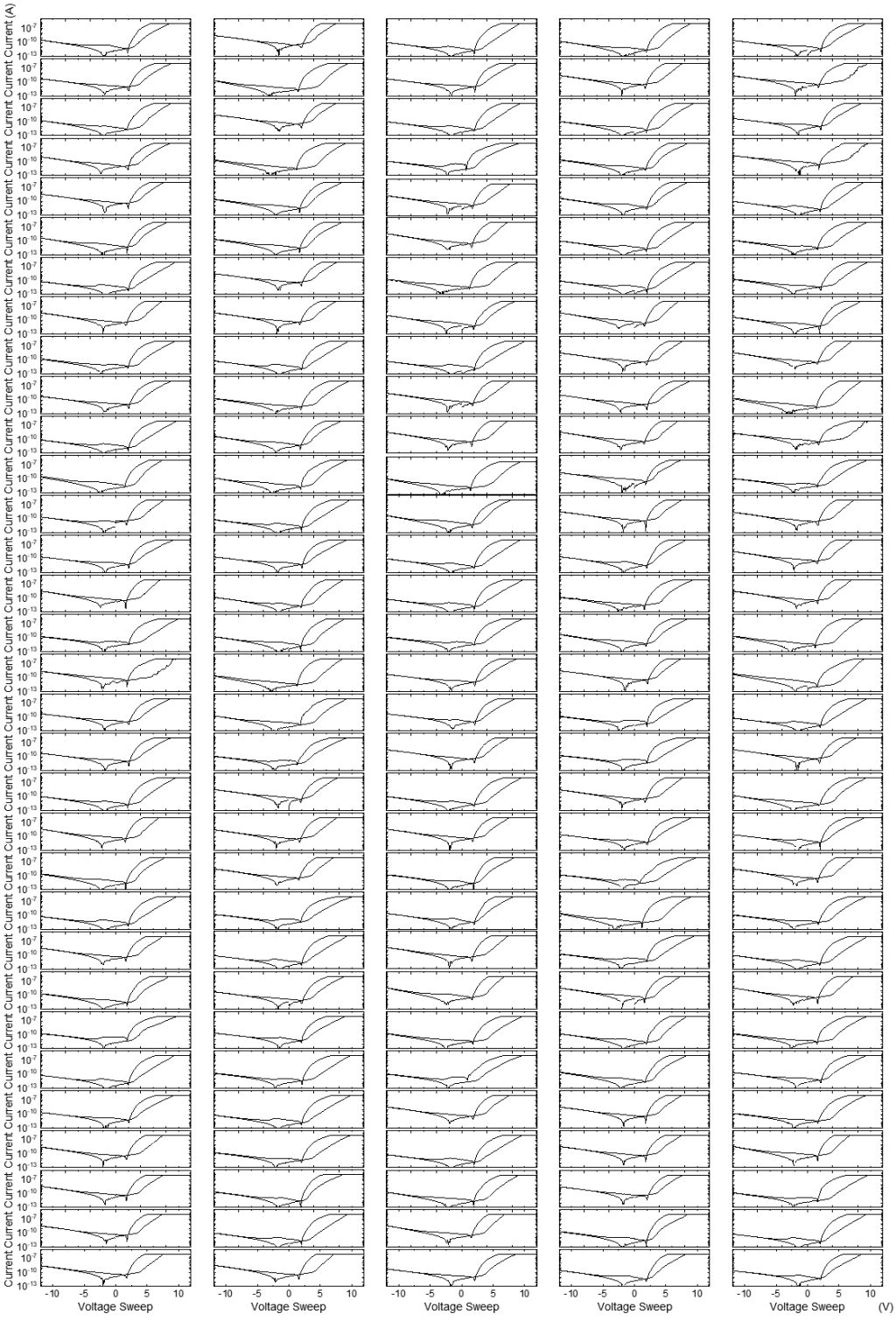


Figure R3 Flow chart for real-time ECG diagnosis system

While the reliability of all stages is important, we focused on proving the reliability of the flexible array hardware, as commercial products were used for the ECG sensor, control PC, and crossbar controller. Specifically, we further examined the device-to-device variation of the flexible array hardware used to drive a 32×5 SLP. Therefore, we additionally measured a total of 160 *I-V* characteristics. These measurements were obtained through DC sweeps, and a compliance current of 1 μA was used. In the measurement results, all 160 devices operated, and we were able to obtain conductance states within the desired range, from 0.1 nS to 15 nS at 3 V. Through this high yield and uniformity results, we would like to emphasize that the reliability of the f-MDPE, which is the main claim of this study.

We revised the device-to-device variation results in Figure S3 as follows.

Revised Supplementary Information (Page 5):



Supplementary Figure 3 The *I-V* curves of 160 devices in the f-DPE array. The voltage was first swept from 0 V to +12 V and back to 0 V, followed by a sweep from 0 V to −12 V and back to 0 V. A compliance current of 1 μA was used during the DC *I-V* measurements.

Also, we elaborated on it in the revised manuscript as follows.

Revised Manuscript (Page 4):

Fig. 2d shows *I-V* curves of 160 cells (32×5) in the array at an I_{CC} of 10^{-6} A, confirming a 100% yield and high array-level uniformity (see Supplementary Fig. 3 in the SI for the 160 raw *I-V* curves).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The author has addressed most of my issues, but there are still a few below.

1. I'm very curious as to the root reason for the different on/off ratios under DC I-V sweeps and pulses, which has occurred in many works.

The amount of set and reset switching in memristors is determined by the **amplitude and period** of the applied voltage. In this regard, even when the applied voltage amplitude is the same, there is a **difference in the period** between a DC I-V sweep and pulse mode, which is the primary cause of the on/off ratio difference.

In a DC sweep mode, the voltage is gradually increased during sweeping and also each voltage is applied for a sufficient time to read the current. As a result, the device can reach a fully switched conductance level at specific voltage points, demonstrating the maximum possible on/off ratio. However, in the case of pulse mode, the pulse period is typically shorter (5 ms in our case for endurance measurements), which is often insufficient to reach the fully switching conductance level. Therefore, the on/off ratio observed during pulse programming typically shows a narrower window compared to that observed in a DC sweep.

2. For Schottky and P-F transport, the only difference in the equations for the functions of I and V is the temperature T, so variable temperature measurements may be required to verify exactly which relationship is met.

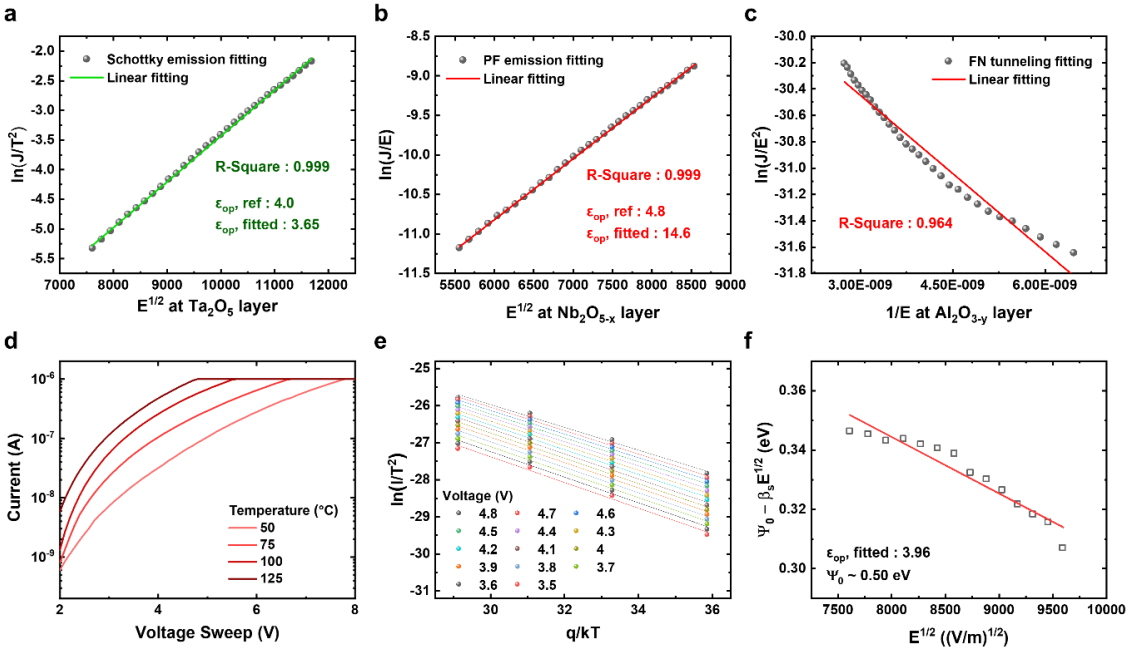
[Author Response]

We appreciate the reviewer's insightful comments on the analysis of the conduction mechanism. As pointed out, we have supplemented the modeling of the conduction mechanism by incorporating additional supporting data on temperature dependence. In addition to the fitting results under RT conditions, we confirmed the temperature-dependent trends of LRS, verifying that the mechanism is thermally dependent.

Furthermore, through an Arrhenius-form plot based on temperature variations, we extracted the activation energy. By referencing the methodology outlined in [J. H. Yoon *et al.*, *Adv. Funct. Mater* **24**, 5086 (2014)], we determined the zero-field Schottky barrier height and Schottky coefficient. Using these, we cross-validated the optical dielectric constant value, which was consistent with the reference for the Ta_2O_5 layer, providing additional evidence supporting Schottky emission at the Ta_2O_5 interface.

In accordance with the reviewer's suggestion, we have revised the conduction mechanism fitting results with additional temperature-dependent analysis in Supplementary Fig. 6 as follows.

Revised Supplementary Information (Page 8):



Supplementary Figure 6 Fitting analysis of possible conduction mechanisms in the f-MDPE. Fitting results of **a** Schottky emission ($\ln(J/T^2)$ vs. $E^{1/2}$) at the Ta_2O_5 layer, **b** Poole-Frenkel emission ($\ln(J/E)$ vs. $E^{1/2}$) at the NbO_x layer, and **c** Fowler-Nordheim tunneling ($\ln(J/E^2)$ vs. $1/E$) at the Al_2O_3 layer. For the fitting, the LRS *I-V* curve at room temperature was fitted in the voltage range of +3.5 V to +6.5 V. The local electric field in each layer was calculated using the thickness and dielectric constant of the oxide layers. The thicknesses were 6 nm for Ta_2O_5, 25 nm for NbO_x, and 7 nm for Al_2O_3. The dielectric constants were 24 for Ta_2O_5, 45 for NbO_x, and 9 for Al_2O_3.[5-7] The linear fitting results, along with the reasonable optical dielectric constant (ε_{op}) value, support that the primary conduction mechanism in this device is Schottky conduction at the Ta_2O_5 interface. In case of the Poole-Frenkel fitting, ε_{op} was significantly higher than the reference value. In case of the FN tunneling, the linear fit was not satisfactory. **d** LRS *I-V* curves measured at various temperatures ranging from 50°C to 125°C. **e** Schottky-type plot of $\ln(I/T^2)$ vs. $1/T$ from the current results in **d**. **f** Schottky barrier height at a given electric field ($\Psi_0 - \beta_s E^{\frac{1}{2}}$, where Ψ_0 and β_s represent the zero-field Schottky barrier height and Schottky coefficient, respectively). The extracted dielectric constant of 3.96 closely matched the fitted result in **a**, while the Ψ_0 value was approximately 0.50 eV.

3. For Comment #6, I suggest that the author should add revision length to allow readers to understand it quickly.

We sincerely appreciate the reviewer's comments. In response, we have provided a more detailed explanation of the relevant content. The manuscript and supplementary information have been elaborated as follows.

Revised Manuscript (Page 7):

Here, we put a compensation constant k that scales the value of each exponential term accordingly, addressing potential issues with scale consistency and ensuring that the neural network operates within a stable and efficient regime (see Supplementary Table 1 for scale between g' and V' with constant k).

Revised Supplementary Information (Page 9):

					$k = 0$	$k = 7$
V'	**Input voltage (V)**		b	c	$V' = \exp(-k + b\sqrt{V - c})$	
	3.5		7.66	1.5	5.07E+4	4.62E+1
	3.4				3.85E+4	3.51E+1
	⋮				⋮	⋮
	2.2				6.07E+2	5.54E-1
	2.1				3.77E+2	3.44E-1
	2.0				2.25E+2	2.05E-1
g'	**Analog states**	g_{ref} **(nS)**	a		$g' = \exp(a + k)$ $(\times 10^9)$	
	16	15	-27.79064		8.52E-4	9.35E-1
	15	14	-27.86691		7.90E-4	8.66E-1
	⋮		⋮			
	3	2	-29.09895		2.30E-4	2.53E-1
	2	1	-30.1967		7.69E-5	8.43E-2
	1	0.1	-32.72548		6.13E-6	6.72E-3

Supplementary Table 1 Comparison of g' and V' for compensation parameter k = 0 and k = 7 during hardware-aware training. The scale between g' and V' is modified by introducing the k. k = 0 can lead to a significant scale mismatch between the g' and V'. Such a mismatch can cause several problems: significant differences in scale can result in unstable gradients during backpropagation, making it difficult to optimize network parameters effectively.

4. For Comment #8, is the circuit of the controller integrated on the PCB board or is it realized through simulation? Besides, the authors should clarify exactly which are obtained by hardware testing and which are obtained by software simulation in the whole application process.

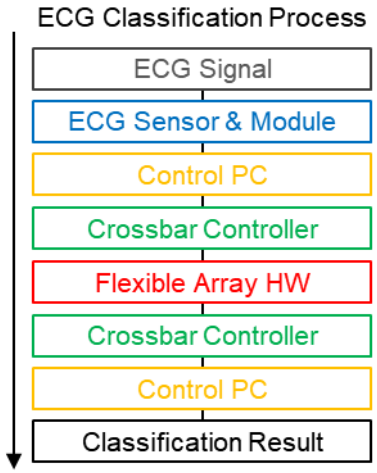
We appreciate the reviewer's concerns regarding the clarity of our explanation. To address this, it is important to first distinguish between the training and inference processes. We utilized *ex-situ* training, which was conducted through software simulation. As a result, the transformation to V' occurred not in the actual circuit but within the computing simulation. The weight matrix obtained through this training process is then transferred to the flexible array hardware, making it ready for inference.

For real-time ECG diagnosis, the procedure follows the steps outlined in the flowchart below. The ECG data collected by the sensor is first gathered on a control PC and preprocessed to align with the operating voltage range of the flexible array hardware. Subsequently, the processed voltage signals are input into the flexible array hardware via the crossbar controller (ArcOne) to obtain current outputs. It is worth noting that no substitution with V' is required during inference. Finally, the current output undergoes classification using a winner-takes-all method, and the result is displayed on the control PC. The manuscript and supplementary information have been elaborated as follows.

Revised Manuscript (Pages 10-11):

Fig. 6a shows a photograph of a laboratory-scale real-time ECG diagnosis system comprising a control PC, an f-MDPE controller, an f-MDPE, and an ECG sensor, where the ECG sensor and the f-MDPE are attached to a human wrist (see Supplementary Fig. 12 for the detailed electrical connection scheme between the f-MDPE and the controller). The f-MDPE was programmed to perform only inference operations based on the weight matrix trained *ex-situ* through software. The control PC receives the ECG signals from the ECG sensor, preprocesses them to the input voltage signal for f-MDPE inference, and sends instructions to the f-MDPE controller. The f-MDPE controller sends the input voltages to the f-MDPE, receives the output currents, and returns them to the PC. Then, the PC determines the category of the ECG signals (see Supplementary Fig. 13 for the flowchart of the real-time ECG diagnosis system).

Revised Supplementary Information (Page 18):



Supplementary Figure 13 A flow chart for real-time ECG diagnosis system as an inference.

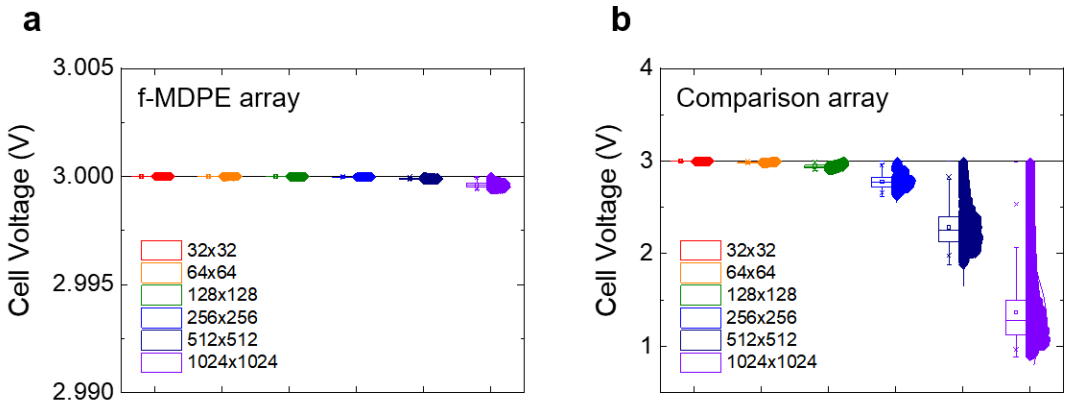
5. What do the authors consider to be the basic factors limiting the scale-up of arrays at both the experimental preparation and hardware testing levels?

First, there is a possibility of issues related to device variation and yield. While we conducted experimental demonstrations using a small 32×10 array, which allowed us to avoid high variation and yield issues, larger arrays are likely to encounter problems with non-uniformity or defective cells. These challenges can be addressed by considering additional ECC design [D. Niu *et al.*, *17th Asia and South Pacific Design Automation Conference* 79, (2012)] or by employing training methods that account for defective cells, such as in-situ training [C. Li *et al.*, *Nat. Commun.* **9**, 2385, (2018)].

Second, when using arrays with tens or hundreds of thousands of cells exceeding 4k, current summation may no longer be feasible within a single bit line. Instead, summation would need to occur within array blocks and then be combined through peripheral circuits. In such scenarios, additional peripheral circuitry would be required, resulting in increased system complexity.

Additionally, there may be an issue with line resistance. The resistance present between each cell of the array can affect the intended vector-matrix multiplication (VMM). However, compared to other devices such as valence change memristors, the charge trap memristor (CTM)—which operates at a higher resistance level (i.e., lower operating current)—may be less affected by issues related to line resistance. At the below, revised Supplementary Fig. 3a presents the read voltage distribution affected by the voltage drop at the line resistance (10 Ω between two adjacent cells), analyzed for different array sizes using random conductance range of 0.1 nS and 15 nS, consistent with the CTM. For comparison, while keeping all other conditions the same, we randomly set the cell conductance between 10 μS and 150 μS and examined the corresponding read voltage distribution (revised Supplementary Fig. 3b). The results indicate that the CTM operates in a high-resistance state, demonstrating better reliability in VMM operations by mitigating the effects of line resistance-induced voltage drops in the scaled-up array. The supplementary information has been elaborated as follows.

Revised Supplementary Information (Page 5):



Supplementary Figure 3 Cell voltages considering line resistance with various array size. **a** Cell voltage distribution of a f-MDPE array with V_{read} = 3 V. **b** Cell voltage distribution of a common memristor array that the conductance between 10 µS and 150 µS with V_{read} = 3 V.

6. For Comment #13, "On the other hand, the latency is more related to the wire capacitance than to the device itself.", this expression is obviously problematic. When the device resistance is large, the line capacitance has rather less effect on the latency than the device itself.

I appreciate the reviewer's comment regarding the use of confusing expressions. In the previous revision letter, the phrase "On the other hand, the latency is more related to the wire capacitance than to the device itself. Therefore, in energy calculations, the reduction in energy caused by wire capacitance can be directly improved through scaling." should be corrected to use "device capacitance" instead of "wire capacitance." The revised statement should read: "On the other hand, the latency is more related to the device capacitance than to the resistance. Therefore, in energy calculations, the reduction in energy caused by device capacitance can be directly improved through scaling."

7. The author has omitted Comment #14.

We are afraid that Comment #14 was not included in the first revision letter.

8. The endurance tests in Fig. 2h lack of more data points throughout the testing period and more devices comparison which can provide a clearer understanding of the device's endurance behavior. Adding more intermediate data points would help to better visualize and confirm the consistency and reliability of the switching performance over the entire testing period. For the methodology, I recommend referring to "https://doi.org/10.1063/5.0227603”.

Thank you for the reviewer's valuable comments. To better visualize the reliability of the switching performance, we have referred to the paper suggested by the reviewer and revised the endurance plot in Fig.2h with a linear time scale.

Revised Figure (Page 22):

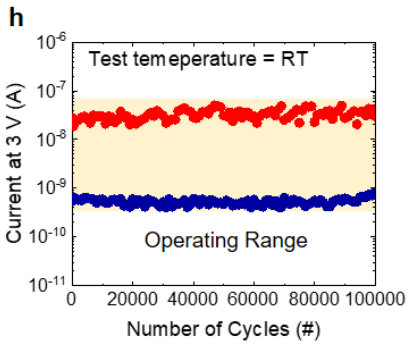


Fig. 2h Endurance performance for up to 10^5 cycles at room temperature

Reviewer #3 (Remarks to the Author):

*Review of the manuscript "**Flexible Self-rectifying Synapse Array for Energy-efficient Edge Multiplication in Electrocardiogram Diagnosis**" authored by Younghyun Lee, Hakseung Rhee, Geunyoung Kim, Woon Hyung Cheong, Do Hoon Kim, Hanchan Song, Sooyeon Narie Kay, Jongwon Lee, and Kyung Min Kim, submitted to Nature Communications. [Manuscript Number: NCOMMS-24-35806A].*

This reviewer was asked to review the work, previous reviewers' comments, and authors' responses. The authors have adequately addressed the previous reviewers' comments through additional experiments and analyses, improving the quality of the manuscript. However, this reviewer found additional concerns which need to be addressed before publication:

*1. In **Figure 6f**, the energy consumption for ECG diagnosis is compared, but only the neural network computation is mentioned. Two questions are:*

• Is the diagnosis of ECG-based arrhythmias typically performed using neural network models such as MLPs or CNNs?

Typical ECG diagnosis primarily relies on the experience and judgment of medical professionals. While it is difficult to quantify the energy consumption of this conventional approach, manually reviewing large volumes of ECG data imposes a significant burden in terms of both accuracy and efficiency. As a result, the application of AI technology for ECG data analysis has recently emerged as a promising approach.

We have revised the corresponding parts to emphasize the purpose of AI-driven ECG diagnosis as follows.

Revised Manuscript (Page 3):

Typical ECG diagnosis primarily relies on the experience and judgment of medical professionals. However, ECG diagnosis requires long-term monitoring of intermittently occurring abnormal signals, imposing a significant burden in terms of both accuracy and efficiency[29]. Therefore, recent efforts have been made to address this issue with the help of AI. However, AI-based analysis often relies on data collected over short periods or requires remote

processing, posing certain limitations. For a more comprehensive diagnosis, ECG signal collection and analysis must be carried out continuously and in real time, making this an optimal application for edge AI.

• Are the neural network computations dominant enough to represent the computational burden of the entire ECG diagnosis process?

We sincerely appreciate the reviewer's valuable comments. Our answer is yes; Neural network computations are dominant to a significant energy consumption burden in the overall ECG diagnosis process.

The overall process can be summarized as follows: 1) ECG sensing, 2) Preprocessing, 3) Computing (using neural networks), and optionally, 4) Wireless transmission (e.g., via Bluetooth). In ECG system-on-chip devices, acquiring a single ECG pulse (assuming a 1-second acquisition) consumes 17.4 μJ [M. Khayatzadeh *et al.*, *IEEE Transactions on Biomedical Circuits and Systems* **7**, 583 (2013)], 6.1 μJ [M. Altini *et al.*, *Proceedings of the 2nd Conference on Wireless Health* **15**, 1 (2011)], and 8.5 μJ [A. Jain *et al.*, *Measurement: Sensors* **24**, 100523 (2022)], respectively. For preprocessing, approximately tens of nJ of energy are consumed [H. Huang *et al.*, *ACM Transactions on Cyber-Physical Systems* **3**, 1 (2019)].

When performing inference for a single ECG pulse using a processor such as a GPU, 42 μJ of energy is consumed (as shown in Supplementary Table 6), making this the most significant energy burden in the overall ECG diagnostic process. However, by utilizing f-MDPE, the energy required for neural network computing can be reduced to 120 nJ, significantly alleviating this energy burden.

Regarding wireless transmission energy, real-time Bluetooth transmission consumes 22.7 mW [E. Nemati *et al.*, *IEEE Communications Magazine* **50**, 36 (2012)]. By implementing edge-based diagnostics and transmitting data only when an arrhythmia is detected, the energy consumed for wireless transmission can be dramatically reduced.

We elaborated discussion in manuscript as below.

Revised Manuscript (Pages 12-13):

In summary, we reported energy-efficient neuromorphic applications utilizing the f-MDPE, characterized by its highly nonlinear and self-rectifying characteristics with a low-current operation, suitable for energy-efficient and reliable vector-matrix multiplication

operations. Also, the device exhibited mechanical durability up to a 5-mm bending curvature, making it suitable for wearable applications. We successfully demonstrated the use of f-MDPE for *in-situ* ECG classification by applying a hardware-aware training method that overcomes hardware characteristics such as non-Ohmic conduction and weight quantization. Furthermore, we simulated the proposed method using scaled-up 1D CNNs, confirming its feasibility for real-time ECG diagnosis with high diagnostic accuracy.

In edge computing, utilizing edge hardware plays a dominant role in reducing overall energy consumption. For example, in real-time ECG diagnosis, the overall process can be overall process can be summarized as follows: 1) ECG signal acquisition, 2) Preprocessing, 3) Inferencing, and optionally, 4) Wireless transmission (e.g., via Bluetooth). A single ECG pulse acquisition (assuming a 1-second acquisition) consumes around 6.1 μJ–17.4 μJ[63-65]. The preprocessing stage requires negligible energy, around tens of nJ[66]. Inferencing using a GPU requires 42 μJ, but this energy consumption can be reduced to 120 nJ by using f-MDPE, significantly alleviating this energy burden. Furthermore, edge computing processes data locally, minimizing external signal transmission. This drastically reduces the high energy consumption required for wireless transmission (real-time Bluetooth transmission typically consumes 22.7 mW[67]). As such, the f-MDPE aligns well with future analog signal processing at edge devices, providing energy-efficient artificial intelligence solutions.

Further, we clarify the manuscript about the energy consumption comparison as below.

Revised Manuscript (Page 12):

Fig. 6f shows the estimated energy consumption for ECG inference on various computing platforms when the 2k model is used. Note that this comparison focuses on the energy consumption of running neural networks, excluding other sources of energy consumption, such as ECG sensors.

2. ECG sensors and arrhythmia diagnosis systems have been commercialized and widely used for many years. It would be valuable to discuss the pros and cons of this hardware compared to existing devices for arrhythmia diagnosis.

We sincerely thank the reviewer for their insightful comments. In this revision, We have revised the conclusion section as follows to clearly convey the advantages and limitations of our approach.

Revised Manuscript (Page 13):

While the f-MDPE shows promise for real-time and long-term signal monitoring due to its energy efficiency and portability, several challenges remain to be addressed for the further advancements. First, for its use as a fully neuromorphic processor, its functionality should be expanded beyond inference to training. To achieve this, a control unit must be integrated, which would require advancements in wearable electronics. Even in this scenario, the low programming energy of f-MDPE would remain a significant advantage. Second, improving the higher operating voltage and slower switching speed of f-MDPE through the development of new materials and process optimizations could further enhance its usability. If these aspects can be enhanced in the future, it would significantly broaden the applicability of f-MDPE across a wider range of applications.

3. The application of AI in medical diagnostics, including arrhythmia diagnosis, generally requires significant reliability and explainability. Although ECG-based arrhythmia diagnosis is relatively straightforward and less constrained by these requirements, reliability remains critical. Discussing the potential reliability issues associated with the memristive hardware in ECG diagnosis would enhance the manuscript.

• While the manuscript states that the hardware is relatively free from sneak current issues, it does not explain voltage drops caused by wire resistance. This issue will likely occur even in the current hardware size (32 × 10) and will become more significant as the array size increases. SPICE-based crossbar array simulations could confirm these concerns. Discussing these challenges and suggesting potential solutions would be beneficial.

• Furthermore, in ***Figure S2****, the manuscript addresses sneak current under the assumption that the crossbar array is used as memory. However, since this manuscript presents the crossbar array as a VMM accelerator, analyzing the impact of leakage current and line resistance in the context of VMM operations is more appropriate.*

We sincerely appreciate the reviewer's valuable comments. As the reviewer pointed out, considering the hardware aspects of VMM accelerator operations is a crucial factor. While the entire bit line is grounded during VMM inference, eliminating the influence of leakage current, the voltage drop caused by wire resistance and analog conductance states during VMM should be taken into account. To analyze this, we conducted simulations using Python. The conductance of each crossbar array cell was randomly assigned a value between 1 nS and 15 nS, while the parasitic line resistance between cells was assumed to be 10 Ohms [S. N. Truong, Materials 12, 4097 (2019)]. A read voltage of 3 V was applied, and the voltage drop was calculated as the array size increased from 32×32 to 1024×1024.

At the below, revised Supplementary Fig. 3a presents the read voltage distribution affected by the voltage drop, analyzed for different array sizes. Additionally, we conducted a similar analysis using low-resistance memristors such as VCM. Keeping all other conditions the same, we randomly set the cell conductance between 10 μS and 150 μS and examined the corresponding read voltage distribution (revised Supplementary Fig. 3b). The results indicate that, unlike VCM, CTM operates in a high-resistance state, demonstrating better reliability in VMM operations by mitigating the effects of line resistance-induced voltage drops. The manuscript and supplementary information have been elaborated as follows.

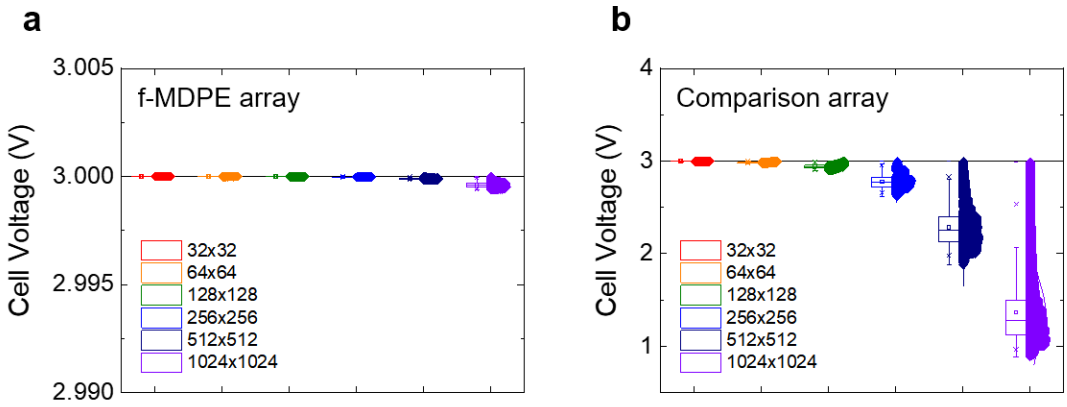
Revised Manuscript (Pages 4-5):

Owing to its self-rectifying characteristics, the device may operate accurately even in a 1k×1k array without sneak path disturbance[34,35] (see Supplementary Fig. 2 in the SI for a discussion on the available array size and read margin of the f-MDPE array). To further evaluate the impact of voltage drop caused by wire resistance in a scaled-up device, we conducted cell voltage distribution simulations (see Supplementary Fig. 3). In the simulation, the conductance of each cell was randomly assigned a value between 1 nS and 15 nS, while the wire resistance between cells was assumed to be 10 Ω[36]. A read voltage of 3 V was applied, and the voltage drop was analyzed across array sizes ranging from 32×32 to 1024×1024. The results indicate that the wire resistance has little impact on f-MDPE operation, suggesting its highly reliable performance even in large-scale arrays.

Revised Manuscript reference (Page 17):

36. Lee, Y. K., et al. Matrix mapping on crossbar memory arrays with resistive interconnects and its use in in-memory compression of biosignals. *Micromachines* **10**, 306 (2019).

Revised Supplementary Information (Page 5):



Supplementary Figure 3 Cell voltages considering line resistance with various array size. **a** Cell voltage distribution of a f-MDPE array with V_{read} = 3 V. **b** Cell voltage distribution of a common memristor array that the conductance between 10 μS and 150 μS with V_{read} = 3 V.

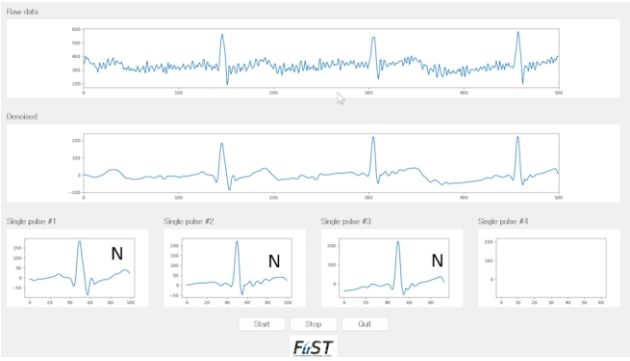
4. In real-time ECG data acquisition, the intervals between beats are not perfectly uniform. Over time, the pulse peaks may shift from the center of the 32-frame window. While ***Table S3*** *mentions preprocessing based on location information from the MIT-BIH dataset,* ***Figure 6a*** *shows real-time diagnosis where location information is unavailable. How can single pulses be extracted for real-time diagnosis, as depicted in* ***Figure 6d****? An additional explanation of the preprocessing method for real-time scenarios would improve the manuscript's clarity.*

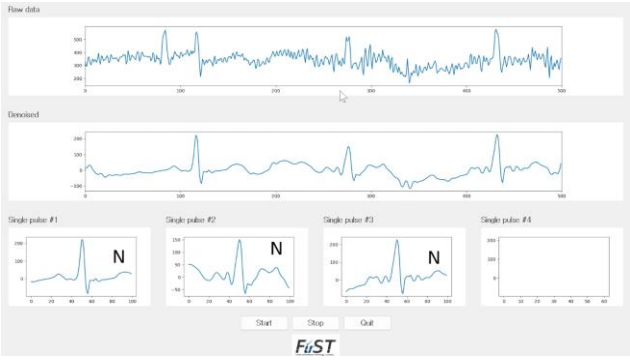
Thank you for your detailed question regarding our work.

In our real-time ECG signal classification process, we also preprocessed the collected raw ECG signals as follows. We temporarily store ECG patterns of approximately every 10 seconds from continuously collected data. From these patterns, a preprocessing step extracts single ECG datasets, typically providing three to four ECG datasets. This preprocessing is performed using software (Python) and includes denoising, resizing, and peak detection.

During peak detection, time information for the start and end points of the ECG signal is obtained. This time information is continuously calibrated throughout data collection, ensuring that even if an abnormal waveform without a distinct peak appears, the normal waveform's time information can be used to extract the abnormal signal.

This process is well shown in the supplementary video. Please see the following captured images for your reference.





As the reviewer mentioned, we have explicitly stated in the manuscript that we preprocessed the collected ECG signals.

Revised Manuscript (Page 11):

We attempted two real-time tests in this system: (1) using the MIT-BIH ECG dataset for reference (w/o ECG sensor part) and (2) using the preprocessed ECG data collected in real-time from ECG sensors. The preprocessing of the real-time ECG data was performed via software (Python) and included signal denoising, resizing, and peak detection to ensure accurate dataset generation (see Supplementary Video for the real-time ECG diagnosis demonstration). … Fig. 6c shows the preprocessed real-time ECG signals collected from the ECG sensor (left panel) and the collected output currents (right panel).

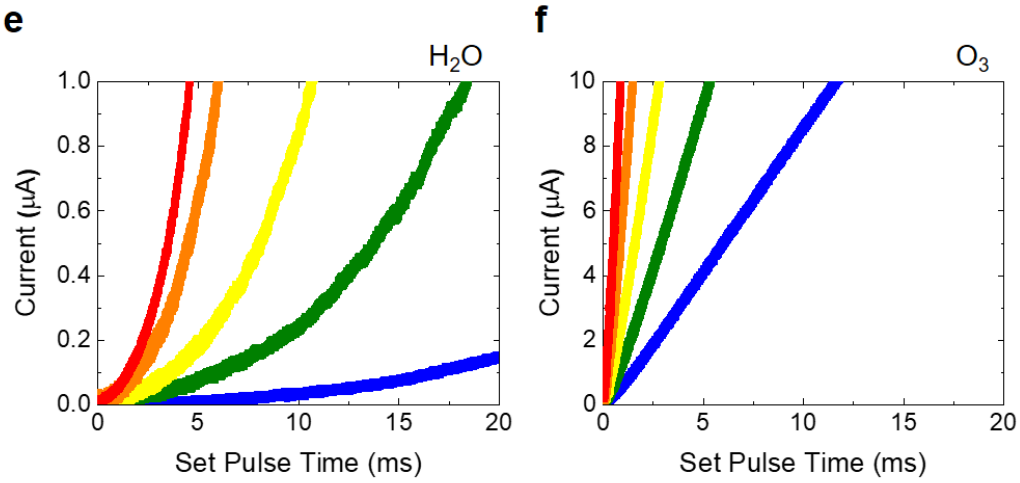
*5. This is a minor comment. In **Figure S1**, the I-V curves indicate that Pt/H2O-Al2O3/Ti device exhibits more significant cycle-to-cycle variation compared to the Pt/O3-Al2O3/Ti device. In contrast, the pulse data suggests the opposite, with a larger variation observed in the Pt/O3-Al2O3/Ti device. What could be the reasons for this discrepancy?*

In the I-V curves, we collected multiple curves, so as the reviewer noted, Fig. 1b and Fig. 1c capture cycle-to-cycle variations.

In contrast, for the pulse sweep in Fig. 1e and Fig. 1f, we collected a single curve for each Vset condition. The scattered data of Fig. 1f is due to the noise in the reading process. This appears more pronounced because of the difference in the x-axis scale range, as Fig. 1f was more magnified compared to Fig. 1e.

In this revision, we have adjusted the scales of Fig. 1e and Fig. 1f to ensure consistency and prevent any confusion, as shown below

Revised Supplementary Information (Page 3):



REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

I think the authors have made proper revision with additional data and discussion to address previous concerns.

This study could provide a impressive demonstration of self-rectifying memristors with application in edge medical computing scenario.

I suggest potential publication as it is.

We sincerely thank for reviewer's thoughtful feedback throughout the review process.

Reviewer #3 (Remarks to the Author):

The manuscript is well modified according to the comments. This reviewer suggests the acceptance of the work in its current form.

We sincerely thank for reviewer's thoughtful feedback throughout the review process.

Review of the manuscript **"Flexible Self-rectifying Synapse Array for Energy-efficient Edge Multiplication in Electrocardiogram Diagnosis"** authored by Younghyun Lee, Hakseung Rhee, Geunyoung Kim, Woon Hyung Cheong, Do Hoon Kim, Hanchan Song, Sooyeon Narie Kay, Jongwon Lee, and Kyung Min Kim, submitted to *Nature Communications*. [Manuscript Number: NCOMMS-24-35806A].

This reviewer was asked to review the work, previous reviewers' comments, and authors' responses. The authors have adequately addressed the previous reviewers' comments through additional experiments and analyses, improving the quality of the manuscript. However, this reviewer found additional concerns which need to be addressed before publication:

1. In **Figure 6f**, the energy consumption for ECG diagnosis is compared, but only the neural network computation is mentioned. Two questions are:

    - Is the diagnosis of ECG-based arrhythmias typically performed using neural network models such as MLPs or CNNs?

    - Are the neural network computations dominant enough to represent the computational burden of the entire ECG diagnosis process?

2. ECG sensors and arrhythmia diagnosis systems have been commercialized and widely used for many years. It would be valuable to discuss the pros and cons of this hardware compared to existing devices for arrhythmia diagnosis.

3. The application of AI in medical diagnostics, including arrhythmia diagnosis, generally requires significant reliability and explainability. Although ECG-based arrhythmia diagnosis is relatively straightforward and less constrained by these requirements, reliability remains critical. Discussing the potential reliability issues associated with the memristive hardware in ECG diagnosis would enhance the manuscript.

    - While the manuscript states that the hardware is relatively free from sneak current issues, it does not explain voltage drops caused by wire resistance. This issue will likely occur even in the current hardware size (32 × 10) and will become more significant as the array size increases. SPICE-based crossbar array simulations could confirm these concerns. Discussing these challenges and suggesting potential solutions would be beneficial.

- Furthermore, in **Figure S2**, the manuscript addresses sneak current under the assumption that the crossbar array is used as memory. However, since this manuscript presents the crossbar array as a VMM accelerator, analyzing the impact of leakage current and line resistance in the context of VMM operations is more appropriate.

4. In real-time ECG data acquisition, the intervals between beats are not perfectly uniform. Over time, the pulse peaks may shift from the center of the 32-frame window. While **Table S3** mentions preprocessing based on location information from the MIT-BIH dataset, **Figure 6a** shows real-time diagnosis where location information is unavailable. How can single pulses be extracted for real-time diagnosis, as depicted in **Figure 6d**? An additional explanation of the preprocessing method for real-time scenarios would improve the manuscript's clarity.

5. This is a minor comment. In **Figure S1**, the I-V curves indicate that Pt/H_2O-Al_2O_3/Ti device exhibits more significant cycle-to-cycle variation compared to the Pt/O_3-Al_2O_3/Ti device. In contrast, the pulse data suggests the opposite, with a larger variation observed in the Pt/O_3-Al_2O_3/Ti device. What could be the reasons for this discrepancy?