Ph. D. DISSERTATION

Complex Data Processing with Novel Memristor Array

by

Yoon Ho Jang

February 2023

Department of Materials Science and Engineering

College of Engineering

Seoul National University

Complex Data Processing with Novel Memristor Array

Advisor: Prof. Cheol Seong Hwang

By
Yoon Ho Jang

A thesis submitted to the Graduate Faculty of Seoul National University

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

Department of Materials Science and Engineering

February 2023

Approved

By

Chairman of Advisory Committee: Sangbum Kim

Vice-chairman of Advisory Committee: Cheol Seong Hwang

Advisory Committee: Min Hyuk Park

Advisory Committee: Kyung Min Kim

Advisory Committee: Hae Jin Kim

Abstract

Recently, with the remarkable development of deep learning, various data have been accumulated. As the structure of big data becomes more diversified and complex, complex data that is difficult to process with existing hardware has emerged. Examples of complex data include sequential data and graph data. Sequential data has characteristics that the current state reflects the input history and the pattern is not constant and difficult to predict. Graph-type data is difficult to be expressed in vector form since graphical data includes the connections between entities. To process such complex data, novel data processing techniques are required.

In the first part of this study, a method for processing time-series data with a nonvolatile memristor is proposed. Recent advances in physical reservoir computing, which is a type of temporal kernel, have made it possible to perform complicated timing-related tasks using a linear classifier. However, the fixed reservoir dynamics in previous studies have limited application fields. In this study, temporal kernel computing was implemented with a physical kernel that consisted of a W/HfO_2/TiN memristor, a capacitor, and a resistor, in which the kernel dynamics could be arbitrarily controlled by changing the circuit parameters. After the capability of the temporal kernel to identify the static MNIST data was proven, the system was adopted to recognize the sequential data, ultrasound (malignancy of lesions), and electrocardiogram (arrhythmia),

that had a significantly different time constant (10^{-7} vs. 1 s). The suggested system feasibly performed the tasks by simply varying the capacitance and resistance. These functionalities demonstrate the high adaptability of the present temporal kernel compared to the previous ones.

In the second part of this study, a method for processing non-Euclidean graphs using self-rectifying memristor arrays is proposed. Many big data have interconnected and dynamic graph structures growing over time. Analyzing these graphical data requires identifying the hidden relationship between the nodes in the graphs, which has conventionally been achieved by finding the effective similarity. However, graphs are generally non-Euclidean, which does not allow finding it. In this study, the non-Euclidean graphs were mapped to a specific crossbar array (CBA) composed of the self-rectifying memristors and metal cells at the diagonal positions. When all bit lines of CBA are connected to the ground, the sneak current is suppressed, and CBA can be used to search for adjacent nodes. When a single bit line is connected to the ground, the sneak current, an intrinsic physical property of the CBA, allows for identifying the similarity function. Sneak current-based similarity function indicates the distance between nodes, the probability that unconnected nodes will be connected in the future, connectivity between communities, and cortical connections in a brain. This work demonstrates the physical calculation methods applied to various graphical problems using the CBA composed of the self-rectifying-memristor based on the HfO_2 switching layer. Moreover, such

applications suffer less from the memristors' inherent issues related to their stochastic nature.

Keywords: Resistive switching memory, ReRAM, Memory, Hafnium oxide, Self-rectifying memristor, Complex data, Kernel, Temporal kernel, Sequential data, Medical diagnosis, Crossbar-array, Sneak current, Graph algorithm, Process-in-memory

Student ID: 2018-24630

Yoon Ho Jang

Table of Contents

List of Tables

pulse width was modified without changing the conditions R_L, C, and pulse height in the condition of Figure 2-9f, an accuracy of 92.4 % was obtained in the combination of '200ns+2μs+5μs'. By combining various kernels or changing pulse conditions for the same kernel machine, the imperfections of one kernel could be compensated for by another kernel, and the accuracy could be improved.

Table 2-5: The accuracy when cycle to cycle variation, cell to cell variation, and both are considered (kernel condition of Figure 2-9f of main text was used). Each variation was calculated based on variation measurement results in Figure 2-1. Up to 1 sigma of each variation was considered, and when both cycle to cycle and cell-to-cell variations were included in the simulation, the accuracy decreased by 0.5 %.

Table 2-6: Results of MNIST recognition using two-layer FCN for the readout layer of the TK system. The table shows the number of training parameters used in each two-layer FCN and the accuracy of the TK system (nBPK = 4). When 196x38x10 FCN was used, 7,828 training parameters were used, and the TK system accuracy was 95.1 %.

Table 2-7: Results of the MNIST recognition while increasing the number of bits processed in the temporal kernel, showing that as nBPK increased, both the size of the used readout layer and the

recognition accuracy decreased.

List of Figures

conductance range (**a**) and a retention time of about 100 days at 25 °C, which is the result of extrapolation based on 60 ~ 150 °C retention data (**b**). Meanwhile, the WHT device has a volatile characteristic in a high conductance region (**c**). This is because the trap depth exerted on the electrons is different according to the conductance state (trapped electron density). In the above case, the trapped electron density was increased by increasing pulse height. Then, the relatively easier de-trapping of the heavily trapped WHT device induced the decay of conductance with time. This can be used as a fading memory.

Figure 2-3: The structure of the 1M1R1C temporal kernel system and the I-V characteristics of the memristor used in the temporal kernel. **a,** The structure of the 1M1R1C temporal kernel system proposed in this study. The temporal kernel system can recognize images in the MNIST database through feature projection and classification. **b,** The I-V curve of the W/HfO_2/TiN memristor. The sweep order is marked in the figure. SET and RESET occurred in the positive bias and the negative bias, respectively, and gradual switching occurred in both switching conditions. Since the filament formation process is not required in this electronic switching device, no electroforming process is seen in the first sweep.

Figure 2-4: Analysis of the AC characteristics, and device structure of the W/HfO_2/TiN memristor. **a,** Changes in conductance of memristor according to pulse number. Pulse number 1~13 correspond to SET pulse, 14~26 correspond to RESET pulse, and read voltage was 0.5 V. The SET and RESET pulse heights were 4 V and -4 V, respectively, and the width of both was 200 μs. **b,** The conductance of the memristor according to the 2.5~4 V SET pulse height. Multilevel switching is possible for both SET and RESET, but the change in conductance according to the pulse number is non-linear (**a**). Also, the change in conductance according to the pulse height is non-linear as the pulse height decreases (**b**). Both nonlinearities were used for the non-linear transformation of the input in the temporal kernel. **c,** Scanning transmission electron microscopy (STEM) cross-sectional image and energy-dispersive x-ray spectroscopy (EDS) analysis results (right portion) of the fabricated W/HfO_2/TiN memristor with a depth profile. **d,** XPS spectra of the W 4f region with a depth profile and fitting results for the W/HfO_2/TiN memristor. The square dot shows the measurement result (Exp), and the black and red lines show the fitting result (Fit) and back ground (BG), respectively. Blue, green, and purple lines show XPS peaks

of tungsten (W), tungsten oxide (WO_3), and tungsten suboxide (WO_x), respectively. The sample was measured immediately after the deposition. **c, d** show that tungsten oxide was generated in the memristor.

Figure 2-5: The effects of the temperature and the cell area on the electrical properties of the device. **a,** The I-V curve at various temperatures (45~105 °C). **b,** The I-V graph of the LRS at various temperatures (45~105 °C). **c, d,** The cell area dependence of the resistance measured in 10 devices in HRS (**c**) and LRS (**d**).

Figure 2-6: The trap depth of the WHT memristor calculated from the time-dependent current-relaxation characteristics of the on and off states at various temperatures. For this test, the current was measured at the 0.5 V read voltage and the temperature was varied from 35 °C to 150 °C. **a, b,** The relaxation curves at various temperatures of the HRS (**a**) and LRS (**b**). Here, the read current was normalized to the initial current at t = 0. The data show that the read current rose (**a**) and decayed (**b**) over time as the trapped electrons were being trapped (**a**) and detrapped (**b**). These relaxation curves were fitted into the stretched exponential function [$f_\beta(t) = Ae^{-\left(\frac{t}{\tau}\right)^\beta} + B$] to attain the time constant (τ) at each temperature. **c, d,** The Arrhenius plots of ln (τ) versus 1/kT of the HRS and LRS cases. The analysis showed 0.45 eV and 0.13

eV activation energy, which correspond to the trap depth for the HRS and LRS, respectively, of the system.

Figure 2-7: The circuit used as a temporal kernel in the experiment, and the V-t graphs obtained from the DUT and CH2 of this circuit. **a,** A temporal kernel circuit composed of a memristor, resistors, and a capacitor. CH1 shows the shape of the input pulse stream, and CH2 shows the voltage applied to a 1M ohm resistor. The voltage across the DUT (green graph) is obtained by subtracting the CH2 voltage from the CH1 voltage. The left panel shows the circuit used in the pulse set (marked by pink) and the right panel shows the circuit used in DC read (marked by blue). **b,** The voltages applied to the memristor with a '0101+reference pulse' (left) and a '1010+reference pulse' (right). **c,** The voltages applied to the corresponding CH2, where the 4 V and 0 V voltage amplitudes represent ‘1’ and ‘0,’ respectively. The voltage across CH2 shows that the charging and discharging rates of the capacitor were asymmetric.

Figure 2-8: Fading memory test of the WHT memristor at the low and high conductance levels. **a**, Response of the 1M1R1C kernel machine to input patterns of '1111', '1010', '1000', and '0001' in the low conductance range. In the low conductance region, the WHT memristor has nonvolatile characteristics, so the effect of the high

signal is accumulated and the fading memory is not implemented. In contrast, the WHT memristor has a volatile characteristic in a high conductance region, and a fading memory is implemented in this region (**b**). **c, d,** Voltage applied to CH1 and CH2 for the input patterns of '1111', '1010', '1000', and '0001' in the low (**c**) and high (**d**) conductance level of the cell 1. During the measurement, a 180 pF capacitor and 390 Ω resistor were used for the 1M1R1C kernel machine. 4 V height 1 μs width pulse was used as the signal pulse and 1 V height 1 μs width pulse was used as the read pulse.

Figure 2-9: Experiment results to analyze the effect of changing parameters on the kernel characteristics in the temporal kernel system. The read current at 0.5 V of the memristor for the pulse stream '0000'~'1111' corresponds to 0~15 in the inset table in **e**. **a,** The read current at 0.5V of the memristor for each input under the conditions of 1 MΩ R_L, 4 V signal pulse height, 100 μs width, 4 V REF pulse height, and 100 μs width. **b-e,** The read current at 0.5 V of the memristor for each input when R_L, pulse width, pulse height, and REF pulse height are changed respectively from the condition of **a**. The various parameter settings for each figure were summarized in Table I. The kernel responses for each input of the temporal kernel optimized for the MNIST recognition are shown in **f**. Responses to inputs showing high prevalence in the dataset were well separated (marked by red circles).

Figure 2-10: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height and a 100 μs REF pulse width.

Figure 2-11: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 120 kΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height with a 100 μs width.

Figure 2-12: The V-t graphs for the '0000'~'1111' inputs under the conditions of 1 MΩ R_L with a 4 V signal pulse height and a 200 μs width, and a 4 V REF pulse height with a 200 μs width.

Figure 2-13: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 3.5 V signal pulse height and a 100 μs width, and a 3.5 V REF pulse height and a 100 μs width.

Figure 2-14: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 3 V REF pulse height, and a 100 μs width.

Figure 2-15: Analysis of the separation of inputs that generated net 1 spikes ('0000', '0001', '0011', '0111', and '1111'). From the conditions of a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height and a 100 μs width, R_L varies from 1 MΩ to 10 kΩ. **a-c,** The V-t graphs for the inputs that generated net 1 spikes when 1 MΩ, 120 kΩ, and 10 kΩ R_L, were used. **d-f,** The read current of the memristor for the '0000~1111' inputs under the conditions in **a-c**. When the 1 MΩ R_L was used, since the voltage distributed to

the memristor was small, SET switching did not occur after the first spike (**a**). Therefore, the responses to the inputs that generated net one spike (0, 1, 3, 7, and 15) were not separated (**d**). As R_L decreased, the voltage distributed to the memristor increased (**b-c**), and thus, the responses to the corresponding inputs were separated (**e and f**).

Figure 2-16: Analysis of the input that caused maximum conductance. **a-b,** The V-t graphs for the '1000' and '1010' inputs under the conditions of a 4 V signal pulse height and a 100 μs width. REF pulse has 4 V height and 100 μs width. 1 MΩ and 120 kΩ R_L were used for **a** and **b**. **c-d,** The read current of the memristor for the '0000~1111' inputs under the conditions in **a-b**. Since the large R_L caused slow discharging, a sufficient interval after the first spike is necessary to generate a spike that can cause large SET switching. Under the conditions in **a**, maximum conductance occurred at the '1000' input due to the slow discharging by the 1MΩ R_L (**c**). On the other hand, under the conditions in **b**, second and third spikes of sufficient magnitude to cause SET switching occurred at the '1010' input due to the fast discharging by the 120 kΩ R_L. Therefore, maximum conductance occurred at the '1010' input (**d**).

Figure 2-17: a, The V-t graphs for the '0000'~'1111' inputs under the conditions of 10 kΩ R_L with a 3.5 V signal pulse height and a 500 ns width, and a 3 V REF pulse height and a 500 ns width. **b,** The read current of the memristor for the '0000'~'1111' inputs. Insufficient charging further increased the separability for the consecutive high signals since the capacitor was not fully charged even though consecutive high signals were applied. This is suitable for situations in which consecutive signals mainly appear, such as in MNIST.

Figure 2-18: Temporal kernels with different time constants (100 ns ~ 1 s). Load resistance, parallel capacitance, and input interval used in each temporal kernel are indicated in each figure. **a-e,** The V-t graphs for the '1000' input under the condition of a 3.5 V signal pulse height. **a-c** represents a temporal kernel with similar characteristics to the temporal kernel in Figure 2-9a of main text, but with a different time constant. **d-f** represents a temporal kernel with similar characteristics to the temporal kernel in Figure 2-9f of main text, but with a different time constant.

Figure 2-19: Result of the I-V curve fitting for the WHT memristor and power consumption in the 1M1R1C kernel machine during processing one input. **a**, I-V curve fitting of the WHT memristor (HRS state) based on the conduction mechanisms. **b**, Power consumption in the 1M1R1C kernel machine (Figure 2-9f kernel condition) during input processing. Since the resistance of the WHT memristor is

dependent on the voltage, the current passing through the memristor was obtained with the HSPICE simulation using the result of the I-V curve fitting in **a**. The energy ($\int_t \ Power(t) \cdot dt$) consumed to process one input was calculated as ~25 pJ.

Figure 2-20: The temporal kernel responses were measured while increasing the number of bits processed in the temporal kernel from 3 bits to 6 bits. **a-d,** The temporal kernel responses for the '000~111', '0000~1111', '00000~11111', and '000000~111111' inputs under the conditions of 10 kΩ R_L with a 3.5 V signal pulse height and a 200 ns width, and a 3 V REF pulse height and a 200 ns width. As the number of bits processed in the temporal kernel increased, the separation of the responses to each input deteriorated.

Figure 2-21: The confusion matrices comparing the recognized digit and the desired digit for the MNIST test dataset (4 situations, from the top left: nBPK = 3 bits to the bottom right: nBPK = 6 bits) showing that the number of correct inferences decreased as the nBPK increased.

Figure 2-22: The automatic medical diagnosis system using the 1M1R1C temporal kernel and the experiment results in the two sections. **a,** A system for diagnosing the malignancy of breast lesions, which is much simpler than in the existing method (inset in **a**). In this system, ultrasonic signals are applied directly to the kernel

machine, so the imaging step is omitted. **b,** V-t graph for one echo line of a benign sample (inset in Figure 2-22b). **c,** A part of the electrocardiogram of a patient with arrhythmia. Long intervals caused by abnormal beats discharged the capacitor, and the conductance of the memristor increased in the next pulse. **d,** Five-minute temporal kernel monitoring based on the ECG of one normal patient (case 1) and two arrhythmic patients (cases 2 and 3). When arrhythmia occurred, the conductance of the memristor increased. Case 3, which had the most severe arrhythmia symptoms, showed the highest conductance.

Figure 2-23: The increase in the conductance of the memristor varied according to the degree of arrhythmia. When arrhythmia was severe, SET switching occurred in the memristor due to long discharging. **a-c,** The ECG-based V-t graphs for three cases of normal, arrhythmia, and severe arrhythmia. The electrical signal of the ECG from the heartbeat was converted into a 2.5 V, 200 ms pulse and applied to the memristor. **d,** The read current of the memristor according to the degree of arrhythmia. The more severe the arrhythmia was, the more the memristor conductance increased.

Figure 2-24: The hardware structure needed to create an array of temporal kernels that can adjust the kernel configuration. **a,** A structure in which the resistors are sequentially connected to several metal

lines. In this structure, the resistance value of the temporal kernel can be adjusted by selecting several metal lines connected to the resistors. **b,** A structure in which memristors are connected in series to the WHT memristors and parallel to the capacitors (**b left panel**). The 1M1R1C circuit can be implemented in a three-dimensional structure by stacking TiN, W metals in multi-layers and depositing a dielectric layer and top electrode in the hole after hole etching (**b right panel**). In this structure, the resistance of the memristor can be set to the desired resistance value using a method such as the incremental step pulse program (ISPP). **c**, Cell area of the diffusive memristor-based reservoir and 1M1R1C kernel. The diffusive memristor-based reservoir is implemented using a passive array composed of memristors. Therefore, $4F^2$ is required per cell (**c left panel**). If the 1M1R1C kernel is implemented with the structure in **a**, a minimum area of $8F^2$ is required per cell when using a vertical pillar transistor (T), and the area increases by $4F^2$ each time a serial resistor is added (**c middle panel**). The structure proposed in **b** requires an area of $4F^2$/cell (**c right panel**). This structure does not require an increase of area/cell even with additional elements (R_L, C) other than the memristor through a 3D integration process.

Figure 2-25: Implementation of various time constants of 1M1R1C kernel using MIS capacitor and WHT memristor (HSPICE simulation).

a, 1M1R1C circuit used in SPICE simulation. **b**, C-V curve of MIS capacitor and I-V curve of WHT memristor used for simulation and their fitting results (red line). In the simulation, an MIS capacitor (sample device) showing a capacitance of 100 pF ~ 2.2 nF was used, and the WHT memristor fitting result of Figure 2-19 was used. **c-g**, V-t graphs of the kernel showing fast discharging (left panel) and slow discharging (right panel) characteristics (pulse width 50 ns ~ 1 ms). **h**, V-t graphs in the kernel condition of 4 ~ 3.5 V pulse height, 2 ~ 1 μs pulse width, and 10 kΩ R_{CH2}. **h** shows the effect of changing pulse height on the capacitance of the TK system.

Figure 3-1. Graph to mCBA mapping. The resistance state of the (n, m) device corresponds to the weight of the (n, m) edge. The (n, n) metal cell represents the zero weight, which is the connection of the node itself.

Figure 3-2. Two operation methods of mCBA. a, Multi-ground method (MGM). **b,** Single-ground method (SGM).

Figure 3-4. Simulation results for two operation methods of mCBA. a, HSPICE array simulation results for MGM at N1. **b**, The adjacency search result of MGM at N1. **c,** HSPICE array simulation results for SGM of N1 to N9. The major and sub-current paths are marked in red and orange, respectively. **d,** Multiple paths between N1 and N9 which are not directly

connected. The shortest and sub-current paths are displayed in red and orange, which correspond to the current flow of **c**.

Figure 3-5. The main current ratio in the SGM at the various mCBA configurations. a-c, mCBA mapping (upper panel) and I-V fitting curves (red lines) for the unit cell memristor (lower panel). The main current ratios for the 9 × 9 and 100 × 100 mapping were 0.60 and 0.45, respectively, when metal cells were placed on the diagonal cells, while self-rectifying cells were placed on the rest cells. **d**, The result of calculating the ratio of $I_{main\ path}$ and I_{output} in 9 × 9, 100 × 100 mCBA under conditions of **a**, **b**, and **c**.

Figure 3-6. mCBA-array fabrication and the electrical analysis of the PAHT memristor. a, Scanning electron microscope (left) images of 9 x 9 mCBA and a cross-section transmission electron microscope image (right) of the PAHT memristor. **b,** I-V characteristic of the PAHT memristor at various set sweep voltages (2.7 V to 3.5 V). The inset of **b** is the PAHT memristor stack schematic. **c,** The surface plot of the three levels of conductance data of 9x9 mCBA.

Figure 3-7. Chemical and physical analysis of the PAHT memristor. a-c, Hf *4f*, O 1s, and Al 3d X-ray photoelectron spectroscopy (XPS) analysis at the Al_2O_3/HfO_2 interface in the PAHT device. d, Energy-dispersive X-ray spectroscopy (EDS) mapping result of the PAHT memristor in cross-section TEM.

the dendrogram. This result can confirm the results of community aggregation according to the algorithm's progress. (inset) Modularity changes according to community agglomeration. After obtaining the modularity according to the branch formation of the dendrogram, the branch is cut-off at the point corresponding to the highest value (≈ 0.37, at iteration 7). In the inset dendrogram, each bar from right to left corresponds to nodes 1 to 9.

Figure 3-30. The whole process of the SGM-based community detection algorithm. a-f, Similarity matrices and community formation at each iteration step. After initially creating the SGM-similarity matrix, the aggregation is shown in the schematic diagram in the pair with the highest value in the matrix. After the aggregation process, the SGM-similarity matrix is updated by calculating a new similarity between nodes and communities, and between communities according to the UPGMA linkage criteria. Finally, the algorithm is repeated until a single community remains (**h**).

Figure 3-31. Algorithm performance evaluation results using various graph data. a, The dendrogram plot according to the sequential community agglomeration in Zachary's karate club, Twitter retweet network, and Books about US politics dataset, and the modularity calculated at each branch of the dendrogram. **b,** Schematics of community detection results at points with

as the input vector to train the feedforward network. The hop number can be identified according to the current level.

Figure 3-36. Performance of the mCBA-based ADHD classification. a, Train and test accuracy per epoch when SGM current vector of 1-hop, 2-hop, and 3-hop pairs were all used as inputs in ADHD classification. **b,** Accuracy and AUC of SGM-based method and existing studies in ADHD classification.

List of Abbreviations

1M1R1C	One Memristor – One Resistor – One Capacitor
3D	Three-Dimensional
ADHD	Attention-Deficit/Hyperactivity Disorder
ALD	Atomic Layer Deposition
AUC	Area Under The Receiver Operating Characteristics Curve
BE	Bottom Electrode
BG	Back Ground
BL	Bit Line
BRS	Bipolar Resistive Switching
C	Capacitor
CBA	Crossbar Array
CC	Current Compliance
CH1	Channel 1
CH2	Channel 2
CNN	Convolutional neural networks
DC	Direct Current
DUT	Device-Under-Test
eBRS	Electronic Bipolar Resistive Switching
ECG	Electrocardiogram
EDS	Energy-Dispersive X-ray Spectroscopy
F	Feature Size
FCN	Fully Connected Network

fMRI	Functional Magnetic Resonance Imaging
HRS	High Resistance State
HSPICE	Hewlett-Simulation Program with Integrated Circuit Emphasis
I	Current
I_{read}	Read Current
ISPP	incremental step pulse program
k	Heuristic Constant
L	Number of Landmarks
LM	Landmark
LRS	Low Resistance State
M	Memristor
MAC	Multiplication and Accumulation
mCBA	Metal-Cell-at-Diagonal CBA
MCV	Memristor Conductance Vector
MGM	Multi-Ground Method
N	Number of Nodes
nBPK	Number of Bits Processed by the Kernel
NC	Neurotypical Controls
NN	Neighbor Node
OSC	Oscilloscope
PAHT	$Pt/Al_2O_3/HfO_2/TiN$ device
PG	Pulse Generator
R	Resistor
RC	Reservoir Computing

ReRAM	Resistive Switching Memory
R_{HRS}	Memristor's High Resistance Level
R_L	Load Resistor
R_{LRS}	Memristor's Low Resistance Level
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
SEM	Scanning Electron Microscope
SGM	Single-Ground Method
SNS	Social Network Service
SPA	Semiconductor Parameters Analyzer
STEM	Scaning Transmission Electron Microscopy
t	Time
T	Transistor
TE	Top Electrode
TEM	Transmission Electron Microscopy
TK	Temporal Kernel
TN	Target Node
V	Voltage
V_{CH1}	Voltage at Channel 1
V_{CH2}	Voltage at Channel 2
V_{DUT}	Voltage of the Device-Under-Test
VMM	Vector-Matrix Multiplication
WHT	W/HfO_2/TiN device
WL	Word Line

XPS	X-ray Photoelectron Spectroscopy
τ	Time Constant

1. Introduction

1.1. Memristor-based Physical Computing for Complex Data Processing

As the amount of information to be processed is rapidly increasing with the advances in deep learning technology, the limited processing efficiency of conventional hardware became a serious problem that impedes performance enhancement in the modern computing system. This motivates the need of exploring new data processing techniques using novel hardware structures to enable the processing of complex data. In this regard, resistive switching random access memory (ReRAM) is a potential candidate for futuristic physical computing implementation. Using the intrinsic physical properties of memristive hardware enables effective data analysis.

Temporal data has a wide range of frequencies, and the kernel characteristics required for each data vary. Recent advances in physical reservoir computing, which is a type of temporal kernel, have made it possible to perform complicated timing-related tasks using a linear classifier. However, the fixed reservoir dynamics in previous studies have limited application fields. This study proposed memristor (M), resistor (R), and capacitor (C)-combined structure showing unique circuit characteristics due to the nonlinear I-V of memristors. The 1M1R1C structure can serve as a kernel capable of processing

various temporal signals. The 1M1R1C temporal kernel was used to identify the static MNIST data, and showed high performance in terms of accuracy, energy efficiency, and processing speed. The system was adopted to recognize the sequential data, ultrasound (malignancy of lesions), and electrocardiogram (arrhythmia), which had a significantly different time constant (10^{-7} vs. 1 s). The suggested system feasibly performed the tasks by simply varying the capacitance and resistance. These functionalities demonstrate the high adaptability of the present temporal kernel compared to the previous ones.[1]-[4]

Another type of complex data is graph data. Graph data differs from other data in that it includes connectivity between entities. Graph data is mostly non-Euclidean type and is difficult to vectorize, making it difficult to process in the existing hardware structure. In this study, graph data was analyzed using the induced sneak current of the self-rectifying memristor crossbar array. The results of implementing various graph algorithms based on memristive CBA and applying them to real-world problems show that the intrinsic properties of crossbars are very effective in analyzing graph structures.

1.2. Objective and Chapter Overview

The objective of the present thesis is focused on complex data processing with memristor-based physical computing. Intrinsic physical properties (R-C delay, I-V nonlinearity, sneak current) of the memristive hardware were used for physical computing.

Chapter 2 describes a new method of sequential data processing using a nonvolatile memristor-based temporal kernel with time constants controllability. A temporal kernel was constructed using memristors (M), resistors (R), and capacitors (C) for effective sequential data processing. The unit cell has a 1M1R1C structure in which a memristor is connected in series with a resistor and a capacitor, and the resistor and capacitor are connected in parallel with each other. The 1M1R1C kernel has the advantage of being applicable to various situations as it can have various time constants through R and C control. 1M1R1C-based MNIST recognition showed high accuracy (90%) with high energy efficiency and fast processing speed. In addition, the 1M1R1C kernel was applied to ultrasound[5] and electrocardiogram-based medical diagnosis[6] with very different time constants (frequency range of 1 to 10 MHz).

Chapter 3 introduces a method for processing non-Euclidean graphs using self-rectifying memristor arrays is proposed. The non-Euclidean graphs were mapped to the metal-cell-at-diagonal crossbar-array (mCBA), composed of the self-rectifying memristors. The sneak current, an intrinsic physical property in

the mCBA, allows identifying the similarity function. Sneak current-based similarity function indicates the distance between nodes, connectivity between communities and nodes, the probability that unconnected nodes will be connected in the future, and the neural activity between cortices. This work shows a feasible demonstration of the memristor-based physical calculation, being applied to various graphical problems.

Finally, in chapter 4, the conclusion of the thesis is made.

1.3. References

[1] Midya, R. *et al.* Reservoir Computing Using Diffusive Memristors. *Adv. Intell. Syst.* **1**, 1900084 (2019).

[2] Du, C. *et al.* Reservoir computing using dynamic memristors for temporal information processing. *Nat. Commun.* **8**, 1–10 (2017).

[3] Moon, J. *et al.* Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2**, 480–487 (2019).

[4] Zhu, X., Wang, Q. & Lu, W. D. Memristor networks for real-time neural activity analysis. *Nat. Commun.* **11**, (2020).

[5] Piotrzkowska-Wróblewska, H., Dobruch-Sobczak, K., Byra, M. & Nowicki, A. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Med. Phys.* **44**, 6105–6109 (2017).

[6] Moody, G. B. & Mark, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **20**, 45–50 (2001).

2. Time-varying data processing with nonvolatile memristor-based temporal kernel

2.1. Introduction

Convolutional neural networks (CNN), which are composed of a convolutional layer and a fully connected layer[1], show outstanding performance in static image processing (recognition and classification).[2], [3] However, when the temporal order of each input vector and the correlation between the input vectors are essential, such as for natural language recognition or translation, a method of processing the input over time is required, and CNN are not suitable for this purpose.[4] Such an event sequence or time-dependent network operation can generally be represented by the relationship between the present network state, the input, and the previous network state.

A typical network with such characteristics is a recurrent neural network (RNN) with the long-short-term memory learning rule,[5] which mitigates the vanishing gradient descent problem of the classical RNN.[6] Nonetheless, these artificial neural networks perform vast amounts of multiplication and accumulation (MAC) operations during the learning and inference steps. When these calculations are performed using the conventional architecture in which the computing unit and memory are separated, even with the latest graphics

processing unit, the cost of achieving the required processing speed and the energy consumption are enormous.[7]

In this regard, the recent upsurge of studies on neural networks that use a memristor-based cross-bar array (CBA) based on Ohm's law and Kirchoff's law is notable.[8]-[13] If the memristor used in such neural networks can process the event-sequence-related and temporal information, it can achieve RNN functionality. An even more desirable functionality is to extract the features of the input information (raw data vector) using a temporal kernel (TK) and feed them to the next classification layer. A representative example of such a computing system is reservoir computing (RC), which is composed of a reservoir and a readout layer (FCN).[14], [15]

The core part of the RC system is the reservoir, where the non-linear transformation of the input signal is performed based on the fading memory properties, and the characteristics of the input signal are projected into a rich enough feature space. The result of the projection is called the *reservoir state*.[16]

The nonlinear dynamic filtering of RC can be regarded as a specific type of a more general TK[17]-[19], in which the time-varying data can be efficiently handled by the fading-memory functionality of the reservoir. Nonetheless, RC may have severe limitations in adapting different time scale of the input data due to its fixed time constant of the specific fading memory function. This may not be the case for other types of TK, based on a physical kernel combined with other circuit elements, as shown in this work. Also, non-fading (or nonvolatile)

memory can be used as the TK because the time-varying input can be encoded into the TK by the effects of the time constant of the entire circuit element. When a memristor is used as the TK, its resistance must be determined by the different input pulse signals with varying amplitudes and the intervals between such input signals. If the input signals have simple and obviously distinguishable patterns, a memristor can sufficiently discern them by assigning different resistance values. However, for complicated and similar input patterns, high separability is required, which is usually challenging to achieve with a given type of memristor.[20], [21] Also, the input signals could have substantially different time constants, which further severely limits the memristor-based temporal kernel (reservoir).[22], [23] In this case, a high-performance kernel machine applicable to diverse circumstances can be created by incorporating additional circuit components.

Recently, various studies were conducted on hardware-based RC systems that use volatile memristors, in which a volatile memristor was used to process a time-varying input.[20]-[23] In those studies, the reservoirs were constructed based on ionic diffusion dynamics (diffusive memristors), in which the the spotaneously decaying conductance of low-resistance state (LRS) of the diffusive memristor provided the fading memory function of a reservoir.

However, there are several limitations in using such reservoir dynamics. Firstly, the duration and interval of the input signal are limited to the time range in which sufficient conductance decay occurs. For this reason, in the previous

studies, it took 1 to 20 ms for one memristor to process 4-bit data, which is insufficient for processing a large amount of data.[20],[21] Secondly, obtaining a reproducible reservoir state could be challenging. An Ag-filament-based diffusive memristor exhibits stochastic switching,[20] so the variation of the reservoir state will be large. Finally, reservoir adaptation could be difficult to achieve, given that the reservoir dynamics are totally determined by the material property, which renders the previous system useful only for applications with a time scale similar to that of the specific memristor.[21]-[23]

In this study, a device based on an electron trap/detrap mechanism was used to solve the aforementioned issues.[24], [25] A W/HfO_2/TiN (WHT) memristor goes into an LRS when the trap is filled with electrons and shifts to a high-resistance state (HRS) when the trapped electrons are detrapped. Since the resistance switching is based on the electron trapping and not the ionic movement, reproducible results can be achieved (Figure 2-1).[26], [27] In addition, since the work functions between the top and bottom electrodes differ only slightly, there is limited built-in potential, so the device has high retention properties (Figure 2-2a, b).[25], [28] Although the WHT memristor has different time constants of operation according to its conductance level (Figure 2-2c), it is insufficient to achieve adaptability with a sufficiently large time constant range. This problem could be solved by combining the memristor with a capacitor (C) and a normal resistor (R). Under this circumstance, the R-C time constant of the circuit can be varied, and the memristor response to the temporal

arrangement of the inputs can be controlled.

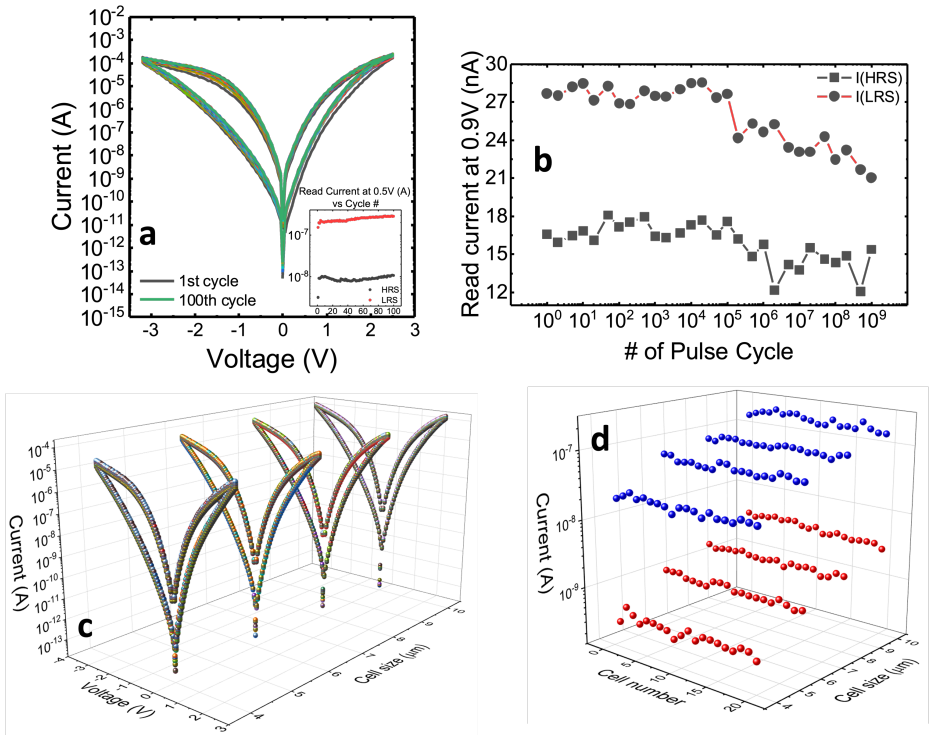


Figure 2-1: Experimental results on device reliability and reproducibility. **a,** Cycle-to-cycle variation of the WHT memristor. Except for the first cycle out of 100 DC cycles (2.5 ~ -3.2 V), there was a slight variation in the I-V curve. The inset of (**a**) shows the read current at 0.5V for each cycle number. **b,** Endurance of the WHT memristor. The WHT memristor showed stable resistive switching behavior during ~10^5 pulse cycles. For endurance measurement, a 3.3 V height 1 μs width SET pulse and -3.35V height 1.5 μs width RESET pulse were used. The read current was recorded with DC read at 0.9 V and a WHT memristor with 4 μm cell size was used for measurement. **c, d,** Cell to cell variation of

the WHT memristor. A total of 80 devices were measured with 20 devices each of 4 μm x 4 μm, 6 μm x 6 μm, 8 μm x 8 μm, and 10 μm x 10 μm. An I-V curve was obtained in each device through a 2.5 V ~ -3.2 V DC cycle (**c**), and the read current was extracted at 0.5 V of each I-V curve (**d**). Data shown in red is read current in HRS and data shown in blue is read current in LRS.

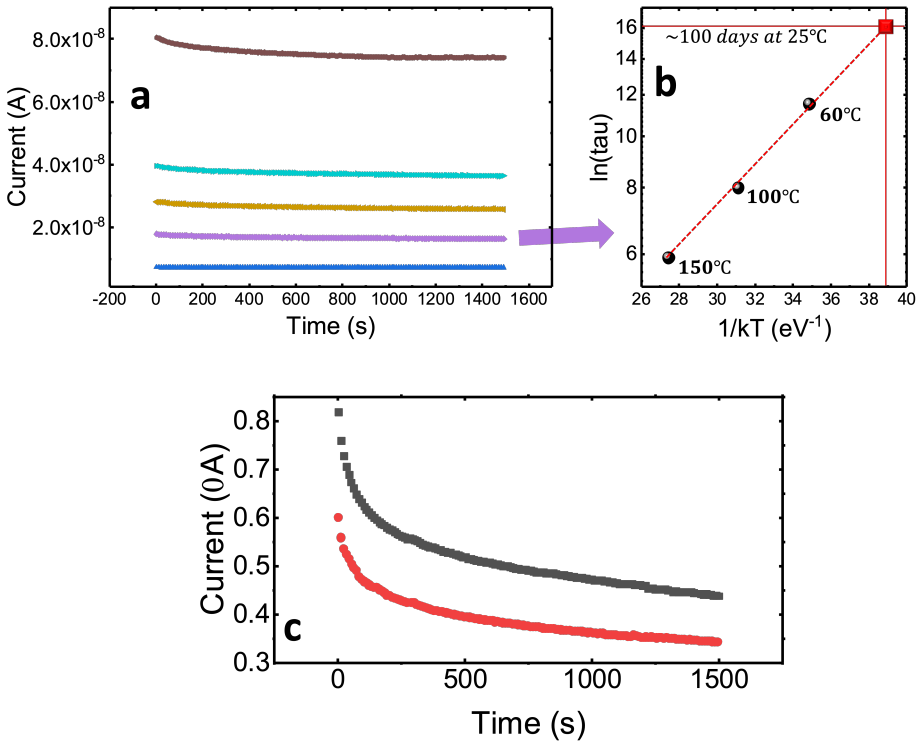


Figure 2-2: Retention measurement result of the WHT device. The WHT device has a nonvolatile characteristic in the low conductance range (**a**) and a retention time of about 100 days at 25 °C, which is the result of extrapolation based on 60 ~ 150 °C retention data (**b**). Meanwhile, the WHT device has a volatile characteristic in a high conductance region (**c**). This is because the trap depth exerted on the electrons is different according to the conductance state (trapped electron density). In the above case, the trapped electron density was increased by increasing pulse height. Then, the relatively easier detrapping of the heavily trapped WHT device induced the decay of conductance with time. This can be used as the fading memory.

2.2. Experimental

The array of cross-bar-type W/HfO_2/TiN memristors was fabricated. A 50 nm-thick TiN layer was sputtered (Endura, Applied Materials) on an SiO_2/Si substrate, and the TiN layer was patterned into a line shape to form a BE. The 2- to 10 μm-wide TiN BEs were patterned using conventional photolithography and the dry-etching system. After the patterning, the residual photoresist was removed with acetone and cleaned sequentially with deionized water. Then 4 nm HfO_2 was deposited using atomic layer deposition (ALD) at a 280 °C substrate temperature using a traveling-wave-type ALD reactor (CN-1 Co. Plus 200). A tetrakis-ethlylmethylamido hafnium (TEMA-Hf) and O_3 were used as precursors for Hf and oxygen, respectively. On the HfO_2 layer, 50-nm-thick W TEs were sputtered using the MHS-1500 sputtering system and patterned into 2- to 10 μm-wide lines using the conventional lift-off process. After the fabrication, the WHT device was analyzed using x-ray photoelectron spectroscopy (XPS, AXIS SUPRA, Kratos) and energy-dispersive x-ray spectroscopy (EDS, JEOL, JEM-ARM200F) to observe the formation of the tungsten oxide layer. Cross-sectional transmission electron microscope (TEM) images of the WHT memristor were observed using scaning transmission electron microscopy (STEM, JEOL, JEM-ARM200F).

2.3. Results and Discussions

Figure 2-3a shows the TK system that can control the kernel dynamics using a memristor, a normal resistor, and a capacitor (1M1R1C). This is a structure in which the reservoir is replaced with a 1M1R1C temporal kernel while maintaining the computing scheme of the RC system. In this TK system, the charging and discharging of the capacitor transforms the signals applied to the device into various forms so that the conductance state of the memristor can be varied depending on the magnitude and sequential arrangement of the input signal (Figure 2-4a, b). The results of input processing in the kernel form a memristor conductance vector (MCV), which becomes the input of the subsequent FCN readout layer. Such a configuration of the TK system allows the arbitrary variation of the response dynamics by adjusting the sizes of the resistor, capacitor, and pulse width, etc. Therefore, the optimized TK system can be configured for tasks with vastly different time scales.

Device analysis. Figure 2-3b shows the measured current-voltage (I-V) curve of the WHT device. During the electrical measurement, the W top electrode (TE) was biased, while the TiN bottom electrode (BE) was electrically grounded. The resistance of the device was changed from HRS to LRS by a positive bias (SET), and reverse switching was achieved by a negative bias

(RESET). In both SET and RESET, gradual switching appeared, as shown in Figure 2-3b and Figure 2-4a, b, which contributed to the high performance of the TK system. Figure 2-4c shows the cross-sectional scanning transmission electron microscopy (STEM) image of the WHT device, which revealed the W TE, the TiN BE, and the 4 nm-thick HfO_2 layer between the TE and BE. Figure 2-4d shows the X-ray photoelectron spectroscopy (XPS) analysis of the W/HfO_2 interface in the WHT device. Analysis of the W peak in the XPS data revealed the presence of tungsten sub-oxide (WO_x, $x<2$) and a WO_3 layer. The energy-dispersive X-ray spectroscopy line scan result (Figure 2-4c, right portion) along the vertical line from TE to BE in the STEM image implies that a thin WO_3 was formed at the W/HfO_2 interface and WO_x ($x << 3$) was formed within the W bulk. Therefore, the WO_x may work as a voltage divider when the voltage is applied to the device, which will cause gradual SET and RESET performance.[29] This is a favorable characteristic, allowing the TK to have various states. Moreover, this WHT device does not have an electroforming step (Figure 2-3b), which also contributed to the stable resistance switching operation (Figure 2-5 and 2-6). W and TiN have similar work functions of ~ 4.5 eV, which may render the energy band profile symmetric.[30], [31] The symmetric energy band profile is unfavorable for fluent electronic bipolar resistive switching (eBRS).[25], [28] However, the WO_3 layer formed at the W/HfO_2 interface can induce a Schottky barrier, whereas the HfO_2/TiN interface constitutes a quasi ohmic contact.[29], [32] Especially, the chemical interaction

between the HfO_2 and TiN layers can produce defects within the HfO_2 layer, which provide the system with the electron traps that are necessary to induce the eBRS mechanism. With the application of the positive bias to the TE, the traps were filled with electrons that were injected from the TiN BE through the quasi-ohmic contact, which switched the device to the LRS. Conversely, when the negative bias was applied, the device switched back to the HRS as the trapped electrons were detrapped, while the electron injection from the TE was blocked by the Schottky barrier at the W/HfO_2 interface.[28] Due to the presence of the WO_x layer, there was no need to set current compliance (CC) during the operation.

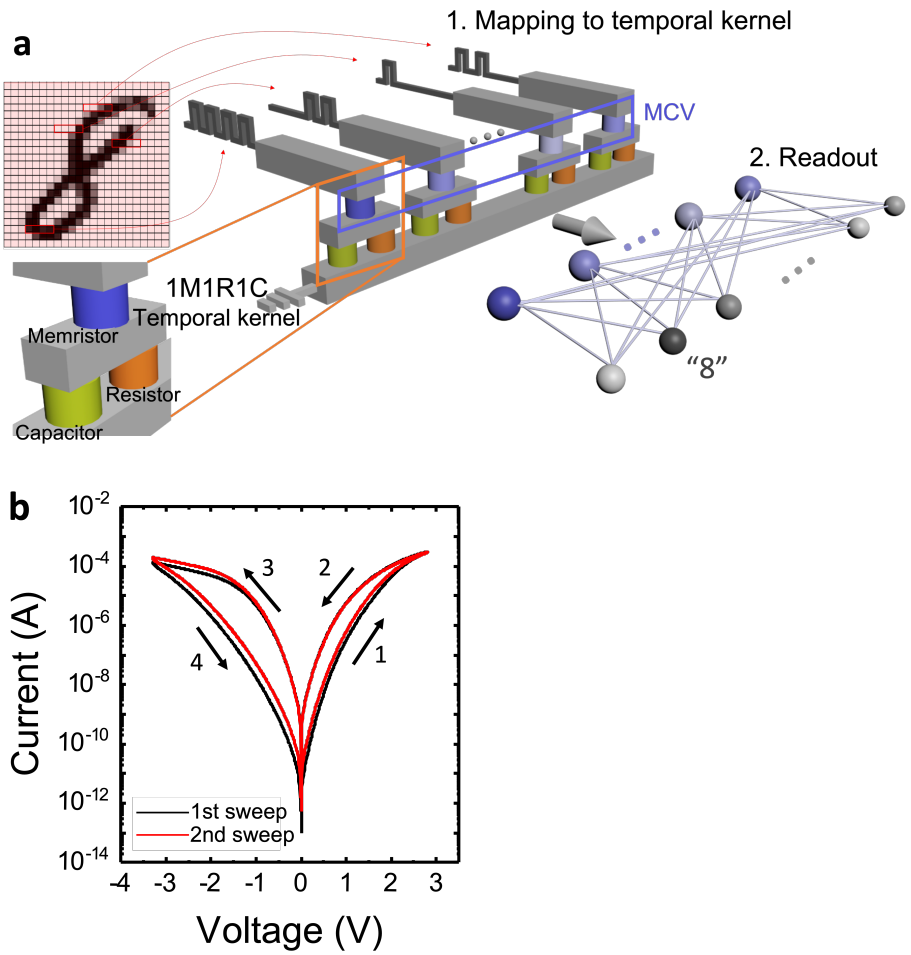


Figure 2-3: The structure of the 1M1R1C temporal kernel system and the I-V characteristics of the memristor used in the temporal kernel. **a,** The structure of the 1M1R1C temporal kernel system proposed in this study. The temporal kernel system can recognize images in the MNIST database through feature projection and classification. **b,** The I-V curve of the W/HfO_2/TiN memristor. The sweep order is marked in the figure. SET and RESET occurred in the positive bias and the negative bias, respectively, and gradual switching

occurred in both switching conditions. Since the filament formation process is not required in this electronic switching device, no electroforming process is seen in the first sweep.

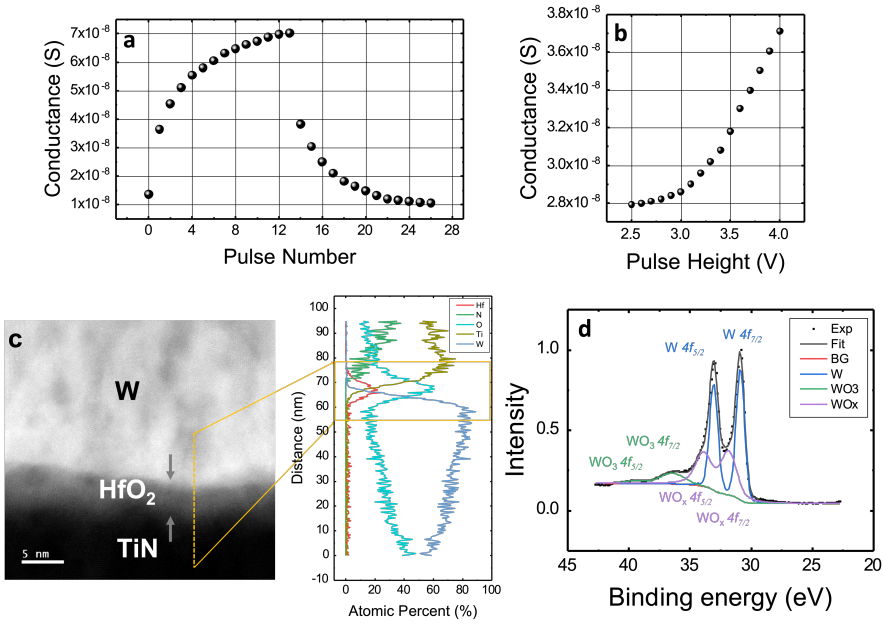


Figure 2-4: Analysis of the AC characteristics, and device structure of the W/HfO_2/TiN memristor. **a,** Changes in conductance of memristor according to pulse number. Pulse number 1~13 correspond to SET pulse, 14~26 correspond to RESET pulse, and read voltage was 0.5 V. The SET and RESET pulse heights were 4 V and -4 V, respectively, and the width of both was 200 μs. **b,** The conductance of the memristor according to the 2.5~4 V SET pulse height. Multilevel switching is possible for both SET and RESET, but the change in conductance according to the pulse number is non-linear (**a**). Also, the change in conductance according to the pulse height is non-linear as the pulse height decreases (**b**). Both nonlinearities were used for the non-linear transformation of the input in the temporal kernel. **c,** Scanning transmission electron

microscopy (STEM) cross-sectional image and energy-dispersive x-ray spectroscopy (EDS) analysis results (right portion) of the fabricated W/HfO_2/TiN memristor with a depth profile. **d,** XPS spectra of the W 4f region with a depth profile and fitting results for the W/HfO_2/TiN memristor. The square dot shows the measurement result (Exp), and the black and red lines show the fitting result (Fit) and back ground (BG), respectively. Blue, green, and purple lines show XPS peaks of tungsten (W), tungsten oxide (WO_3), and tungsten suboxide (WO_x), respectively. The sample was measured immediately after the deposition. **c, d** show that tungsten oxide was generated in the memristor.

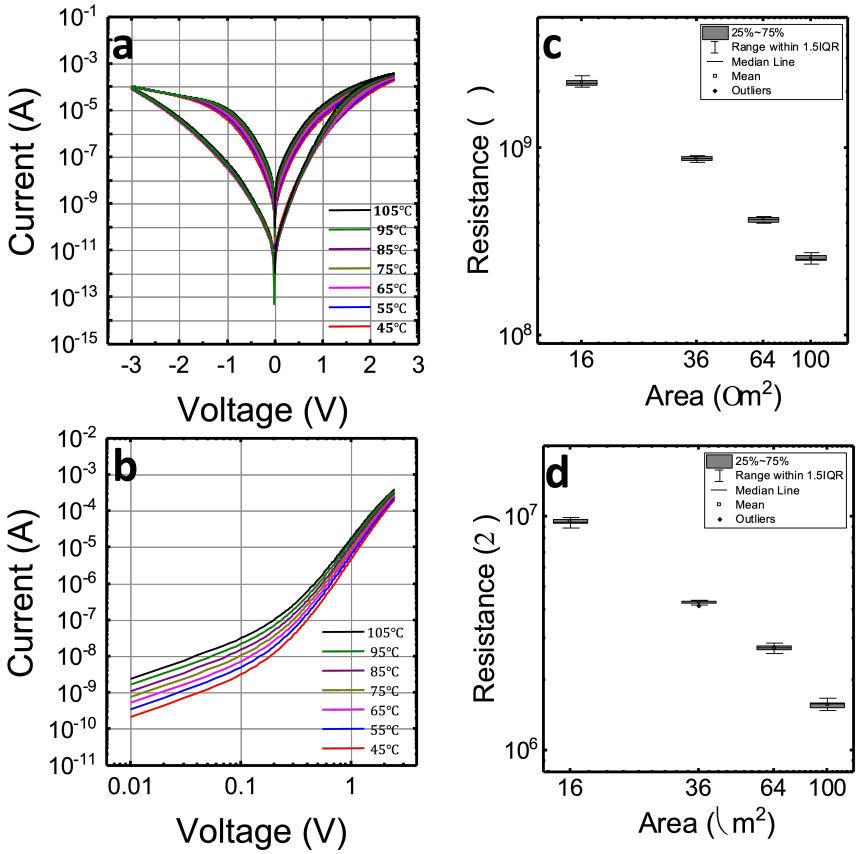


Figure 2-5: The effects of the temperature and the cell area on the electrical properties of the device. **a,** The I-V curve at various temperatures (45~105 °C). **b,** The I-V graph of the LRS at various temperatures (45~105 °C). **c, d,** The cell area dependence of the resistance measured in 10 devices in HRS (**c**) and LRS (**d**).

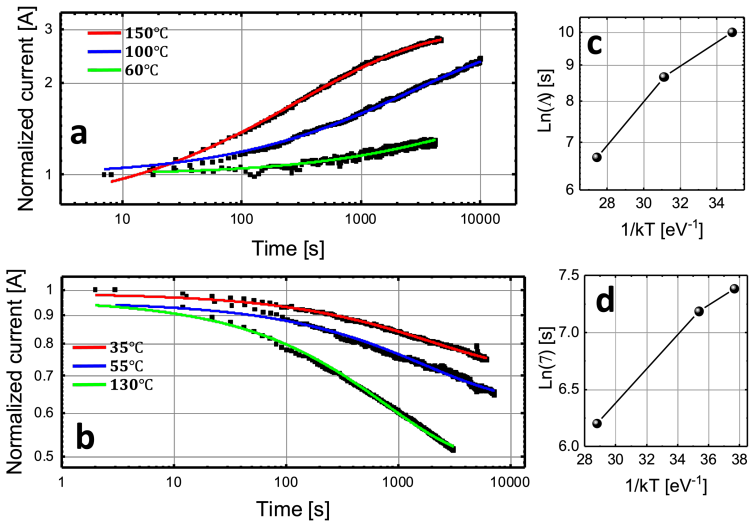


Figure 2-6: The trap depth of the WHT memristor calculated from the time-dependent current-relaxation characteristics of the on and off states at various temperatures. For this test, the current was measured at the 0.5 V read voltage and the temperature was varied from 35 °C to 150 °C. **a, b,** The relaxation curves at various temperatures of the HRS (**a**) and LRS (**b**). Here, the read current was normalized to the initial current at t = 0. The data show that the read current rose (**a**) and decayed (**b**) over time as the trapped electrons were being trapped (**a**) and detrapped (**b**). These relaxation curves were fitted into the stretched exponential function [$f_\beta(t) = Ae^{-\left(\frac{t}{\tau}\right)^\beta} + B$] to attain the time constant (τ) at each temperature. **c, d,** The Arrhenius plots of ln (τ) versus 1/kT of the HRS and LRS cases. The analysis showed 0.45 eV and 0.13 eV activation energy, which correspond to the trap depth for the HRS and LRS, respectively, of the system.

Temporal kernel generation. We implemented the TK by configuring the circuit, as shown in Figure 2-7a. Pulse streams were generated by a pulse generator (PG), where input signal '1' is converted to a high level, and '0' is converted to a low level. These pulse streams were delivered to channel 1 (CH1) and channel 2 (CH2) of an oscilloscope (OSC). A 50 Ω resistor was assigned to CH1, which allowed monitoring of the input pulse shape. In CH2, a 1 MΩ resistor was connected to the device-under-test (DUT, the memristor) in series. From the estimated voltage from the CH2 resistor, the voltage applied to the DUT was inferred. Since the oscilloscope fixes the size of the CH2 resistor at 1 MΩ, the overall series resistance to the memristor was adjusted by connecting a load resistor (R_L), as shown in the figure. Also, a capacitor was connected to the CH2 resistor in parallel, which stored the charge supplied by the applied pulse voltage. In this specific experimental setup, its value was fixed at 180 pF, but the dynamic time constant of the TK system was varied by changing R_L and the capacitance. The measurement consisted of two steps. In the first step, a pulse was generated at the PG, which caused SET switching in the memristor, while the circuit part with the semiconductor parameter analyzer (SPA) was deactivated (Figure 2-7a left panel). In the second step, the conductance state of the memristor was read through the DC sweep using the SPA, while the other parts of the circuit were deactivated (Figure 2-7a right panel). To compose the temporal kernel circuit, the WHT device with an area of 10 μm × 10 μm was connected to the pulse generator (PG, Agilent 81110A) and an oscilloscope

(OSC). A 1M1R1C circuit was constructed by adding a load resistor to the circuit and setting the resistance values of CH1 and CH2 in the OSC to 50 Ω and 1 MΩ, respectively. A semiconductor parameter analyzer (SPA, Hewlett-Packard 4145B) was connected to the WHT device to monitor the DC sweeps. To process the static and sequential data, the device states after the pulse streams were measured. After the measurement, the device was reset to the HRS state and the process was repeated. The TK state was constructed based on the recorded device states, and the readout layer was trained based on it.

Figure 2-7b shows the voltages transients over the memristor with a '0101+reference pulse' (left) and a '1010+reference pulse' (right), and Figure 2-7c shows the corresponding voltage transients read at CH2. In these operations, 4 V, 200 μs, and 0 V, 200 μs pulses were programmed to represent '1' and '0', respectively. The initial resistance of the WHT memristor was set to 50 MΩ when measured at 0.5 V. The role of the last reference pulse is explained as follows. The left panels of Figure 2-7b and c show that since the first signal was '0', no voltage appeared up to 0.2 ms. When the first '1' signal was applied, the DUT showed a peak of up to ~ 3.5 V due to the involvement of the capacitive charging current, and it decayed to ~ 1.5 V after the capacitor charging was completed. At the same time, the CH2 voltage showed a corresponding gradual increase in the capacitor voltage, which was saturated at ~ 2.5 V. When the second '0' signal came in, the capacitor was discharged and the reverse current flowed into the DUT, which made its voltage negative, while

the CH2 showed gradual decay of the capacitor voltage. It was noted from the CH2 voltage that the capacitor was not completely discharged during the 0.2 ms duration of '0' signal, so when the subsequent '1' signal came in, the capacitive charging current was not as high as in the previous '1' signal case (where the DUT voltage peaked only up to ~ 2.5 V). Such an effect can be more evidently seen with the subsequent '1' signal (the reference pulse), as there was almost no peak in the DUT. Therefore, in this case, the effective number of SET pulses applied to the DUT was only two (the first and second '1' among the total three '1's in the '01011' sequence). After the entire pulse sequence was over, the memristor resistance was 28.2 MΩ.

In the case of the right panels in Figure 2-7b and c, in contrast, each of the 1 signals is separated by 0 signals, and all the three '1's in the '10101' sequence are effective, and they switched the DUT to the SET state, which made its resistance 26.7 MΩ, despite the application of the same number of set pulses (three) in the two cases. It should be noted, however, that the last two peaks had a lower effect in decreasing the memristor resistance than the first one due to its lower peak height, which was induced by the incomplete discharging of the capacitor during the intervening '0' pulse cycle. This is not a demerit but actually a merit of this TK system, which allowed even higher separability and adaptability. Therefore, this TK system can recognize not only the different input pulse numbers but also their timing. Figure 2-7b and c show several notable features. First, due to the built-in asymmetry of the band profile

of the WHT memristor, the resistance at the positive bias of ~ 2.5 V was ~ 100 times lower than that at the negative bias of ~ 1.5 V. Therefore, the charging was much faster than the discharging. This is the first factor that allows the TK system to have higher separability and adaptability. Second, the capacitance and R_L can be arbitrarily taken to vary the charging and discharging times, which can eventually affect the effectiveness of the voltage pulse application to the memristor. Third, the input voltage pulse height and duration are another knob that can further change the TK dynamics. These features rendered the TK system flexible and adaptable to the various requirements, as shown in the next sections. Without the last reference pulse, such a systematic variation and examination of the memristor state control would have been improbable.

The WHT memristor in this study shows both nonvolatile and volatile memory properties, when its conductance is low and high, respectively. In this study, the WHT memristor was operated within the conductance range showing nonvolatile characteristics, but outside that range, the WHT device shows fading conductance state (Figure 2-2c). Therefore, depending on the operation scheme, the 1M1R1C kernel can also perform a reservoir function, and the results are shown in Figure 2-8. In this study, time series data were processed based on the unique characteristics of 1M1R1C, not the fading memory.

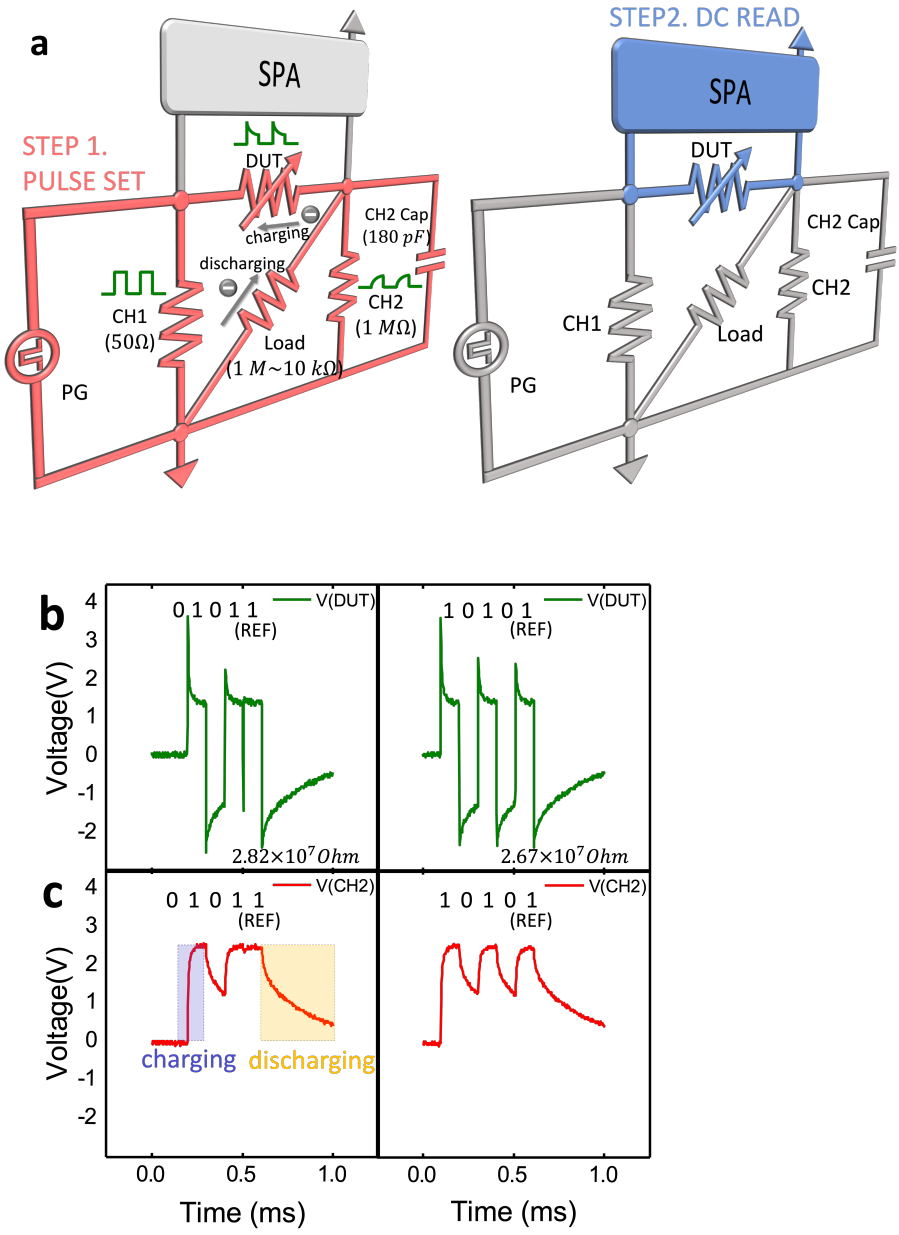


Figure 2-7: The circuit used as a temporal kernel in the experiment, and the V-t graphs obtained from the DUT and CH2 of this circuit. **a,** A temporal kernel circuit composed of a memristor, resistors, and a capacitor. CH1 shows the shape of the input pulse stream, and

CH2 shows the voltage applied to a 1M ohm resistor. The voltage across the DUT (green graph) is obtained by subtracting the CH2 voltage from the CH1 voltage. The left panel shows the circuit used in the pulse set (marked by pink) and the right panel shows the circuit used in DC read (marked by blue). **b,** The voltages applied to the memristor with a '0101+reference pulse' (left) and a '1010+reference pulse' (right). **c,** The voltages applied to the corresponding CH2, where the 4 V and 0 V voltage amplitudes represent ‘1’ and ‘0,’ respectively. The voltage across CH2 shows that the charging and discharging rates of the capacitor were asymmetric.

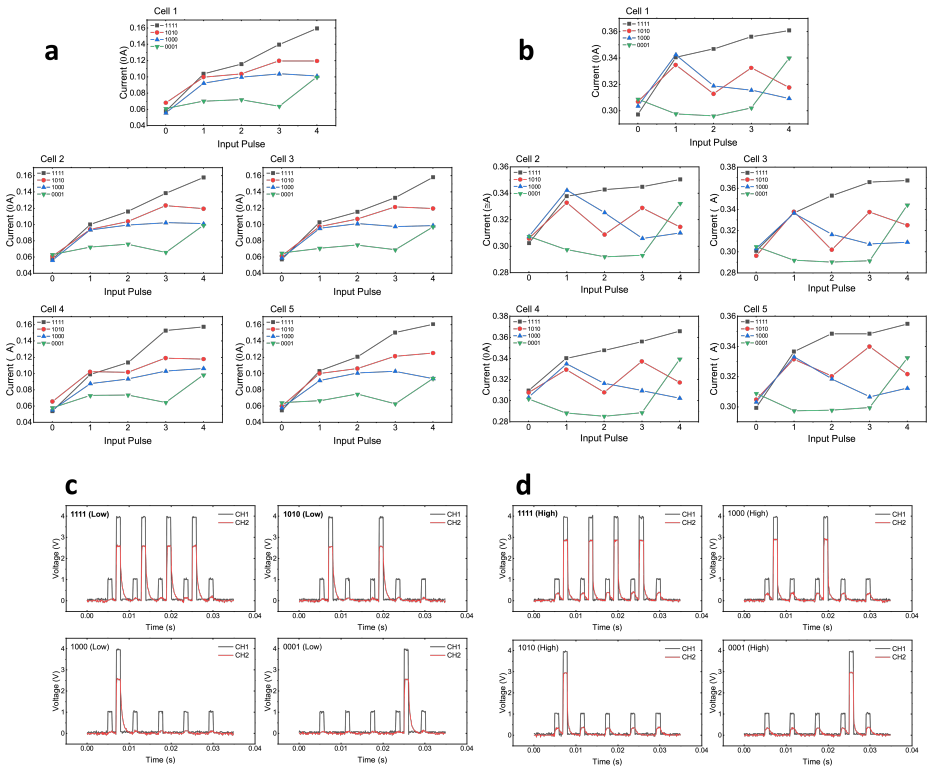


Figure 2-8: Fading memory test of the WHT memristor at the low and high conductance levels. **a**, Response of the 1M1R1C kernel machine to input patterns of '1111', '1010', '1000', and '0001' in the low conductance range. In the low conductance region, the WHT memristor has nonvolatile characteristics, so the effect of the high signal is accumulated and the fading memory is not implemented. In contrast, the WHT memristor has a volatile characteristic in a high conductance region, and a fading memory is implemented in this region (**b**). **c, d,** Voltage applied to CH1 and CH2 for the input patterns of '1111', '1010', '1000', and '0001' in the low (**c**) and high (**d**) conductance level of the cell 1. During the measurement, 180 pF capacitor and 390 Ω resistor were used for the 1M1R1C kernel

machine. 4 V height 1 μs width pulse was used as signal pulse and 1 V height 1 μs width pulse was used as the read pulse.

Modifying the temporal kernel dynamics. In this TK system with the given WHT memristor property and capacitance, R_L and the pulse height/duration were varied to examine the separability of the memristor. The capacitance could also be varied, but it was fixed in this experiment section. Figure 2-9 shows several examples of the different degrees of separability of the TK system when these parameters were varied. The examples show the current value read at 0.5 V after the 16 different input patterns, from '0000' to '1111', were programmed to the PG, with the additional reference pulse added last. Since the output current depends on the initial resistance, the resistance of the WHT memristor in this experiment was reset to a constant value (50 MΩ at 0.5 V) before measurement. The x-axis numbers correspond to the different input patterns described in the inset table in Figure 2-9e, and the different parameters, such as R_L, the input pulse, and the reference pulse, for each graph in Figure 2-9 are summarized in Table 2-1. It should be noted that in Figure 2-9, the y-axis scales of each graph were varied to easily compare them. All the detailed pulse responses and analyses are included in Figures 2-10 to 2-14. In Figure 2-9a, wherein R_L = 1 MΩ, the signal pulse = 4 V, 100 μs, and the reference pulse = 4 V, 100 μs, the five patterns, '0000', '0001', '0011', '0111', and '1111' are not clearly distinguished (an analysis of the separation of these inputs is shown in Figure 2-15). It was also noted that the '1000' pattern resulted in the highest memristor conductance, although there were only two SET pulses (the first 1 and the reference pulse at the last SET pulse). This is because the reference

pulse induced the highest peak voltage to the memristor because the interval between the two pulses, during which the capacitor was fully discharged, was the longest (the details are shown in Figure 2-16).

Of the six graphs in Figure 2-9, Figure 2-9c shows well the critical features of this TK system. The only difference of Figure 2-9c from Figure 2-9a is the pulse length [100 μs (in a) vs. 200 μs (in c)]. As the pulse width increases, the capacitor discharging during the 0 input increased, and the subsequent '1' induced a higher peak voltage. The conductance levels in Figure 2-9c can be clearly grouped into three levels, which are determined by the number of 1's immediately after the '0' (not the total number of '1'). For example, '0000' has only one 1 after 0 (the reference pulse), so it induced the lowest conductance. Interestingly, '1111' has the same low conductance even though it had five 1 inputs (including the reference pulse). This is because the only effective '1' was the first one because all the other '1's do not have the preceding '0's, so they cannot produce peak voltage.

Another characteristic and most desirable setting could be seen in Figure 2-9f, in which R_L was decreased to 10 kΩ and the pulse width was decreased to 200 ns. This setting makes the capacitor charging per one voltage pulse ('1' signal) insufficient and its discharging during the '0' signals faster. Overall, this makes the memristor conductance more linearly dependent on the total number of '1's, as shown in Figure 2-9f (an example of insufficient charging and details of the effects are included in Figure 2-17). A short pulse length is also beneficial

to rapidly process the input vectors.

By appropriately changing both the C and R_L, the kernel characteristics obtained in Figure 2-9 could be implemented at different time scales. Additional kernels are configured as the time constants in Figure 2-18. Based on the analysis of the effect of each parameter change, a kernel condition suitable for the task is determined through kernel adaptation, and ex-situ training is performed, which is followed by inference.

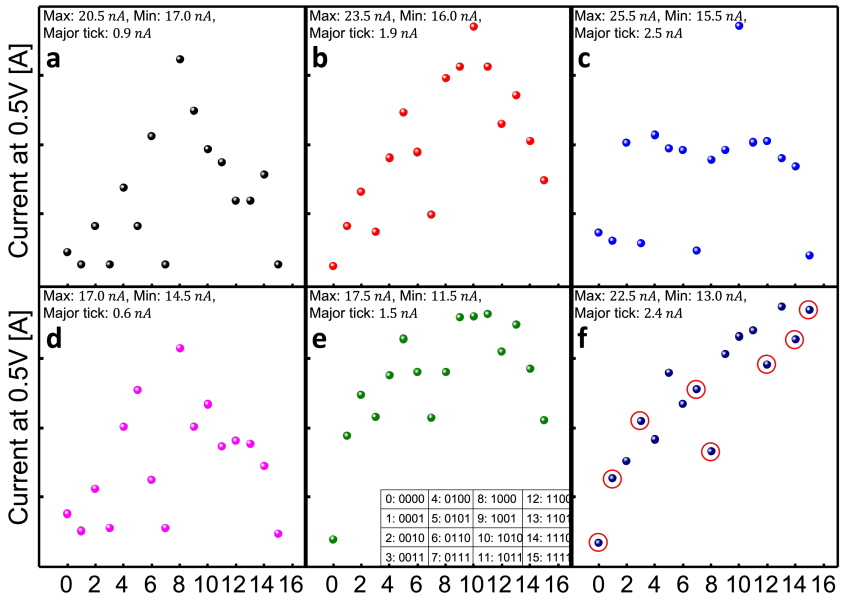


Figure 2-9: Experiment results to analyze the effect of changing parameters on the kernel characteristics in the temporal kernel system. The read current at 0.5 V of the memristor for the pulse stream '0000'~'1111' corresponds to 0~15 in the inset table in **e**. **a,** The read current at 0.5V of the memristor for each input under the conditions of 1 MΩ R_L, 4 V signal pulse height, 100 μs width, 4 V REF pulse height, and 100 μs width. **b-e,** The read current at 0.5 V of the memristor for each input when R_L, pulse width, pulse height, and REF pulse height are changed respectively from the condition of **a**. The various parameter settings for each figure were summarized in Table I. The kernel responses for each input of the temporal kernel optimized for the MNIST recognition are shown in **f**. Responses

to inputs showing high prevalence in the dataset were well separated (marked by red circles).

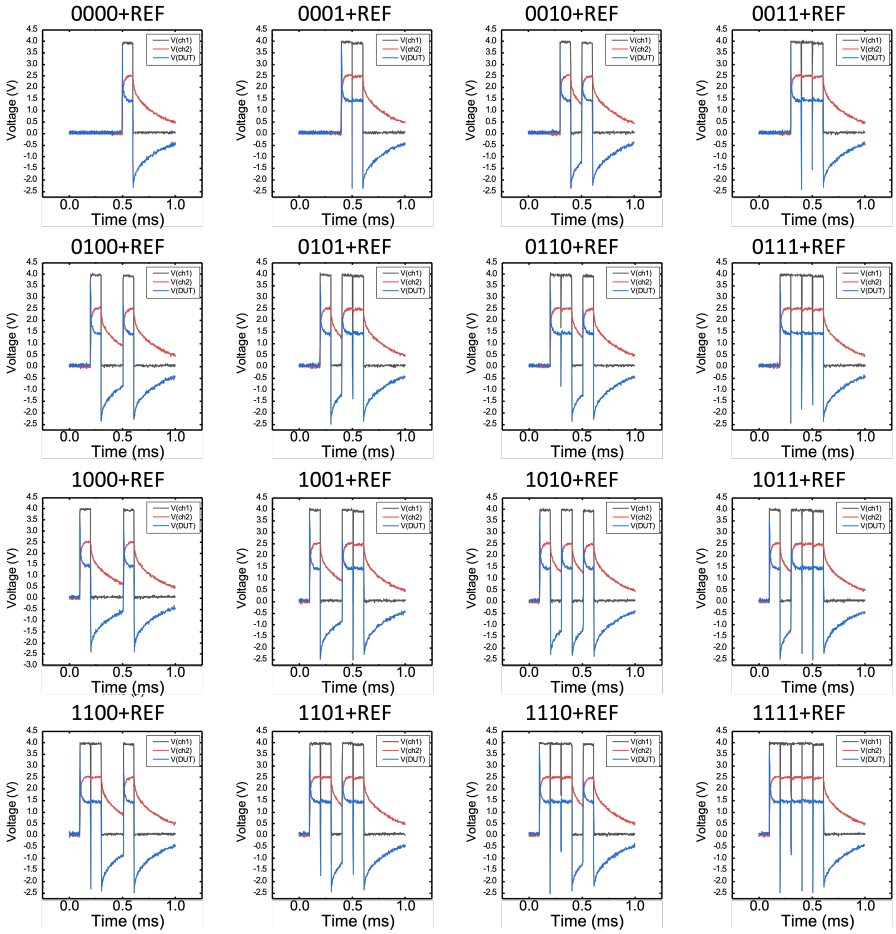


Figure 2-10: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height and a 100 μs REF pulse width.

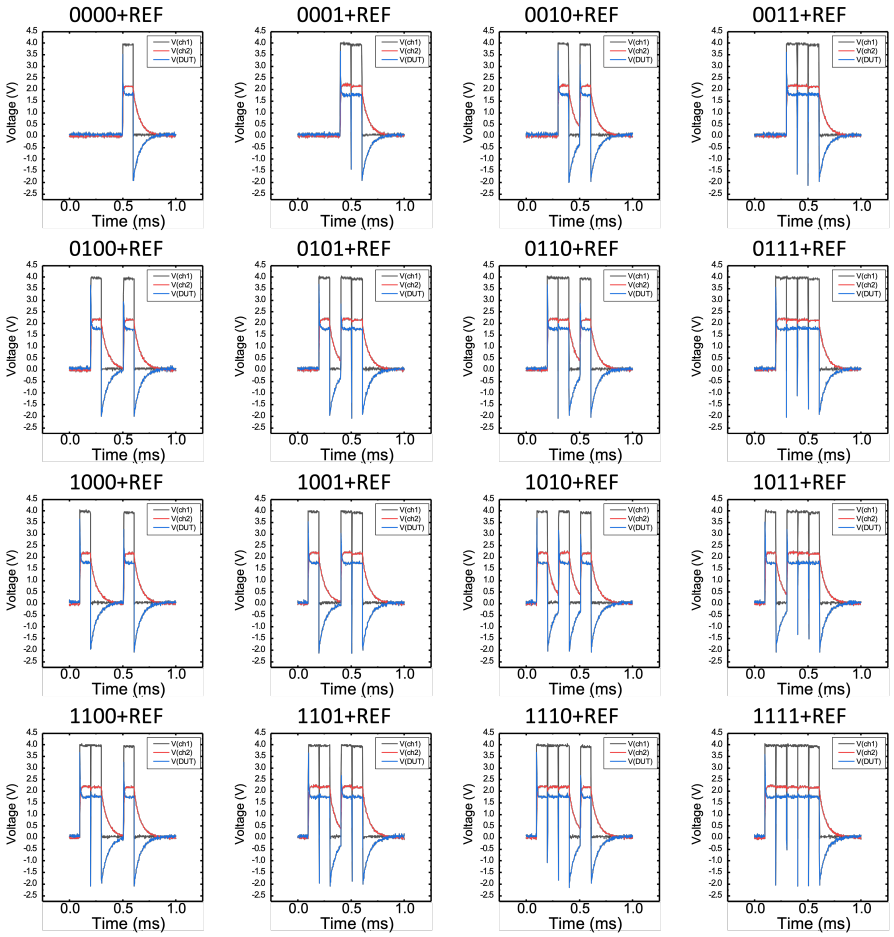


Figure 2-11: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 120 kΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height with a 100 μs width.

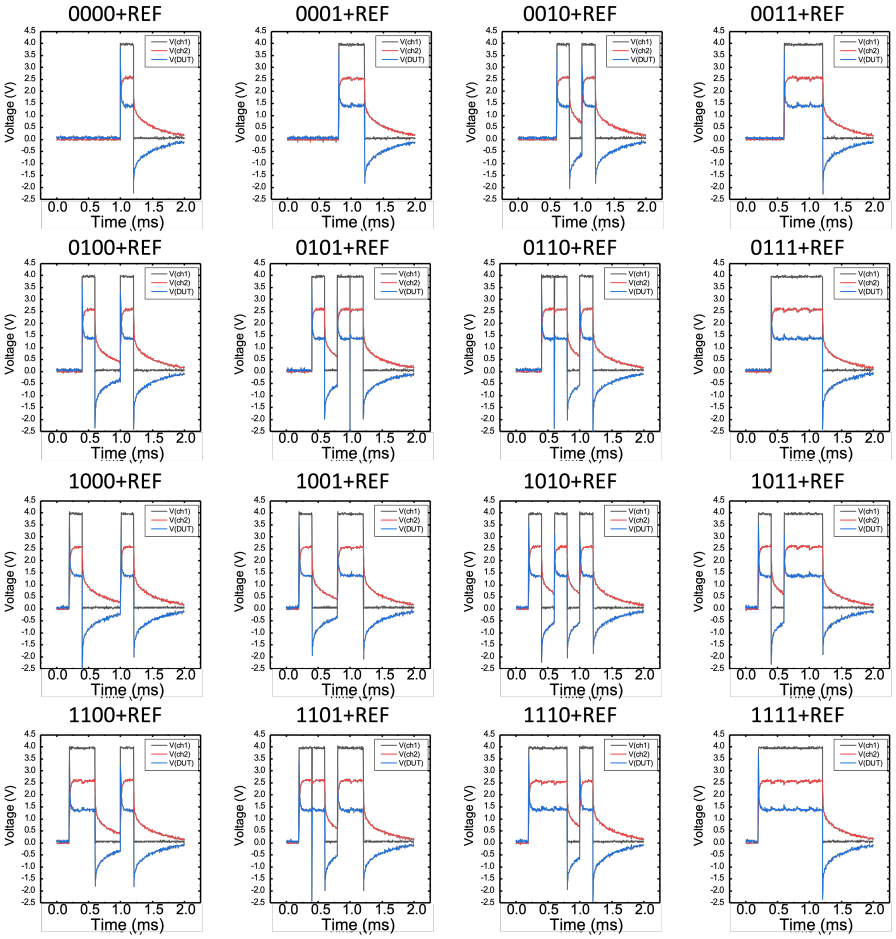


Figure 2-12: The V-t graphs for the '0000'~'1111' inputs under the conditions of 1 MΩ R_L with a 4 V signal pulse height and a 200 μs width, and a 4 V REF pulse height with a 200 μs width.

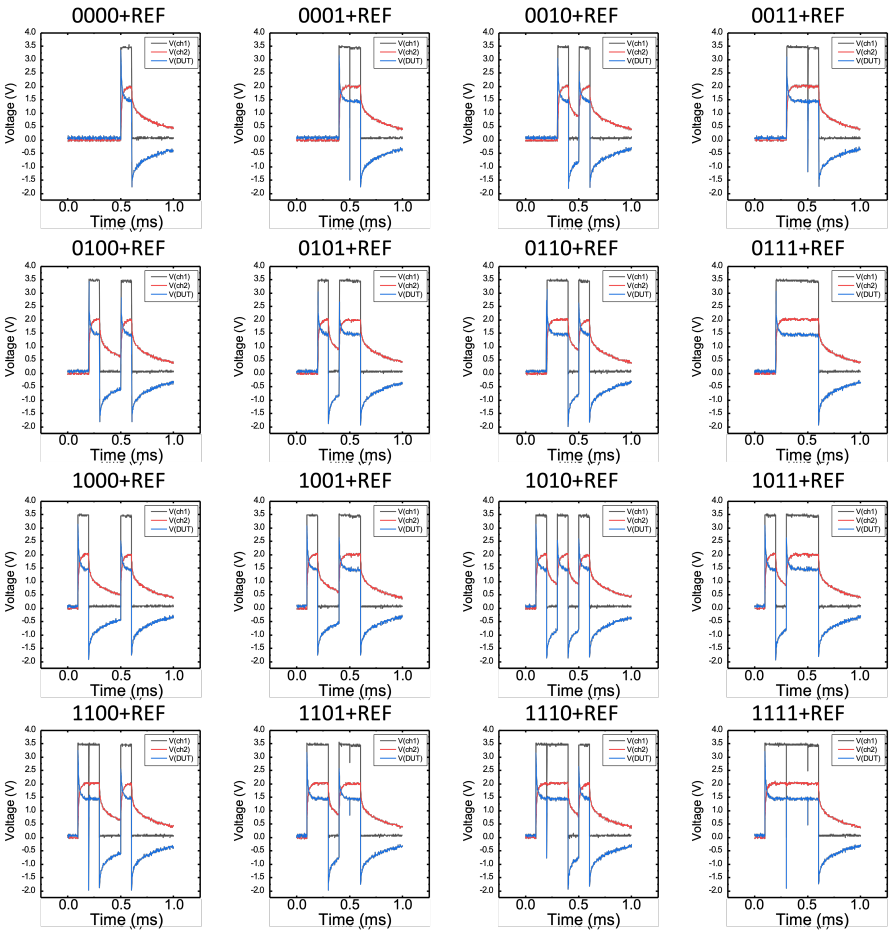


Figure 2-13: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 3.5 V signal pulse height and a 100 μs width, and a 3.5 V REF pulse height and a 100 μs width.

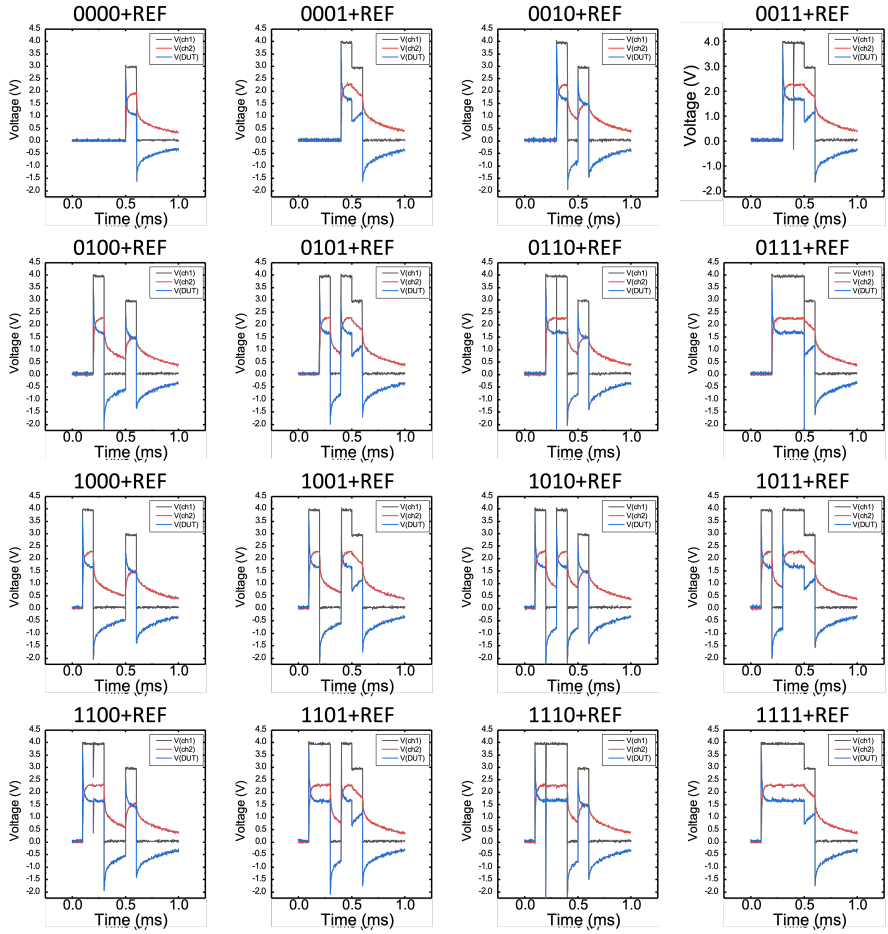


Figure 2-14: The V-t graphs for the '0000'~'1111' inputs under the conditions of a 1 MΩ R_L with a 4 V signal pulse height and a 100 μs width, and a 3 V REF pulse height and a 100 μs width.

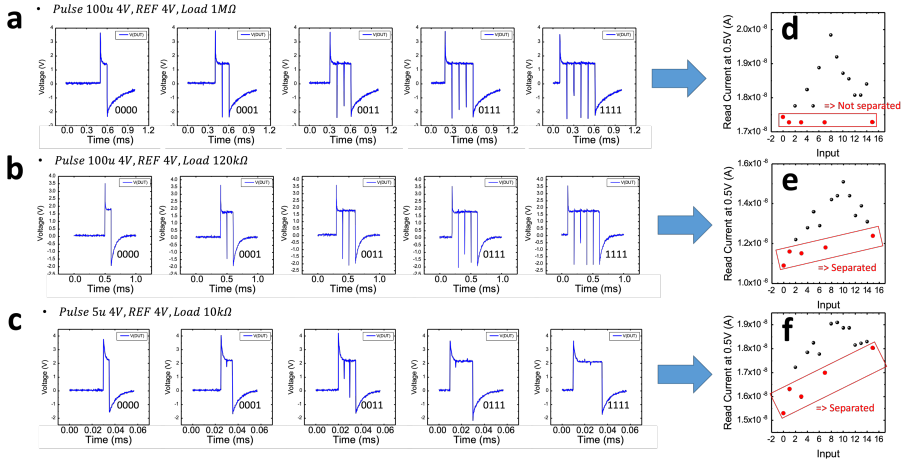


Figure 2-15: Analysis of the separation of inputs that generated net 1 spikes ('0000', '0001', '0011', '0111', and '1111'). From the conditions of a 4 V signal pulse height and a 100 μs width, and a 4 V REF pulse height and a 100 μs width, R_L varies from 1 MΩ to 10 kΩ. **a-c,** The V-t graphs for the inputs that generated net 1 spikes when 1 MΩ, 120 kΩ, and 10 kΩ R_L, were used. **d-f,** The read current of the memristor for the '0000~1111' inputs under the conditions in **a-c**. When the 1 MΩ R_L was used, since the voltage distributed to the memristor was small, SET switching did not occur after the first spike (**a**). Therefore, the responses to the inputs that generated net one spike (0, 1, 3, 7, and 15) were not separated (**d**). As R_L decreased, the voltage distributed to the memristor increased (**b-c**), and thus, the responses to the corresponding inputs were separated (**e and f**).

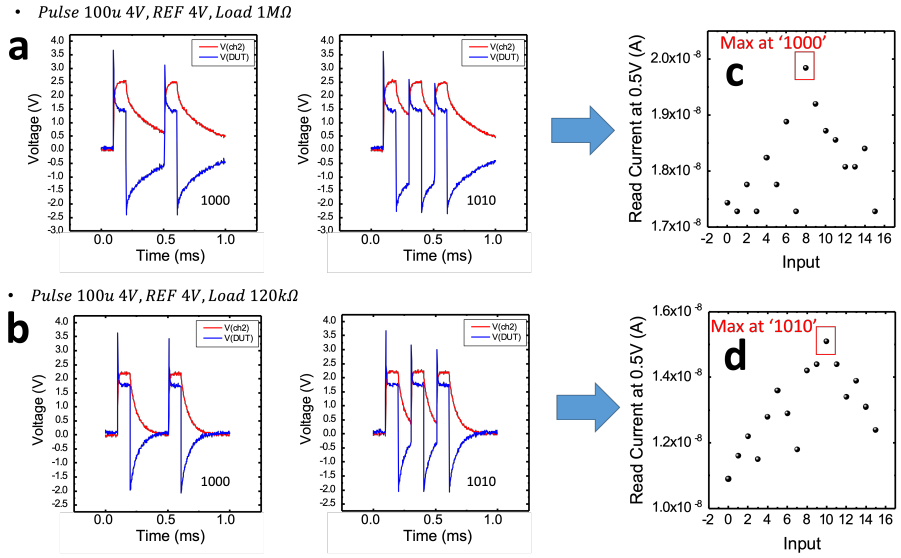


Figure 2-16: Analysis of the input that caused maximum conductance. **a-b,** The V-t graphs for the '1000' and '1010' inputs under the conditions of a 4 V signal pulse height and a 100 μs width. REF pulse has 4 V height and 100 μs width. 1 MΩ and 120 kΩ R_L were used for **a** and **b**. **c-d,** The read current of the memristor for the '0000~1111' inputs under the conditions in **a-b**. Since the large R_L caused slow discharging, a sufficient interval after the first spike is necessary to generate a spike that can cause large SET switching. Under the conditions in **a**, maximum conductance occurred at the '1000' input due to the slow discharging by the 1MΩ R_L (**c**). On the other hand, under the conditions in **b**, second and third spikes of sufficient magnitude to cause SET switching occurred at the '1010' input due to the fast discharging by the 120 kΩ R_L. Therefore, maximum conductance occurred at the '1010' input (**d**).

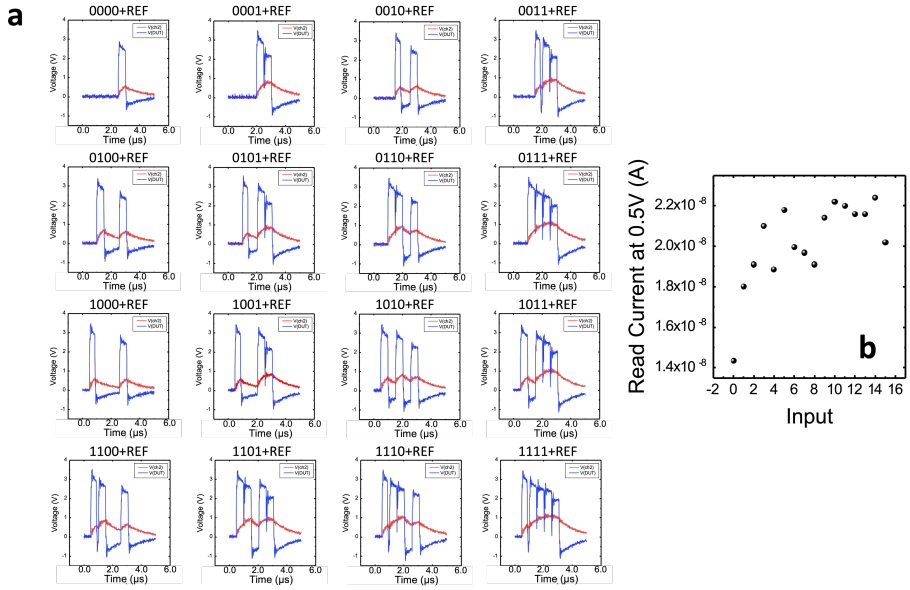


Figure 2-17: a, The V-t graphs for the '0000'~'1111' inputs under the conditions of 10 kΩ R_L with a 3.5 V signal pulse height and a 500 ns width, and a 3 V REF pulse height and a 500 ns width. **b,** The read current of the memristor for the '0000'~'1111' inputs. Insufficient charging further increased the separability for the consecutive high signals since the capacitor was not fully charged even though consecutive high signals were applied. This is suitable for situations in which consecutive signals mainly appear, such as in MNIST.

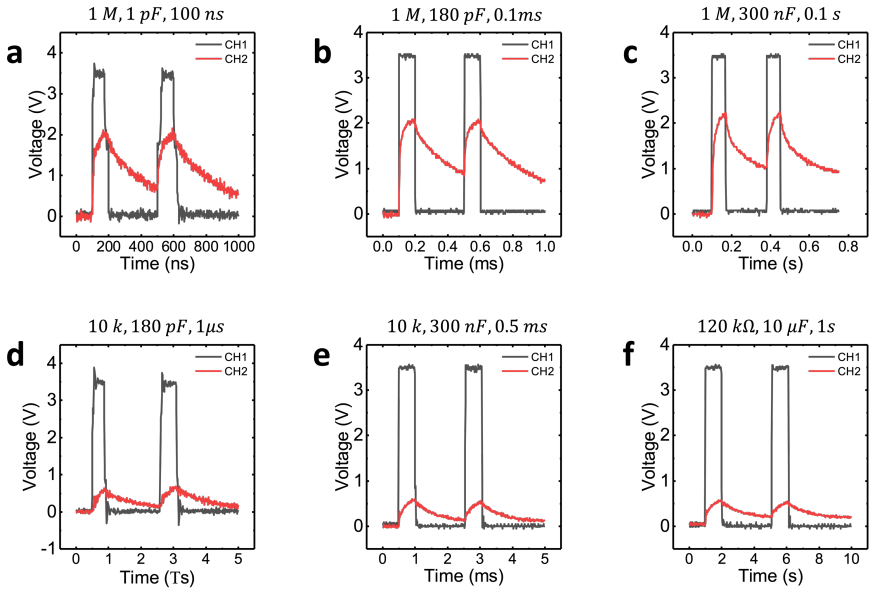


Figure 2-18: Temporal kernels with different time constants (100 ns ~ 1 s). Load resistance, parallel capacitance, and input interval used in each temporal kernel are indicated in each figure. **a-e,** The V-t graphs for the '1000' input under the condition of a 3.5 V signal pulse height. **a-c** represents a temporal kernel with similar characteristics to the temporal kernel in Figure 2-9a of main text, but with a different time constant. **d-f** represents a temporal kernel with similar characteristics to the temporal kernel in Figure 2-9f of main text, but with a different time constant.

TK Condition	R_L	Signal Pulse	REF Pulse
a	$1\ M\Omega$	$4\ V, 100\ \mu s$	$4\ V, 100\ \mu s$
b	$120\ k\Omega$	$4\ V, 100\ \mu s$	$4\ V, 100\ \mu s$
c	$1\ M\Omega$	$4\ V, 200\ \mu s$	$4\ V, 100\ \mu s$
d	$1\ M\Omega$	$3.5\ V, 100\ \mu s$	$3.5\ V, 100\ \mu s$
e	$1\ M\Omega$	$4\ V, 100\ \mu s$	$3\ V, 100\ \mu s$
f	$10\ k\Omega$	$3.5\ V, 200\ ns$	$3\ V, 200\ ns$

Table 2-1: The temporal kernel conditions (R_L, signal pulse, and REF pulse) used in Figure 2-9a-e

Task Optimization: MNIST. To perform the task of recognizing digit images in the Modified National Institute of Standards and Technology (MNIST) Database[33], the kernel dynamics were optimized to implement a TK system suitable for the task. To do this, the raw MNIST data set, composed of 784 pixels (28 x 28), had to be reconfigured to meet the requirement of this specific TK system, which is basically a binary system (0 and 1 inputs). Therefore, the data in the 784 pixel images were binarized and chopped by 4 bits, which resulted in 196 4-bit input signals. To make the task analysis more efficient, the frequency of the appearance of inputs in the dataset was investigated, and it was confirmed that '0000' appeared most frequently, followed by '1111,' '1000,' '0011,' and '0001' (Table 2-2). Therefore, in this task-optimized TK system, the task was performed effectively by setting the operation parameters so that the TK system could readily separate the responses to the inputs with a high frequency of appearance rather than separating the responses to all the 16 inputs. The data points indicated by the red circle in Figure 2-9f correspond to these frequently appearing signal sets. Accordingly, the 196 4-bit input image data were converted to the 196-membered MCV, where the measurements were performed on a single 1M1R1C circuit, based on Figure 2-9f. Using the 50,000 training images in the MNIST data set, 50,000 training MCVs were generated. These MCVs were used to train the 196 x 10 FCN (weights and biases), which were generated in a PyTorch simulation. The logistic regression algorithm was used to train the readout layer for the MNIST recognition and breast lesion

classification. The TK state (**x**) in the form of an n × 1 vector (n = 784 ~ 112 for the MNIST recognition and n = 510 for the breast lesion classification) was multiplied by the weight matrix (**W**) of the readout layer to yield the weighted sum (**z**).

$$\mathbf{z} = \mathbf{W}^T \cdot \mathbf{x} \quad (1)$$

The weighted sum was applied to the following softmax function to yield an output ($\hat{\mathbf{y}}$).

$$\hat{\mathbf{y}}_j = \sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{n} e^{z_k}} \; for\; j = 1, \ldots, n. \quad (2)$$

The sum of the elements of the output vector became 1 and the output of the softmax function was perceived as a 'probability.' The cross-entropy loss was used for the loss function, which is defined as:

$$loss = -\frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{y}_i log(\hat{\mathbf{y}}_i) + (1 - \boldsymbol{y}_i) log(1 - \hat{\mathbf{y}}_i)] \;, \quad (3)$$

wherein N is the number of samples, and $\boldsymbol{y}_i$ is the target output for input $\boldsymbol{x}_i$. To minimize the loss, a gradient-descent-based Adam optimizer[34] was identically used for the readout layer and 784 × 10 FCN. Full-batch-type learning of the readout layer and 784 × 10 FCN was performed in PyTorch. The trained TK system was used to infer the 10,000 MNIST test images, and the achieved accuracy was 90.1 % (see Table 2-3 and Table 2-4, 2-5 for the results of combining various kernels and the results of considering cycle-to-cycle and cell-to-cell variations). When one hidden layer composed of 200 neurons is added to the FCN, the accuracy was increased to 96.5 %.

This kernel machine took 200 ns of time and ~25 pJ of energy (Figure

2- 19) to process one input pulse, which is $10^3 \sim 10^4$ times shorter and 100 ~ 400 times lower than in the previous studies.[20]-[22] Table 2-3 shows the comparison with other RC results using the diffusive memristors and the software-based single-layer FCN. This study focuses on the only memristive TK system that performs kernel adaptation and that showed the best performance in terms of accuracy and latency. Table 2-6 shows the results for the case where the 2-layer FCN is used as the readout layer, and when 196×38×10 FCN is used, it offers 95.1 % accuracy. The number of training parameters in this network (7,828) is slightly smaller than that of the software-based FCN (7,840). The readout network size of the TK system could be further decreased as the number of bits processed by the kernel (nBPK) increases, for as long as the separability for the higher nBPK is guaranteed. Figure 2-20 shows the different read currents for the 3 to 6 bits (8 to 64 input patterns). Obviously, the separability decayed as the nBPK increased, but they were still be used to recognize the MNIST data set because not all the input patterns mattered equally. Table 2-7 shows the variation in the test accuracy of the MNIST data set using the same method as above, but with different nBPKs. As the nBPK increased from 3 to 6, which was accompanied by a decrease in the required memristor number from 252 to 112, the accuracy decreased from 90.7% to 86.3% (the confusion matrices are included in Figure 2-21), which is not much lower than in the software-based FCN (784 x 10). The next section demonstrates the most crucial merit of this TK system by showing its capacity to process time-

series data using medical diagnostic data.

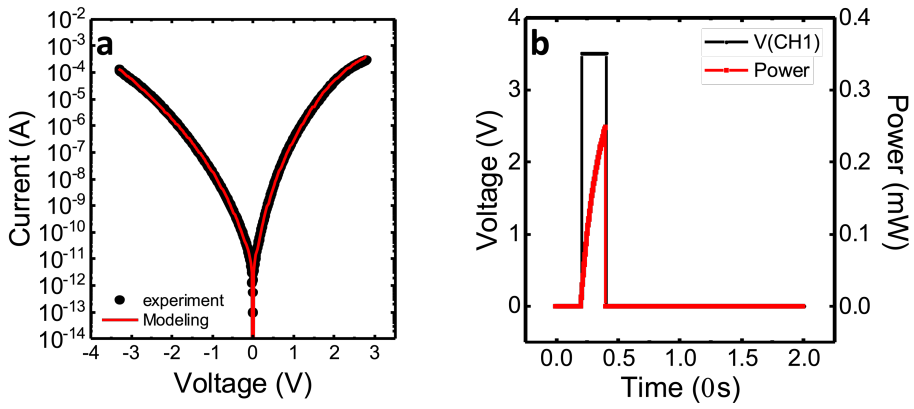


Figure 2-19: Result of the I-V curve fitting for the WHT memristor and power consumption in the 1M1R1C kernel machine during processing one input. **a**, I-V curve fitting of the WHT memristor (HRS state) based on the conduction mechanisms. **b**, Power consumption in the 1M1R1C kernel machine (Figure 2-9f kernel condition) during input processing. Since the resistance of the WHT memristor is dependent on the voltage, the current passing through the memristor was obtained with the HSPICE simulation using the result of the I-V curve fitting in **a**. The energy ($\int_t Power(t) \cdot dt$) consumed to process one input was calculated as ~25 pJ.

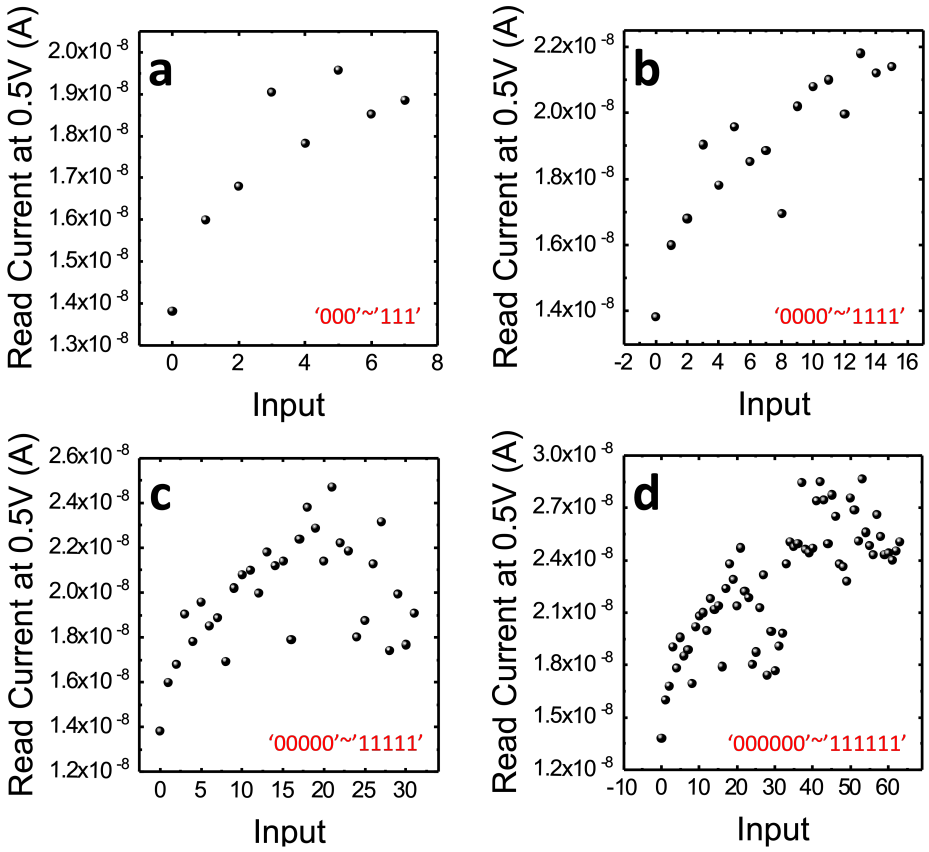


Figure 2-20: The temporal kernel responses were measured while increasing the number of bits processed in the temporal kernel from 3 bits to 6 bits. **a-d,** The temporal kernel responses for the '000~111', '0000~1111', '00000~11111', and '000000~111111' inputs under the conditions of 10 kΩ R_L with a 3.5 V signal pulse height and a 200 ns width, and a 3 V REF pulse height and a 200 ns width. As the number of bits processed in the temporal kernel increased, the separation of the responses to each input deteriorated.

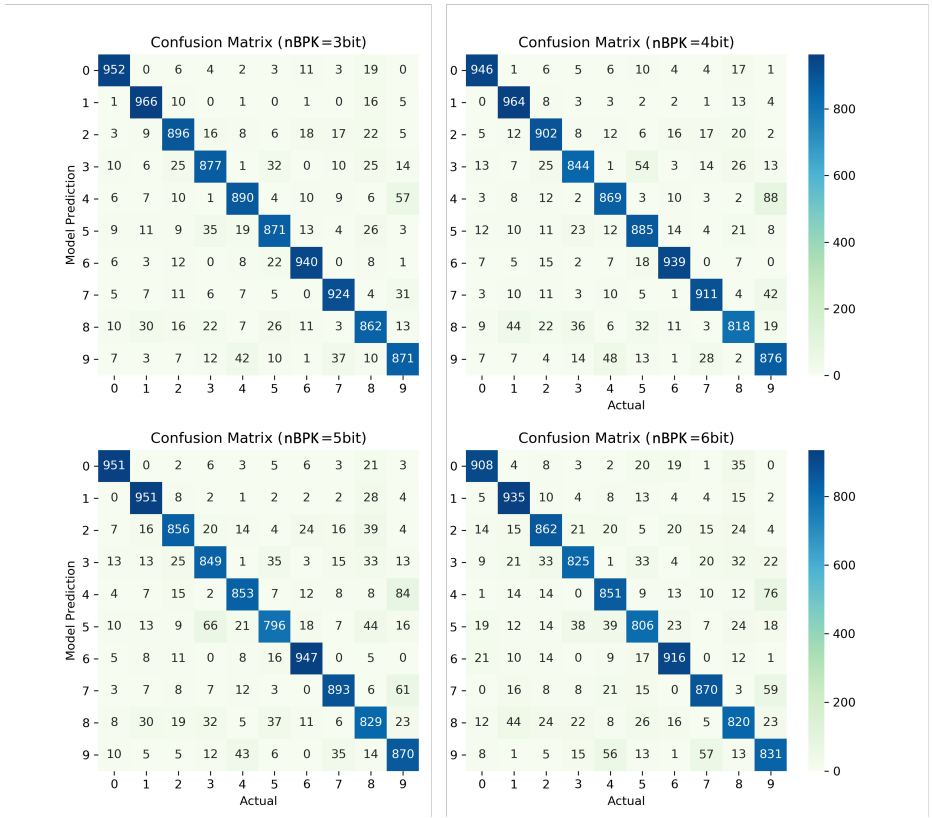


Figure 2-21: The confusion matrices comparing the recognized digit and the desired digit for the MNIST test dataset (4 situations, from the top left: nBPK = 3 bits to the bottom right: nBPK = 6 bits) showing that the number of correct inferences decreased as the nBPK increased.

Inputs	# of Inputs	Percentage [%]
0000	9,078,931	77.2017
0001	355,307	3.0213
0010	24,104	0.2049
0011	356,362	3.0302
0100	23,171	0.1970
0101	806	0.0068
0110	105,121	0.8938
0111	276,838	2.3540
1000	364,691	3.1011
1001	11,919	0.1013
1010	959	0.0081
1011	8,580	0.0729
1100	346,940	2.9501
1101	9,044	0.0769
1110	262,350	2.2308
1111	534,877	4.5482
Total	11,760,000	100

Table 2-2: The frequency of the appearance of inputs in the preprocessed MNIST dataset, in which '0000' appeared overwhelmingly, followed by the inputs '1111', '1000', '0011', '0001', '1100', '0111' and '1110' in the table (Due to the nature of the picture, the pixels were continuously blanked or filled in most cases. Therefore, inputs with consecutive high or low signals mainly appeared, and there were a few inputs with alternating high and low signals such as '1010' and '0101'.)

Group	Accuracy	Latency in kernel	Kernel Adaptation	Network Size	Image Size	Etc.
This Study	90 % (95.1% - two layer)	1 μs	O	196x10 (196x38x10)	28x28	
Wei. D. Lu	85 %	10 $\mathrm{m}s$	X	88x10	22x20	14,000/2,000 Training/Test set
Joshua Yang	83 %	1 $\mathrm{m}s$	X	220x10	22x20	In situ training
Software (784x10 FCN)	91 %	-	-	784x10	28x28	

Table 2-3: Comparison of the results of the MNIST recognition using memristive temporal kernel computing systems[21],[20] and a software-based system[1] (single-layer FCN), showing very fast processing and the highest accuracy in this study

Kernel	Accuracy	Input vector	Readout Layer
a + b	90.7%	196x10	392x10
a + f	91.8%	196x10	392x10
b + f	91.7%	196x10	392x10
a + b + f	91.7%	196x10	588x10
Kernel	Accuracy	Input vector	Readout Layer
200 ns + 2 us	91.4%	196x10	392x10
200 ns + 5 us	91.6%	196x10	392x10
2 us + 5 us	91.5%	196x10	392x10
200 ns + 2 us + 5 us	92.4%	196x10	588x10

Table 2-4: Results of MNIST recognition using various kernel combinations. For the recognition, kernel conditions of Figure 2-9a, b, and f of main text were used. A combination of 'Figure 2-9a+Figure 2-9f' showed an accuracy of 91.8%. For a 196x10 input vector, two kernels processed the input, and a 392x10 readout layer was used (588x10 readout for the 3 kernels). On the other hand, when the pulse width was modified without changing the conditions R_L, C, and pulse height in the condition of Figure 2-9f, an accuracy of 92.4 % was obtained in the combination of '200ns+2μs+5μs'. By combining various kernels or changing pulse conditions for the same kernel machine, the imperfections of one kernel could be compensated for by another kernel, and the accuracy could be improved.

Variation	Accuracy
no variation	90.1%
cycle to cycle	89.7%
cell to cell	90.0%
cycle to cycle + cell to cell	89.6%

Table 2-5: The accuracy when cycle to cycle variation, cell to cell variation, and both are considered (kernel condition of Figure 2-9f of main text was used). Each variation was calculated based on variation measurement results in Figure 2-1. Up to 1 sigma of each variation was considered, and when both cycle to cycle and cell to cell varaition were included in the simulation, the accuracy decreased by 0.5 %.

Readout layer	Training parameters	nBPK	Accuracy
196 x 38 x 10	7,828	4	95.1 %
196 x 50 x 10	10,300	4	95.5 %
196 x 110 x 10	22,660	4	96.1 %
196 x 200 x 10	41,200	4	96.5 %

Table 2-6: Results of MNIST recognition using two-layer FCN for the readout layer of TK system. The table shows the number of training parameters used in each two-layer FCN and the accuracy of the TK system (nBPK = 4). When 196x38x10 FCN was used, 7,828 training parameters were used, and the TK system accuracy was 95.1 %.

nBPK	Readout Layer Size	Accuracy
3	252x10	90.7%
4	196x10	90.1%
5	140x10	88.1%
6	112x10	86.3%

Table 2-7: Results of the MNIST recognition while increasing the number of bits processed in the temporal kernel, showing that as nBPK increased, both the size of the used readout layer and the recognition accuracy decreased.

Task Optimization: Medical Diagnosis. Medical diagnosis often requires analyzing time-varying data and making a quick diagnosis, but there are inevitable limitations such as high dependence on operators and high variability across different medical institutions. For a more accurate and objective medical diagnosis, a universal diagnosis system adaptable to various situations is essential. Automatic medical diagnosis using deep learning has considerable potential, and several studies have been conducted on it,[35]-[37] but most of them rely on the conventional image classification method, such as CNN. This means that the traditional medical diagnosis produces data images and analyzes them later, mostly *ex-situ*. This study suggests a method for *in-situ* medical diagnosis in real-time using a 1M1R1C kernel. The diagnostic application consists of two sections. The first section is breast cancer diagnosis using ultrasound images, and the second section is arrhythmia diagnosis based on electrocardiogram (ECG) results. These two applications have vastly different operating signal frequencies (MHz to Hz). In this study, a system for efficient medical diagnosis was implemented by optimizing the TK system for each task.

1) Diagnosis of malignancy in breast lesions. Breast cancer is the most common cancer in women. Ultrasound is used to diagnose and monitor this disease. In contrast to the conventional CNN, where the preprocessed images are identified, the proposed TK system in this study directly uses ultrasonic raw data without an imaging process, as shown in Figure 2-22a. In the conventional ultrasound diagnosis, the ultrasound is transmitted to the piezoelectric material, where

electrical signals are generated. The signal processor processes these signals to generate an ultrasound image, which the operator analyzes to diagnose the disease. However, if the TK can directly process the ultrasonic signal, the imaging process can be skipped, and an automatic diagnosis will be made at the readout layer. Therefore, this system makes real-time diagnosis simpler than in the existing ultrasound diagnosis.

The dataset used in the experiment consisted of an open-access database of raw ultrasound signals acquired from malignant and benign breast lesions.[38] Each sample consisted of 510 ultrasound (10 MHz) echo lines. After they were preprocessed for measurement convenience, they were converted into pulse streams and applied to the memristor (Methods section). Figure 2-22b shows the results of the voltage-time (V-t) measurement for one echo line of a benign sample (inset in Figure 2-22b). The test set consisted of 36 samples randomly extracted out of the total 100 samples, and the training set consisted of the remaining 64 samples. The readout was performed by repeating the process of randomly extracting the test set from the entire dataset 30 times, and an average accuracy of 94.6% was obtained.

This method has two main advantages over the existing ultrasound diagnosis using CNN. First, diagnosis is performed using a much simpler system without a pre-imaging process. Second, one of the major difficulties in ultrasound analysis is the presence of artifacts.[35] CNN may have difficulty in recognizing such artifacts because it performs learning and inference with the

information on the artifacts. Using 1M1R1C, even with additional stimulation by artifacts, the capacitor only maintains the charging state. Therefore, the kernel state is determined by the overall contour rather than by fine artifacts, and it can show higher performance.

2) Real-time arrhythmia diagnosis. Arrhythmia is a condition in which the heart has an irregular rhythm or an abnormal heart rate. Since malignant arrhythmia can cause sudden death due to a heart attack,[39] real-time ECG monitoring and diagnosis are required. The purpose of this experiment is to implement a system capable of real-time diagnosis of arrhythmia in response to an electrical signal caused by a heartbeat. For the experiment, a part of the MIT-BIH arrhythmia database[40] was used, and a task-optimized kernel was utilized to distinguish between arrhythmia and normal cases. A TK capable of responding to a signal with a frequency of 0.8 to 1.2 Hz was constructed using a 1 μF capacitor parallel to CH2. In this case, a simple temporal kernel machine composed of only one 1M1R1C kernel could be used. Figure 2-22c shows a part of the ECG of a patient with arrhythmia. The electrical signal is generated at approximately 0.8-s intervals, and then arrhythmia occurs at 1.6 s (marked by a red arrow). When an electrical signal from a heartbeat is applied to the kernel machine, the capacitor maintains a high charging level at a normal beat. When an arrhythmia occurs, the capacitor is discharged at a longer interval than in the normal case, and SET switching occurs in the memristor by the next pulse (Figure 2-23). Since this kernel responds only to arrhythmia, the memristor conductance can

reflect the pulse of the arrhythmia patient in real-time. Figure 2-22d shows the results of 5-minute TK monitoring based on ECG data of normal (case 1) and arrhythmic (cases 2 and 3) patients. In cases 2 and 3, 49 and 81 arrhythmias occurred, respectively. As a result, the conductance of the TK monitoring in case 3 was the highest, and the memristor conductance was clearly distinguished according to the degree of arrhythmia. This single TK system was able to detect different arrhythmia conditions in real-time with low energy using a simple 1M1R1C circuit.

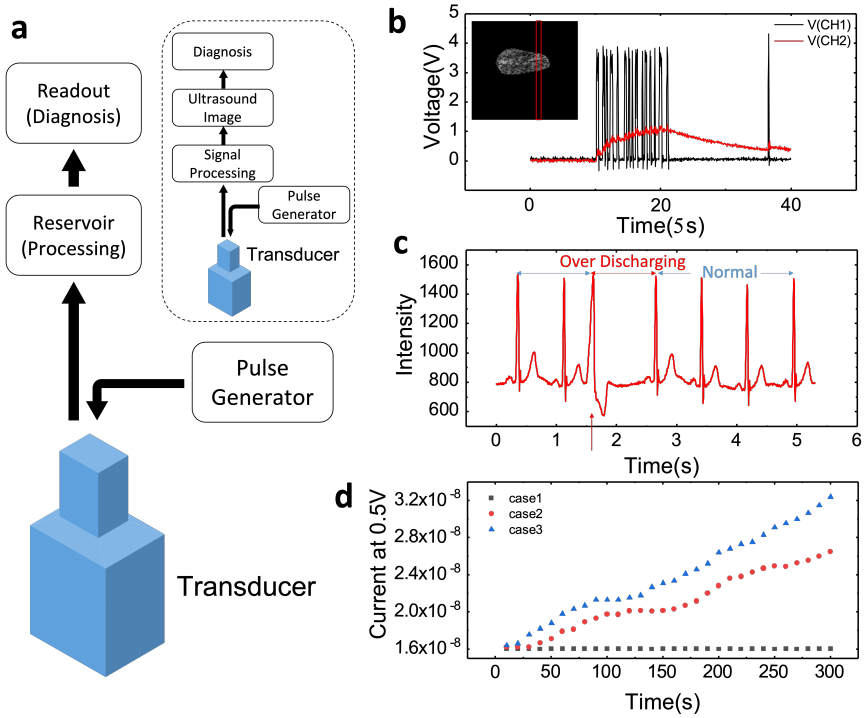


Figure 2-22: The automatic medical diagnosis system using the 1M1R1C temporal kernel and the experiment results in the two sections. **a,** A system for diagnosing the malignancy of breast lesions, which is much simpler than in the existing method (inset in **a**). In this system, ultrasonic signals are applied directly to the kernel machine, so the imaging step is omitted. **b,** V-t graph for one echo line of a benign sample (inset in Figure 2-22b). **c,** A part of the electrocardiogram of a patient with arrhythmia. Long intervals caused by abnormal beats discharged the capacitor, and the conductance of the memristor increased in the next pulse. **d,** Five-minute temporal kernel monitoring based on the ECG of one normal patient (case 1) and two arrhythmic patients (cases 2 and

3). When arrhythmia occurred, the conductance of the memristor increased. Case 3, which had the most severe arrhythmia symptoms, showed the highest conductance.

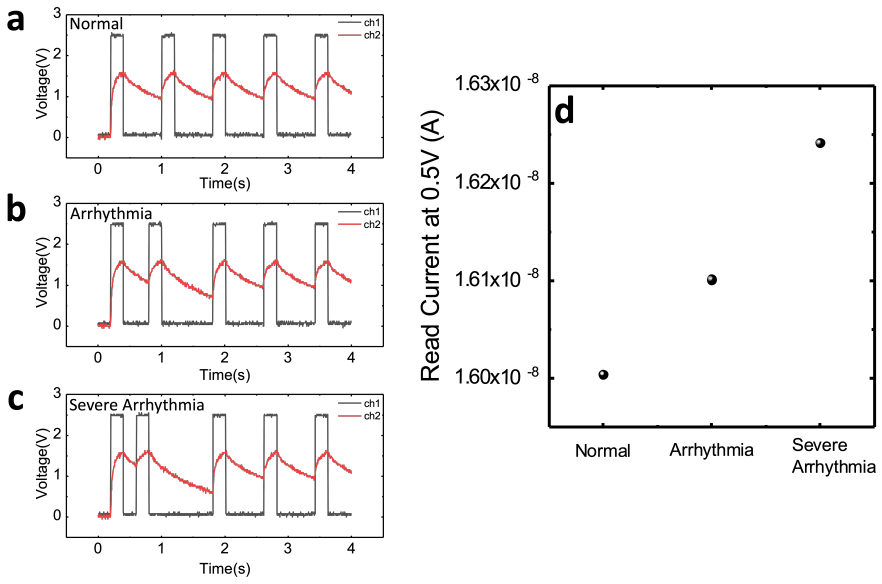


Figure 2-23: The increase in the conductance of the memristor varied according to the degree of arrhythmia. When arrhythmia was severe, SET switching occurred in the memristor due to long discharging. **a-c,** The ECG-based V-t graphs for three cases of normal, arrhythmia, and severe arrhythmia. The electrical signal of the ECG from the heartbeat was converted into a 2.5 V, 200 ms pulse and applied to the memristor. **d,** The read current of the memristor according to the degree of arrhythmia. The more severe the arrhythmia was, the more the memristor conductance increased.

2.4. Conclusion

In this study, an TK system with high kernel separability and dynamics controllability was demonstrated using a W/HfO_2/TiN memristor. A dynamic kernel was generated by composing a 1M1R1C circuit. From asymmetric charging/discharging of the capacitor caused by the memristor, separability, which is the basic property of the TK, was achieved. In addition, the manner in which the kernel reacted to the input signal was modified by changing various parameters such as the load resistor, capacitance, pulse width, and pulse height. Using these characteristics, the TK system was optimized to perform static data-based MNIST recognition applications and sequential data-based medical diagnoses (ultrasound diagnosis and ECG-based diagnosis). For the MNIST recognition, a task-optimized system was used to improve the separability of the inputs that frequently appeared in the dataset. Furthermore, the tradeoff between the reduction of the readout layer size and the performance was confirmed by increasing the nBPK. TK system-aided diagnosis was conducted for two situations with contrasting input frequencies (1 Hz and 10 MHz). By implementing a kernel configuration suitable for each task (kernel adaptation), the excellent performance was achieved. In particular, the most crucial point of this study is its demonstration that dynamic signals with vastly different time constants can be well distinguished by changing the resistor or capacitor added to the circuit using only one type of memristor.

The two types of hardware needed to implement the 1M1R1C TK system and analysis on the area/cell are shown in Figure 2-24. In both cases, using a metal-insulator-semiconductor capacitor, the capacitance can be adjusted by modifying the R and pulse height (Figure 2-25). Therefore, it is expected that the fabrication of the hardware for the array configuration will be simple and that the TK dynamics can easily be changed even in the fabricated hardware.

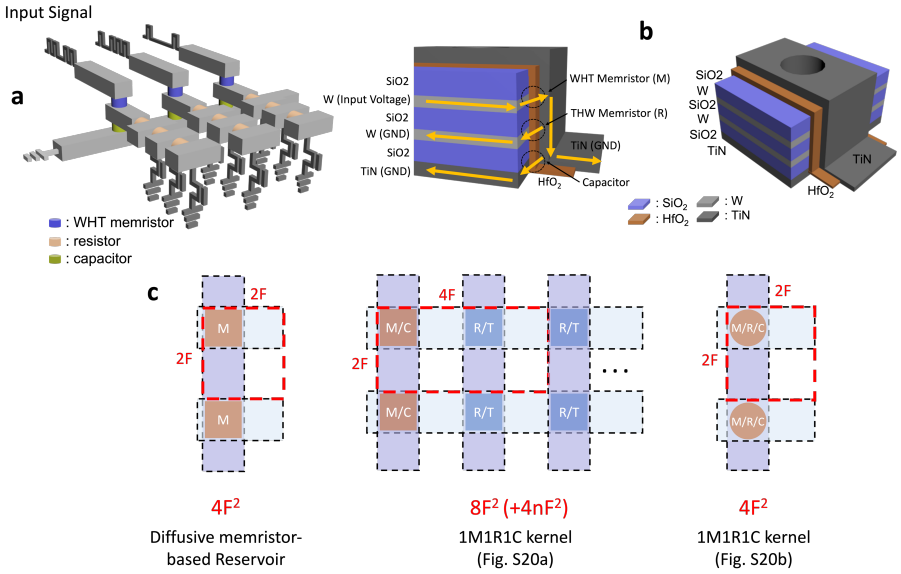


Figure 2-24: The hardware structure needed to create an array of temporal kernels that can adjust the kernel configuration. **a,** A structure in which the resistors are sequentially connected to several metal lines. In this structure, the resistance value of the temporal kernel can be adjusted by selecting several metal lines connected to the resistors. **b,** A structure in which memristors are connected in series to the WHT memristors and parallel to the capacitors (**b left panel**). The 1M1R1C circuit can be implemented in a three-dimensional structure by stacking TiN, W metals in multi-layers and depositing a dielectric layer and top electrode in the hole after hole etching (**b right panel**). In this structure, the resistance of the memristor can be set to the desired resistance value using a method such as the incremental step pulse program (ISPP). **c**, Cell

area of the diffusive memristor-based reservoir and 1M1R1C kernel. The diffusive memristor-based reservoir is implemented using a passive array composed of memristors. Therefore, $4F^2$ is required per cell (**c left panel**). If the 1M1R1C kernel is implemented with the structure in **a**, a minimum area of $8F^2$ is required per cell when using a vertical pillar transistor (T), and the area increases by $4F^2$ each time a serial resistor is added (**c middle panel**). The structure proposed in **b** requires an area of $4F^2$/cell (**c right panel**). This structure does not require an increase of area/cell even with additional elements (R_L, C) other than the memristor through a 3D integration process.

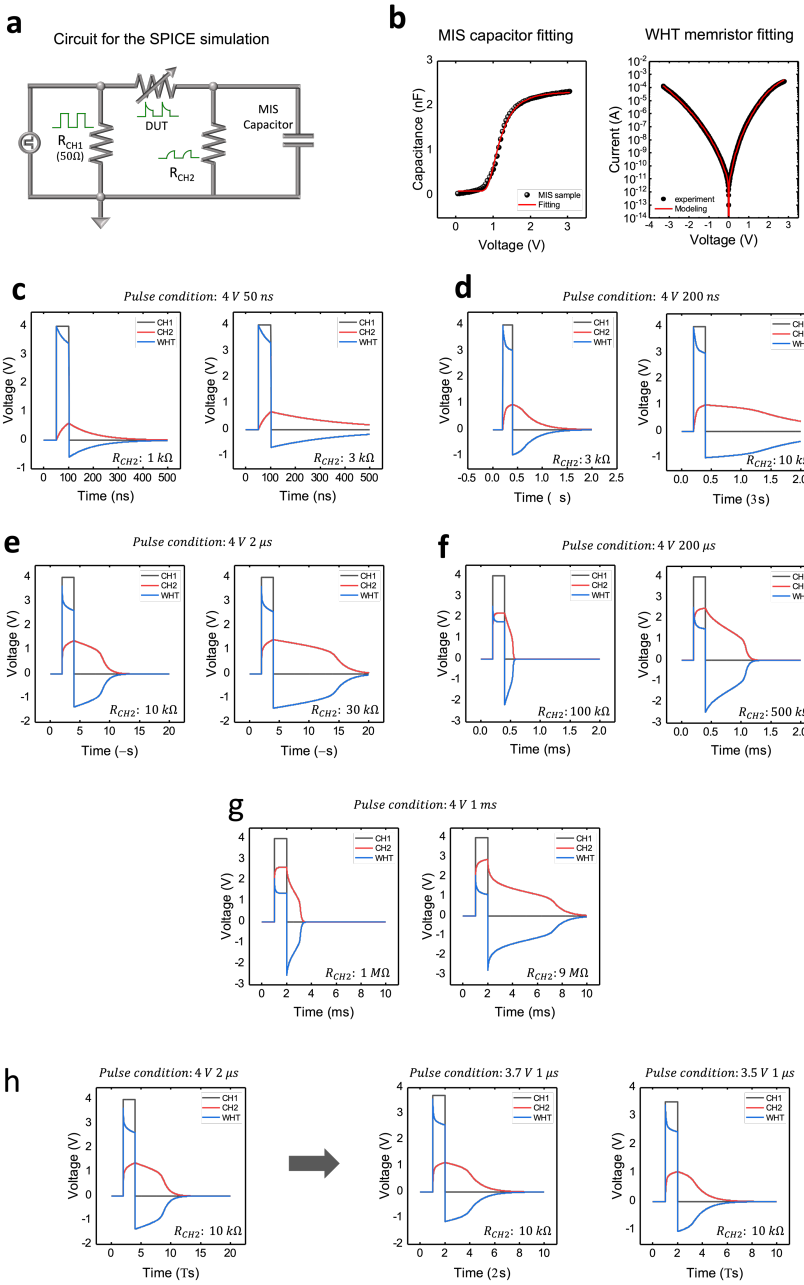


Figure 2-25: Implementation of various time constants of 1M1R1C kernel using MIS capacitor and WHT memristor (HSPICE simulation). **a,** 1M1R1C circuit used in SPICE simulation. **b**, C-V curve of MIS capacitor and I-V curve of WHT memristor used for

simulation and their fitting results (red line). In the simulation, an MIS capacitor (sample device) showing a capacitance of 100 pF ~ 2.2 nF was used, and the WHT memristor fitting result of Figure 2-19 was used. **c-g**, V-t graphs of the kernel showing fast discharging (left panel) and slow discharging (right panel) characteristics (pulse width 50 ns ~ 1 ms). **h**, V-t graphs in the kernel condition of 4 ~ 3.5 V pulse height, 2 ~ 1 μs pulse width, and 10 kΩ R_{CH2}. **h** shows the effect of changing pulse height on the capacitance of the TK system.

2.5. References

[1] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323 (1998).

[2] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* (2017) doi:10.1145/3065386.

[3] Dong, C., Loy, C. C., He, K. & Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016) doi:10.1109/TPAMI.2015.2439281.

[4] Wang, Z., Yan, W. & Oates, T. Time series classification from scratch with deep neural networks: A strong baseline. in *Proceedings of the International Joint Conference on Neural Networks* (2017). doi:10.1109/IJCNN.2017.7966039.

[5] Hochreiter, S. Long Short-Term Memory. **1780**, 1735–1780 (1997).

[6] Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertainty, Fuzziness Knowlege-Based Syst.* **6**, 107–116 (1998).

[7] Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[8] Kim, G. H. *et al.* 32 × 32 Crossbar Array Resistive Memory Composed

of a Stacked Schottky Diode and Unipolar Resistive Memory. *Adv. Funct. Mater.* **23**, 1440–1449 (2013).

[9] Jeong, D. S. & Hwang, C. S. Nonvolatile Memory Materials for Neuromorphic Intelligent Machines. *Adv. Mater.* **30**, 1–27 (2018).

[10] Kim, K. M. *et al.* Low-Power, Self-Rectifying, and Forming-Free Memristor with an Asymmetric Programing Voltage for a High-Density Crossbar Application. *Nano Lett.* **16**, 6724–6732 (2016).

[11] Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).

[12] Lee, Y. K. *et al.* Matrix mapping on crossbar memory arrays with resistive interconnects and its use in in-memory compression of biosignals. *Micromachines* (2019) doi:10.3390/mi10050306.

[13] Kim, Y. *et al.* Novel Selector-Induced Current-Limiting Effect through Asymmetry Control for High-Density One-Selector–One-Resistor Crossbar Arrays. *Adv. Electron. Mater.* **5**, 1–11 (2019).

[14] Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.* (2002) doi:10.1162/089976602760407955.

[15] Jaeger, H. The 'echo state' approach to analysing and training recurrent neural networks. *GMD Rep.* (2001).

[16] Lukoševičius, M. & Jaeger, H. Reservoir computing approaches to

recurrent neural network training. *Comput. Sci. Rev.* **3**, 127–149 (2009).

[17] Schrauwen, B., Verstraeten, D. & Van Campenhout, J. An overview of reservoir computing: Theory, applications and implementations. *ESANN 2007 Proc. - 15th Eur. Symp. Artif. Neural Networks* 471–482 (2007).

[18] Miller, J. & Broersma, H. *Computational Matter*. (2015). doi:10.1145/2739482.2764939.

[19] Fortune, E. S. & Rose, G. J. Short-term synaptic plasticity as a temporal filter. *Trends Neurosci.* **24**, 381–385 (2001).

[20] Midya, R. *et al.* Reservoir Computing Using Diffusive Memristors. *Adv. Intell. Syst.* **1**, 1900084 (2019).

[21] Du, C. *et al.* Reservoir computing using dynamic memristors for temporal information processing. *Nat. Commun.* **8**, 1–10 (2017).

[22] Moon, J. *et al.* Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2**, 480–487 (2019).

[23] Zhu, X., Wang, Q. & Lu, W. D. Memristor networks for real-time neural activity analysis. *Nat. Commun.* **11**, (2020).

[24] Kim, K. M. *et al.* A detailed understanding of the electronic bipolar resistance switching behavior in Pt/TiO2/Pt structure. *Nanotechnology* (2011) doi:10.1088/0957-4484/22/25/254010.

[25] Shao, X. L. *et al.* Electronic resistance switching in the Al/TiOx/Al structure for forming-free and area-scalable memory. *Nanoscale* **7**,

11063–11074 (2015).

[26] Lu, Y. *et al.* An electronic silicon-based memristor with a high switching uniformity. *Nat. Electron.* **2**, 66–74 (2019).

[27] Kwon, S. *et al.* Structurally engineered nanoporous Ta2O5-x selector-less memristor for high uniformity and low power consumption. *ACS Appl. Mater. Interfaces* **9**, 34015–34023 (2017).

[28] Kim, Y. *et al.* Nociceptive Memristor. *Adv. Mater.* **30**, 1–7 (2018).

[29] Ryu, J. J. *et al.* Fully 'Erase-free' Multi-Bit Operation in HfO 2 -Based Resistive Switching Device. *ACS Appl. Mater. Interfaces* **11**, 8234–8241 (2019).

[30] Ang, S. S. Titanium nitride films with high oxygen concentration. *J. Electron. Mater.* **17**, 95–100 (1988).

[31] Müller, E. W. Work function of tungsten single crystal planes measured by the field emission microscope. *J. Appl. Phys.* **26**, 732–737 (1955).

[32] Yoon, J. H. *et al.* Highly uniform, electroforming-free, and self-rectifying resistive memory in the Pt/Ta2O5/HfO2-x/TiN structure. *Adv. Funct. Mater.* **24**, 5086–5095 (2014).

[33] Lecun Yann, Cortes Corinna & Burges Christopher. THE MNIST DATABASE of Handwritten Digits. *Courant Inst. Math. Sci.* (1998).

[34] Konur, O. Adam Optimizer. *Energy Education Science and Technology Part B: Social and Educational Studies* (2013).

[35] Liu, S. *et al.* Deep Learning in Medical Ultrasound Analysis: A Review.

Engineering **5**, 261–275 (2019).

[36] Cui, R. & Liu, M. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Comput. Med. Imaging Graph.* **73**, 1–10 (2019).

[37] Arena, P., Basile, A., Bucolo, M. & Fortuna, L. Image processing for medical diagnosis using CNN. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.* **497**, 174–178 (2003).

[38] Piotrzkowska-Wróblewska, H., Dobruch-Sobczak, K., Byra, M. & Nowicki, A. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. *Med. Phys.* **44**, 6105–6109 (2017).

[39] To, U. E. *et al.* S d d c a. **345**, 1473–1482 (2001).

[40] Moody, G. B. & Mark, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **20**, 45–50 (2001).

3. Graph analysis with multi-functional self-rectifying memristive crossbar array

3.1. Introduction

A memristive crossbar array (CBA) is exploited as high-density nonvolatile memory or storage. [1–6] It also showed great potential in implementing the hardware of diverse neuromorphic networks as the synaptic weight-representing device or temporal/physical kernels. [7–10] For both applications, the passive configuration of the CBA renders the sneak current a severe problem. Sneak current flows in the plurality of parallel paths where the minimum resistance is formed. For the standard memory and several neuromorphic applications, the adverse effects from the sneak current flow were suppressed by adopting a selector or transistor.[1],[11],[12]

Nonetheless, the full potential of the memristive CBA has not been exploited yet. Especially there could be other applications than the standard memory and synaptic devices, which may utilize the parallel configuration of the array with the sneak current. The CBA may be used to solve mathematically complicated graphical problems. Moreover, considering the two-dimensional (or even three-dimensional) layout of the CBA, it may find an even higher potential for graphical problems.

A graph is a data structure that models a set of nodes connected by edges.

Several critical problems, such as the traveling salesman problem, can be intuitively represented in graph form. [13],[14] Therefore, they are gaining greater attention in the contemporary computing fields, such as understanding social networks, molecular structures, virus transmission networks, and the World Wide Web.[15–19] They have also been used in social science and biology.[20–24] This work exploits another potential of CBA utilizing the sneak current to solve several challenging problems, which the graphs can represent. These attempts include path-finding, link prediction, community detection problems, and brain network analysis, which have customarily been attempted by the software algorithms based on the appropriate similarity function.[25–28] However, these similarity function is not always optimal, and the software codes cannot find the solutions for several graphical problems without pre-processing, especially when the graphs are non-Euclidean.[29],[30]

Solving such challenging problems using the physical mechanism, such as utilizing the sneak current in CBA, could be a feasible option, as shown in this work. For this purpose, however, the sneak current must not be allowed to flow arbitrarily but in a controlled manner.

3.2. Experimental

The array of crossbar Pt/Al_2O_3/HfO_2/TiN memristors was fabricated through the following procedure. First, a 50 nm-thick TiN layer was sputtered (Endura, Applied Materials) on a SiO_2/Si substrate, and the TiN layer was patterned into a line shape to form a bottom electrode. The 2- to 10-μm-wide TiN BEs were patterned using conventional photolithography and the dry-etching system. Then, 4-nm-thick HfO_2 and 4-nm-thick Al_2O_3 were sequentially deposited using atomic layer deposition (ALD) at a 280 °C substrate temperature using a traveling-wave-type ALD reactor (CN-1 Co. Plus 200). A Tetrakis(dimethylamino)hafnium, trimethylaluminum, and O_3 were used as precursors for Hf, Al, and oxygen, respectively. Finally, 30-nm-thick Pt top electrodes were deposited using an electron beam evaporator (Sorona, SRN-200i) and patterned into 2- to 10-μm-wide lines using the conventional lift-off process. After the fabrication, the PAHT device was analyzed using x-ray photoelectron spectroscopy (XPS, AXIS SUPRA, Kratos) and energy-dispersive x-ray spectroscopy (EDS, JEOL, JEM-ARM200F) to observe the interfacial layer formation. Cross-sectional images of the PAHT memristor were examined using scanning transmission electron microscopy (STEM, JEOL, JEM-ARM200F).

The DC I-V characteristics of a single device were measured using the semiconductor parameter analyzer (SPA, HP4145B). During the single

device measurement, the top electrode (TE) was biased, and the bottom electrode (BE) was grounded. The AC pulse measurement was performed using SPA, pulse generator (PG, Agilent 81110A), and oscilloscope (OSC, Tektronix TDS 684C). The measurement of 9 × 9 mCBA was conducted in the setup of the 9 × 1 custom multiprobe (MS-TECH), switch matrix (Keithley 708A), and SPA (HP 4155B). During the mCBA measurement, TE and BE were used as a word line and a bit line, respectively, and were contacted by two 9 × 1 multiprobe. All electrical measurements were carried out with an interface based on LabViEW™.

3.3. Results and Discussions

3.3.1. SELF-RECTIFYING MEMRISTOR AND METAL CELL AT DIAGONAL CBA

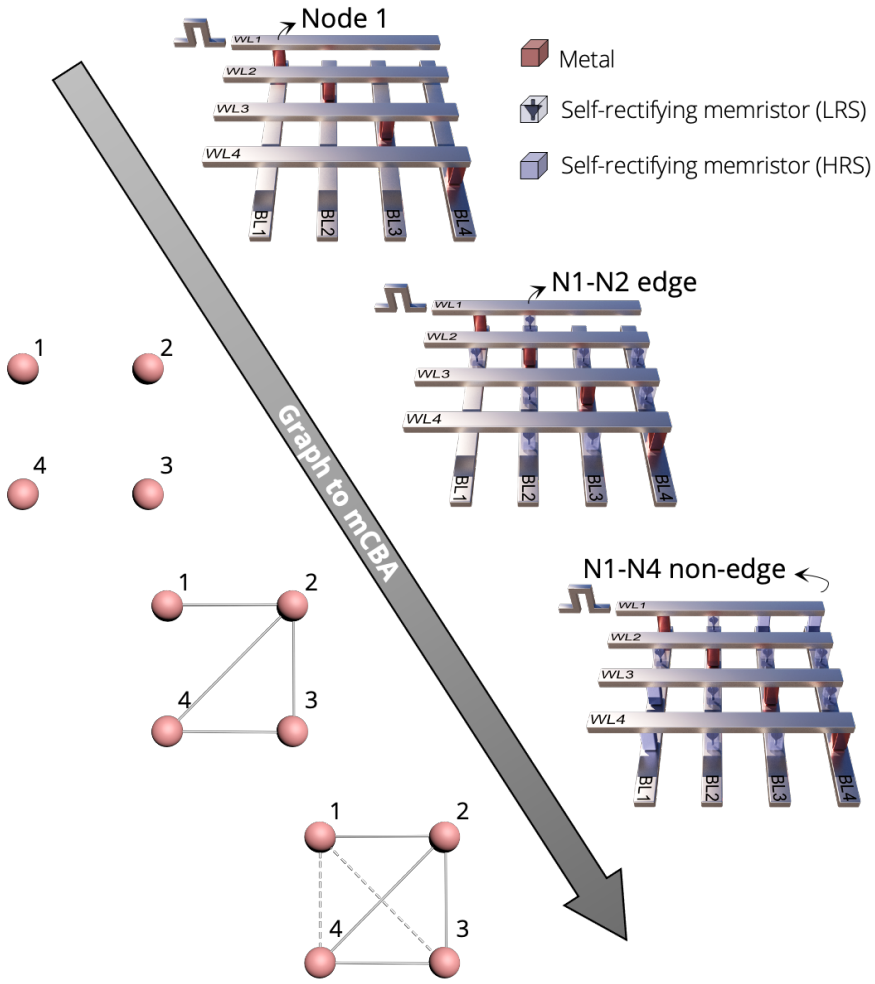


Figure 3-1. Graph to mCBA mapping. The resistance state of the (n, m) device corresponds to the weight of the (n, m) edge. The (n, n) metal cell represents the zero weight, which is the connection of the node itself.

Figure 3-1 shows how a non-Euclidean graph is mapped onto the hypothetical CBA. The connected edge weights are mapped to the low resistance of the memristors, while the unconnected edges are represented by the high resistance of the corresponding memory cells. The cells at the diagonal positions, (n, n), denote the connections between the node itself, represented by the shorted circuit at those locations. For example, the graphic representation of node 1 is implemented in the hardware CBA by allocating the metal vias at (1, 1) locations. Such a specific CBA is named "metal cell at diagonal CBA (mCBA)," which provides a crucial effect in implementing the desired sneak current-based graph algorithm using the self-rectifying memristors.

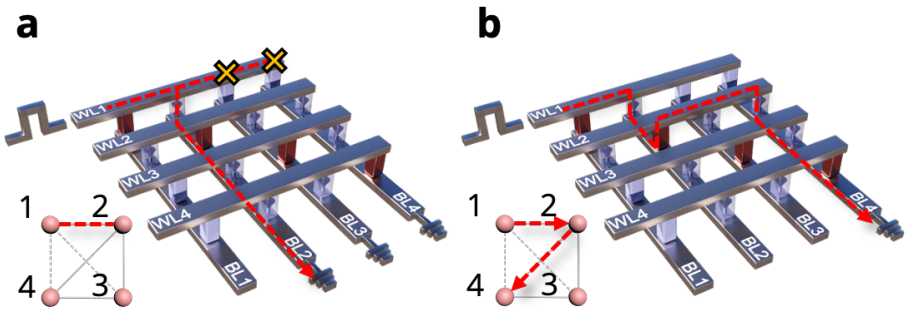


Figure 3-2. Two operation methods of mCBA. a, Multi-ground method (MGM). **b,** Single-ground method (SGM).

Figure 3-2 shows the mCBA-based methods to extract information in the non-Euclidean graph. The multi-ground method (MGM, Figure 3-2a) and the single-ground method (SGM, Figure 3-2b) are used as an adjacency search function that searches for nearby nodes, and a similarity function that represents the relationship between the two nodes, respectively. For example, in the MGM of word line 1 (WL_1), current in bit line 2 (BL_2) means that the adjacent node of node 1 of the inset graph is node 2. The current flow in the SGM between WL_1 and BL_4 indicates that node 1 arrives at node 4 through node 2. Figure 3-3 shows the general method of how the mCBA solves graphical problems. 1) Graph to mCBA mapping. 2) Analysis of similarity between graph nodes. (for all or part of total pairs) 3) Deriving desired results, such as the distance, probability of link formation, community formation order, and connectome classification, based on the identified similarity. As an example, a weighted,

undirected, non-Euclidean graph consisting of 9 nodes and 16 edges can be mapped to 9x9 mCBA of Figure 3-3. In the mCBA, 9 red rectangles on the diagonal represent the metal (9 nodes), and blue rectangles represent various RRAM device states (The lighter the color, the lower the resistance, the lower the weight), and dark rectangles represent the RRAM device in HRS. (non-edges) The mapped weights are implemented to the multi-resistance states of the self-rectifying memristor. In mCBA, to which the graph is mapped, the similarity between nodes can be extracted without the graph embedding process. Various applications such as path-finding, link prediction, community detection, and connectome analysis can effectively be performed based on the extracted information.

Figure 3-4a, c shows the HSPICE simulation results of MGM and SGM at the graph in Figure 3-3. Figure 3-4a shows MGM, in which the WL_1 corresponding to the source, or selected node in the graph, is biased, while all other BLs are grounded. MGM finds out the adjacency nodes, directly connected to the source. (Figure 3-4b) In this case, no sneak current is allowed. Figure 3-4c shows the SGM, where the selected WL_1 is biased, with only the target BL_9 being grounded (all other BLs are floated). By this connection, SGM can delineate the hidden information related to the connection between the source and the target by the sneak current through the metal. In addition, the main current path of SGM corresponds to the shortest path of the graph. (red dash and dot lines in Figure 3-4d)

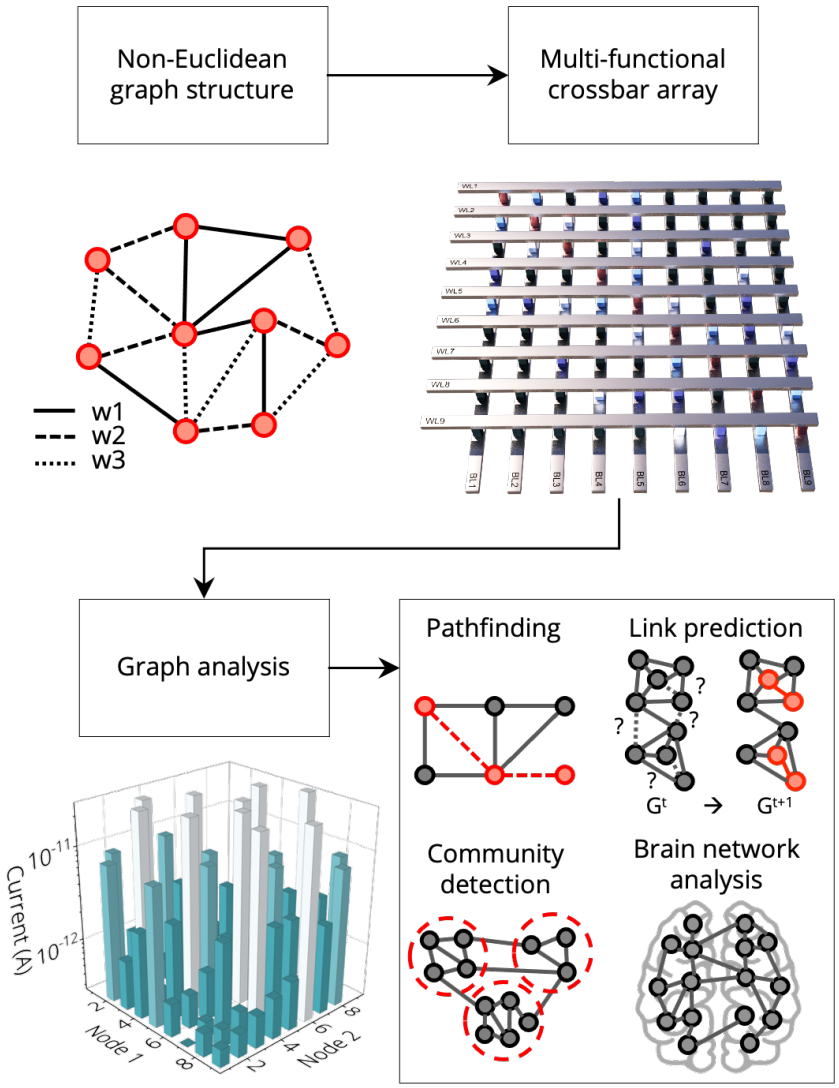


Figure 3-3. The process of analyzing non-Euclidean graphs with mCBA and the implementable applications.

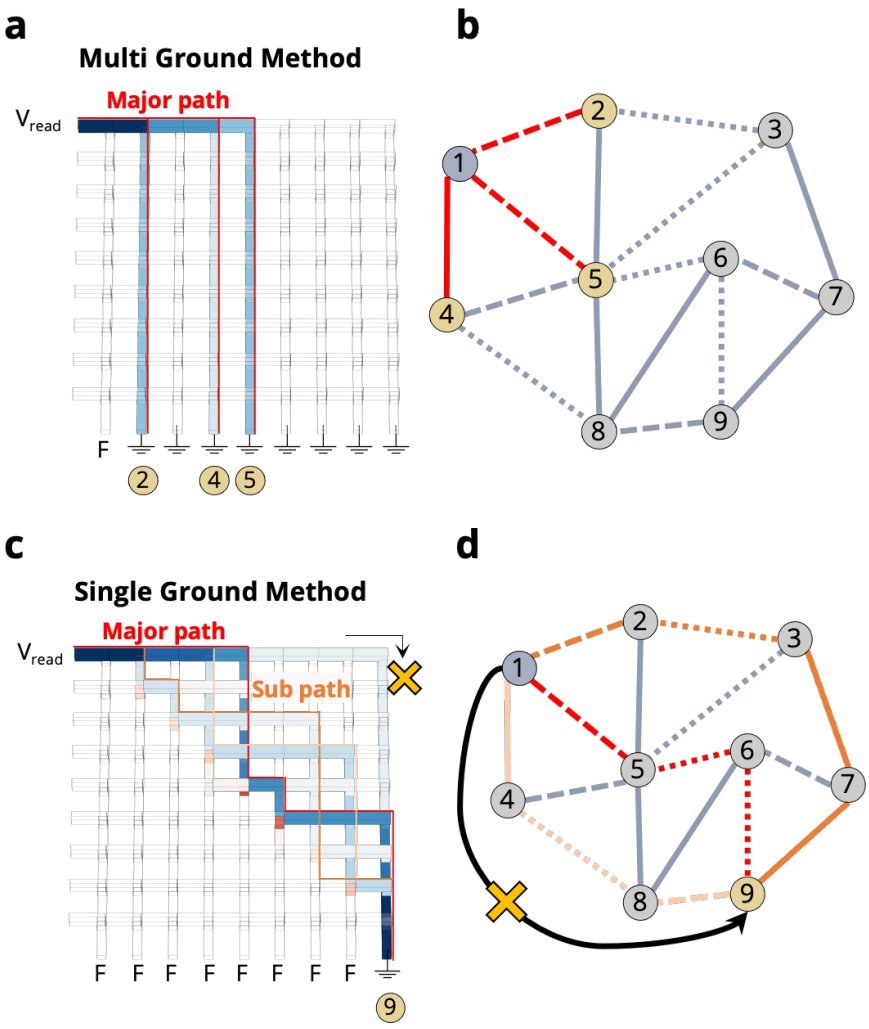
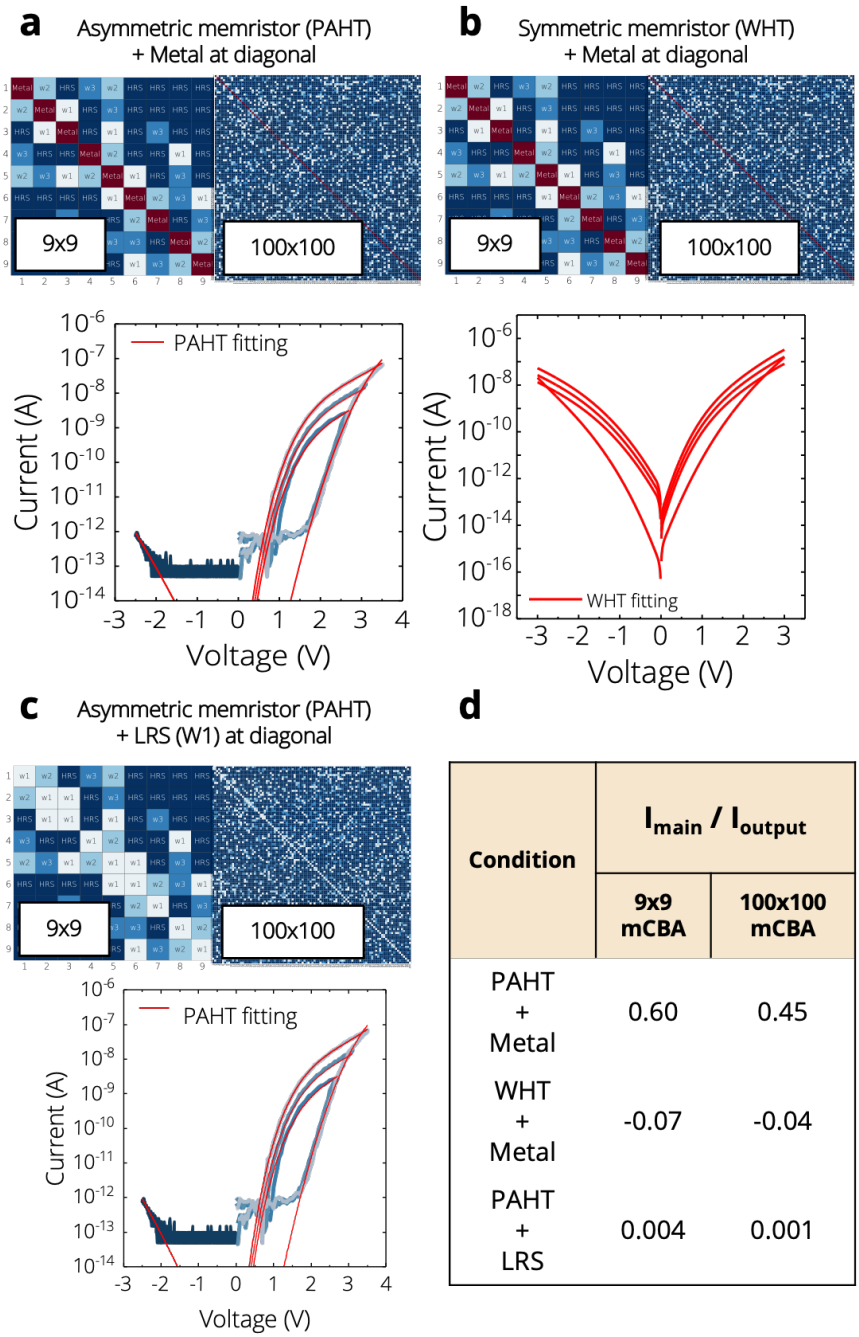


Figure 3-4. Simulation results for two operation methods of mCBA. a, HSPICE array simulation results for MGM at N1. **b**, The adjacency search result of MGM at N1. **c,** HSPICE array simulation results for SGM of N1 to N9. The major and sub-current paths are marked in red and orange, respectively. **d,** Multiple paths between N1 and N9 which are not directly connected. The shortest and sub-current paths are displayed in red and orange, which correspond to the current flow of **c**.



Condition	I_{main} / I_{output}	
	9x9 mCBA	100x100 mCBA
PAHT + Metal	0.60	0.45
WHT + Metal	-0.07	-0.04
PAHT + LRS	0.004	0.001

Figure 3-5. The main current ratio in the SGM at the various mCBA configurations. a-c, mCBA mapping (upper panel) and I-V fitting

curves (red lines) for the unit cell memristor (lower panel). The main current ratios for the 9 × 9 and 100 × 100 mapping were 0.60 and 0.45, respectively, when metal cells were placed on the diagonal cells, while self-rectifying cells were placed on the rest cells. **d**, The result of calculating the ratio of $I_{main\ path}$ and I_{output} in 9 × 9, 100 × 100 mCBA under conditions of **a**, **b**, and **c**.

Figure 3-5 shows the simulated SGM results when the WL_1 is biased with 1 V, and BL_9 and BL_{100} are grounded in various 9 × 9 and 100 × 100 CBA configurations, respectively. In the simulation, the finite wire line resistance of 50 Ω was considered due to the possible process issues. (The calculated line resistance was ~ 2 Ω calculated based on the resistivity of Pt [23] and TiN [23] and the dimension of the line in the fabricated mCBA) However, The line resistance did not significantly affect the read and write operations due to the $Pt/Al_2O_3/HfO_2/TiN$ (PAHT) memristor's high resistance level (R_{LRS} = 26 GΩ at V_{Read} = 1 V). In Figure 3-5a, due to the presence of parallel sneak paths, the current flows along diverse routes, and the ratio of current following the main path to all current paths is 0.60. This presence of the main current path was possible due to the metal cells at (n,n) positions, although the adopted PAHT memristor has a self-rectifying property. The sneak current can also flow when the sneak path involves the (n,n) positions; otherwise, the suppressed reverse current of the self-rectifying PAHT will not allow the sneak current to flow.

An interesting finding was that no meaningful current flow was achieved using the symmetrical memristor, i.e., non-rectifying I-V characteristics, which is supposed to have a higher sneak current (Figure 3-5b). This abnormal behavior could be owing to an overlap of the anti-directional sneak current flows at specific cells. Therefore, it was concluded that the mCBA configuration with the self-rectifying memristor was the most useful hardware to implement the suggested graph algorithms.

Besides, when the (n,n) cells were programmed to even the lowest resistance of the PAHT memristor, with the identical weight distribution at other cells, the ratio became negligible (0.004, Figure 3-5c). This result is due to the suppressed current flow under the reverse bias condition of the PAHT memristor, even with the lowest resistance. (Figure 3-5d).

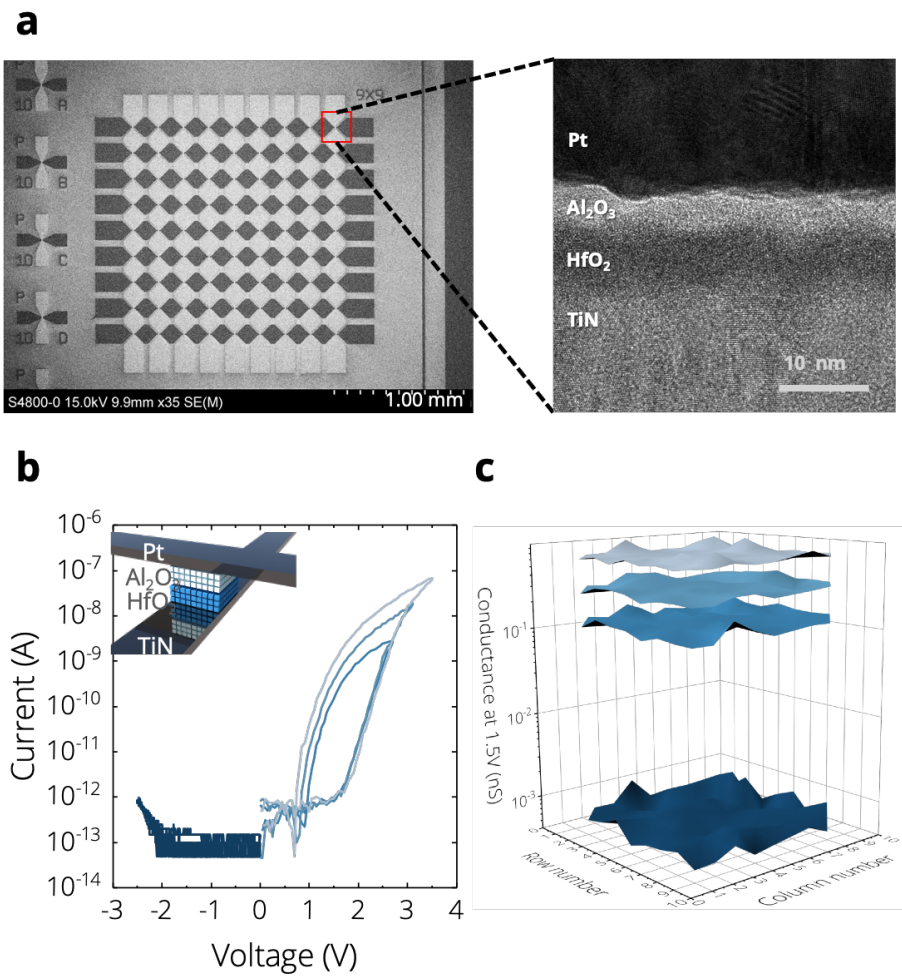


Figure 3-6. mCBA-array fabrication and the electrical analysis of the PAHT memristor. a, Scanning electron microscope (left) images of 9 x 9 mCBA and a cross-section transmission electron microscope image (right) of the PAHT memristor. **b,** I-V characteristic of the PAHT memristor at various set sweep voltages (2.7 V to 3.5 V). The inset of **b** is the PAHT memristor stack schematic. **c,** The surface plot of the three levels of conductance data of 9x9 mCBA.

The adopted PAHT self-rectifying memristor is composed of the top electrode Pt/4nm-Al_2O_3/4nm-HfO_2/bottom electrode TiN structure, where the Al_2O_3 and HfO_2 layers were grown by the atomic layer deposition, while the Pt and TiN layers were grown by the electron-beam evaporation and reactive sputtering, respectively. Figure 3-6a shows the scanning electron microscope (SEM, left panel) and the cross-section transmission electron microscope (TEM, right panel) images of 9 x 9 mCBA.

Figure 3-6b shows the current-voltage (I-V) curve of the PAHT memristor at various maximum sweep voltages (2.7, 3.1, and 3.5 V) during the set switching. The PAHT memristor has stable counterclockwise bipolar resistive switching behavior and exhibits multi-states, self-rectifying, forming-free, and gradual switching characteristics. At 1.5 V of reading voltage, the conductance of the PAHT memristor was continuously increased from 0.01 nS to 0.62 nS with the 2.7 ~ 3.5 V of set voltage. Figure 3-6c shows the distribution of the HRS and three LRS conductance values of 9 x 9 mCBA cells.

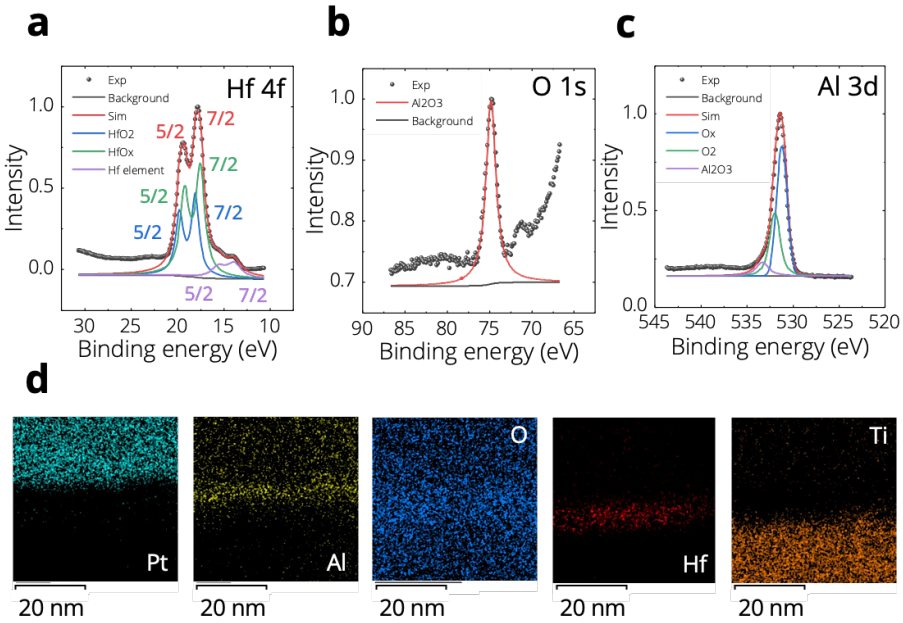


Figure 3-7. Chemical and physical analysis of the PAHT memristor. a-c, Hf *4f*, O 1s, and Al 3d X-ray photoelectron spectroscopy (XPS) analysis at the Al_2O_3/HfO_2 interface in the PAHT device. d, Energy-dispersive X-ray spectroscopy (EDS) mapping result of the PAHT memristor in cross-section TEM.

The physical and chemical structures of the PAHT devices are reported in Figure 3-7. The X-ray photoelectron spectroscopy (XPS) analysis of the Al_2O_3/HfO_2 interface in the PAHT device revealed the presence of Hafnium sub-oxide (HfO_x) and Hf elements. HfO_x is expected to be formed by the thermal ALD process of Al_2O_3 on the HfO_2 layer, while the metallic Hf could be induced by the in-situ etching of the top electrode during the XPS analysis. The ALD process of Al_2O_3 made the HfO_2 layer have a high trap density, which enabled the device to have a high On/Off ratio and long retention (Figure 3-8) [6]. In addition, Al_2O_3 in the device acts as a voltage divider to suppress the abrupt resistive switching of HfO_2. It also forms a high Schottky barrier at the interface with upper Pt, which makes the device have good rectifying properties.

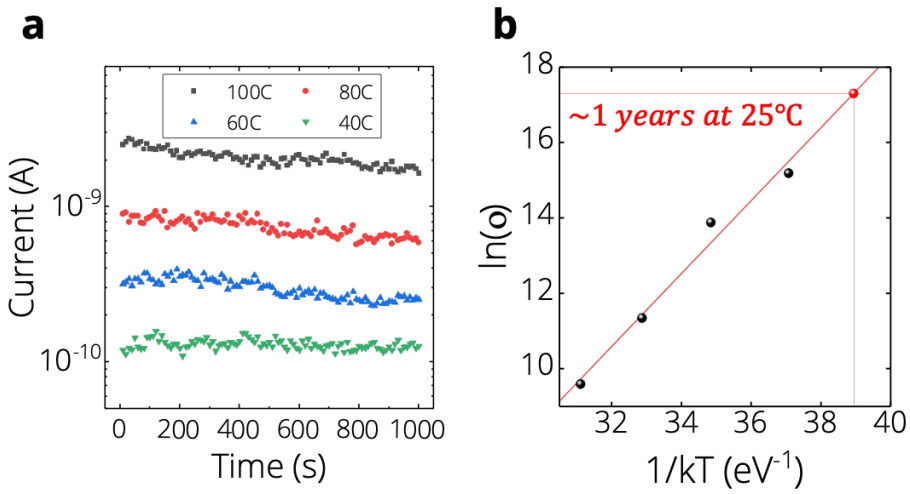


Figure 3-8. Retention of the PAHT memristor. a, Retention of the PAHT memristor measured at various temperatures (40 ~ 100 °C). **b**, Arrhenius plots of ln (τ) versus 1/kT of the LRS retention. A retention time of ~ 1 year (relaxation time from LRS level to HRS level) was obtained at room temperature by extrapolating retention data at 40 ~ 100 °C.

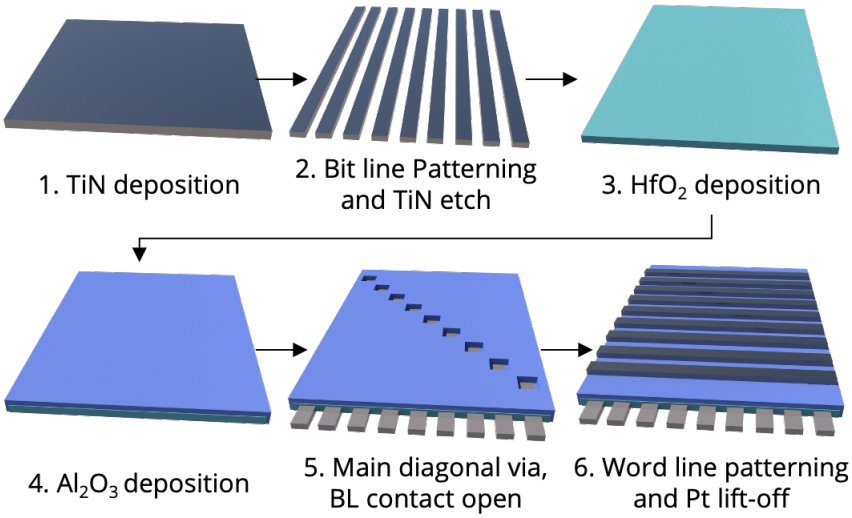


Figure 3-9. The process flow of the mCBA fabrication.

Figure 3-9 shows the fabrication process of the PAHT mCBA. To shorten the main diagonal cells with the top electrode material, the switching layer at those locations was etched during the BL contact open step of the conventional CBA process.

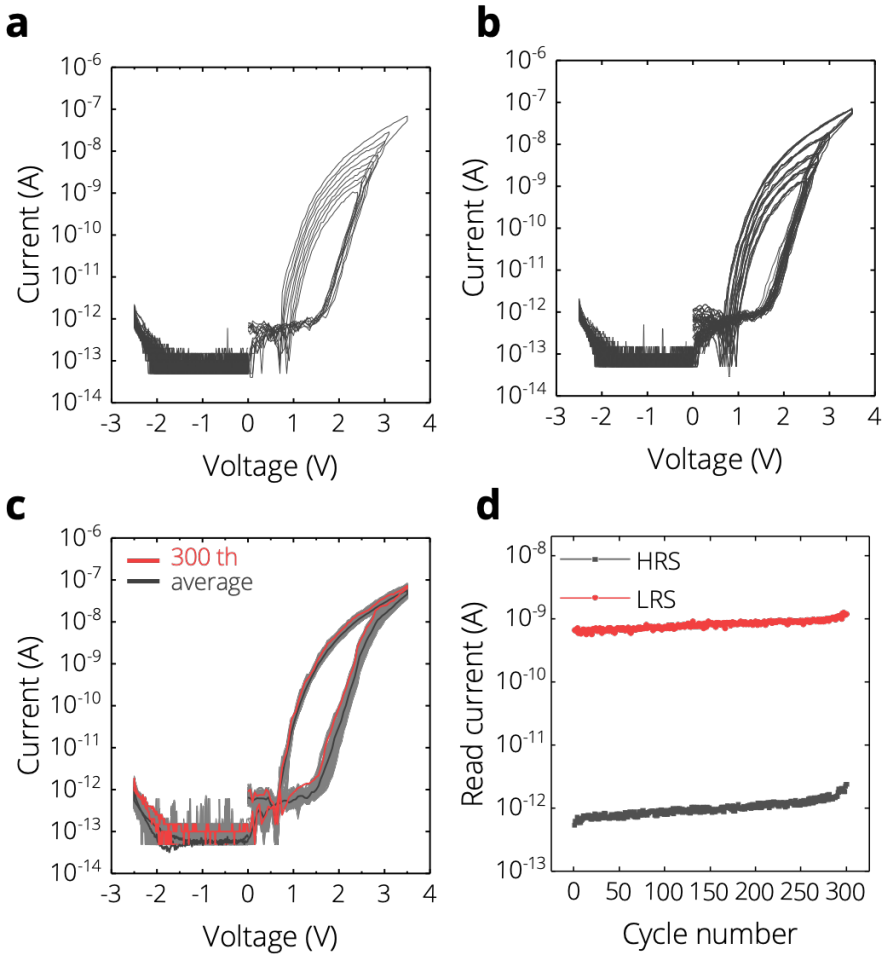


Figure 3-10. Multi-level and dc cycle results of the PAHT memristor. a, I-V curve when the DC sweep (SET) voltage is set to 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, and 3.5 V (9 states). **b**, Result of the 300 DC cycle of the PAHT memristor (set sweep: 3.5 V, reset sweep: -2.5 V).

I-V curves of the PAHT memristor for more than three states are included in Figure 3-10a. Figure 3-10b shows the results of the 300 consecutive I-V curves of the PAHT memristor, showing the low cycle-to-cycle variation.

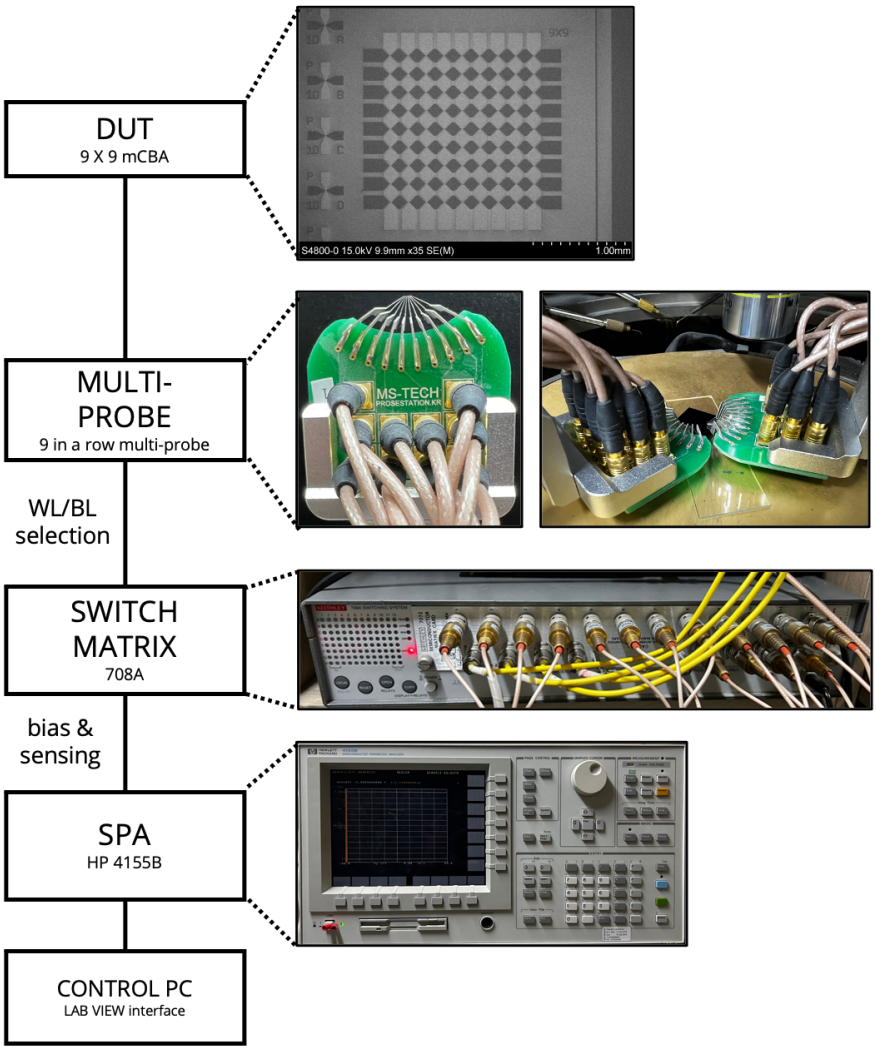


Figure 3-11. Measurement setup for the 9x9 mCBA. Flow chart of the 9x9 mCBA measurement. The 9x9 mCBA was measured in the setup of the 9x1 custom multiprobe, switch matrix, and semiconductor parameter analyzer.

The detailed measurement setup and the measurement flow are described in Figure 3-11.

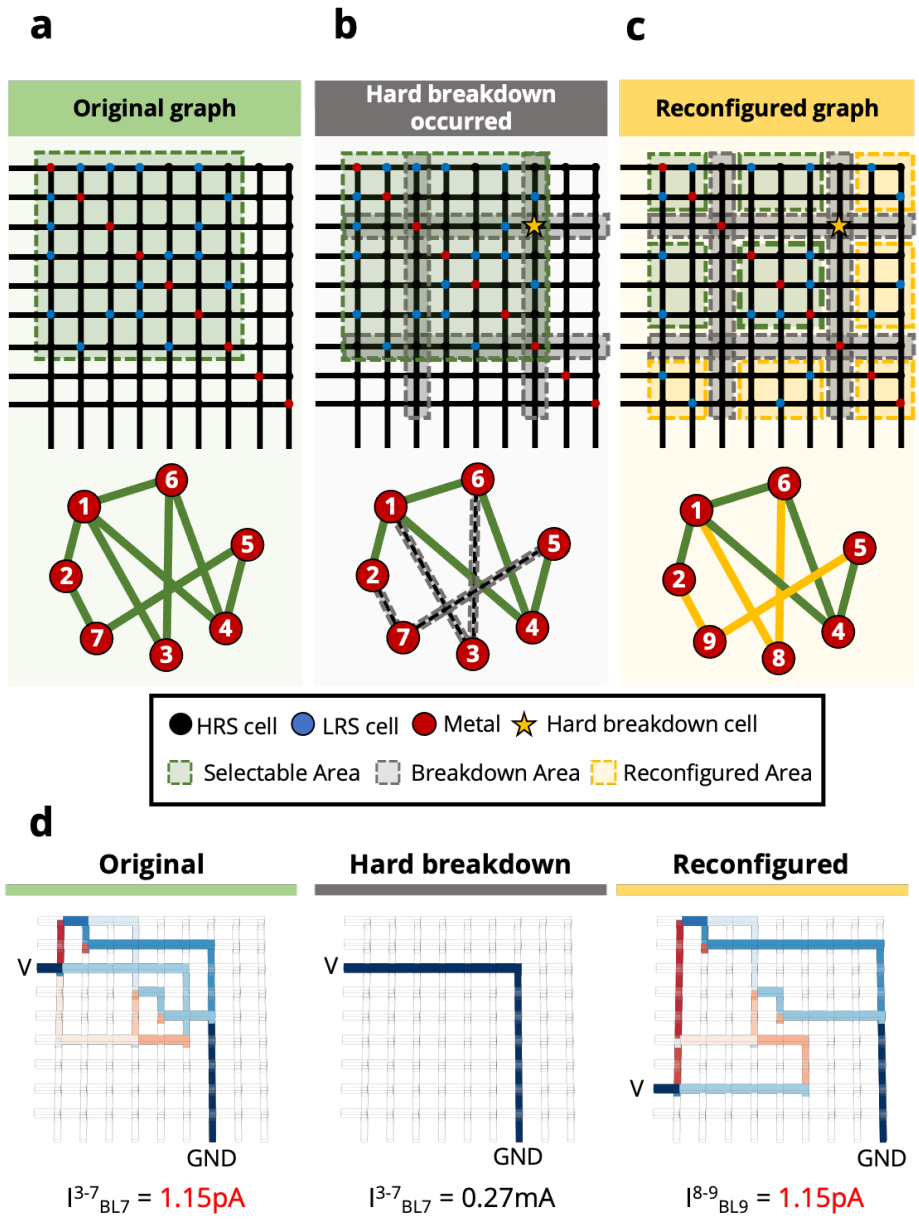


Figure 3-12. Reconfigurability of the mCBA. a, 9x9 mCBA (upper panel) to which the graph of the lower panel was mapped. **b**, Affected area of the mCBA (upper panel) and affected edges of the graph (lower panel) when a hard breakdown occurs in the (3, 7) cell of

the array. **c**, Results of remapping the affected part in mCBA (upper panel) and the recovered graph (lower panel). For the restoration, the edge data connected to nodes 3 and 7 are moved to BL_8, BL_9, WL_8, and WL_9, and the cells of BL_3, BL_7, WL_3, and WL_7 are changed to HRS. **d**, The current path and value of the SGM in the original graph, breakdown case, and the restored graph.

On the other hand, an issue in which a cell expressing an edge becomes inoperable may occur during mCBA operation. If hard breakdown occurs in the mCBA cell, it will cause problems in SGM-based graph data analysis. In this case, the original graph data can be restored by remapping the affected edge data based on the reconfigurability of the mCBA (Figure 3-12).

3.3.2. PATH-FINDING PROBLEM

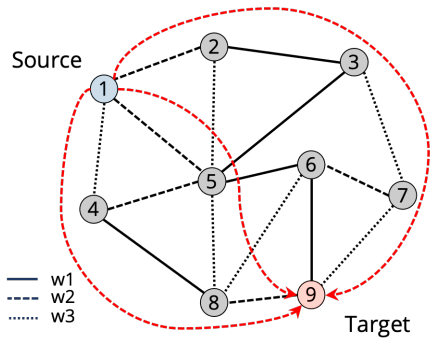


Figure 3-13. An example weighted network. The red arrow indicates various paths from node 1 to node 9.

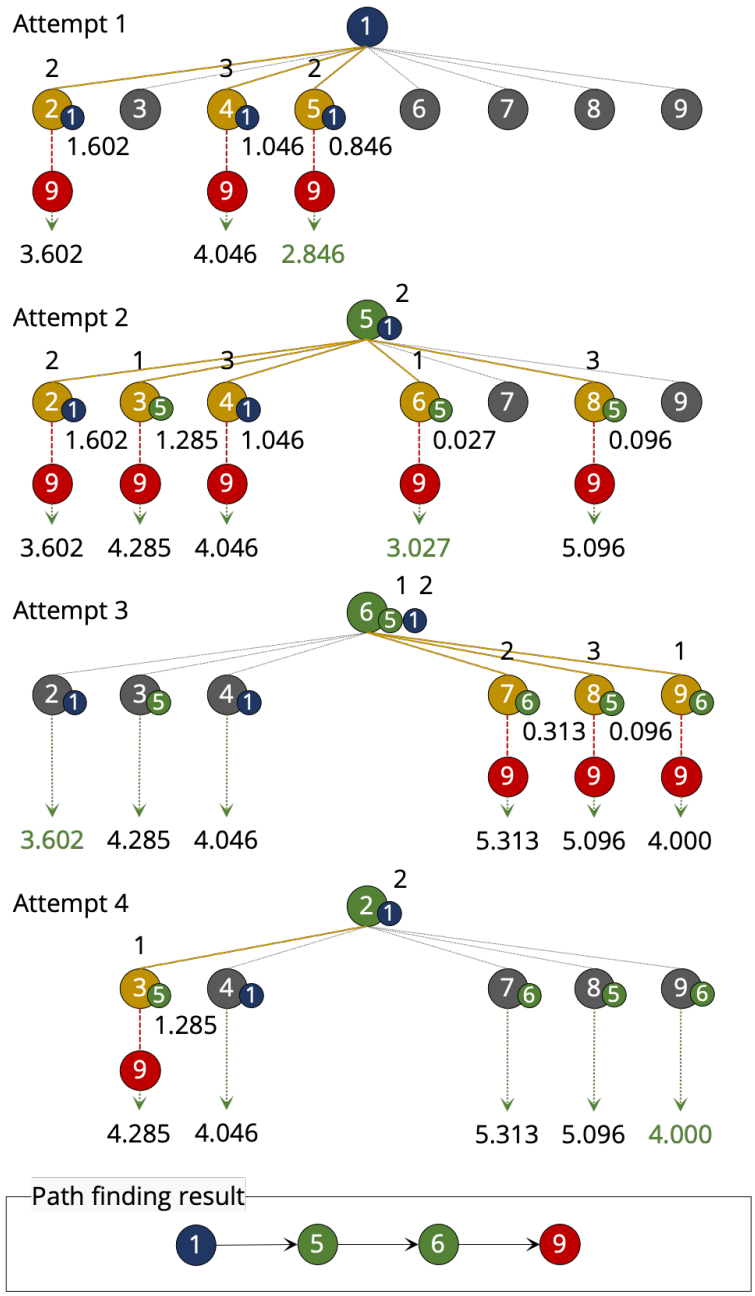


Figure 3-14. mCBA-based pathfinding algorithm. a, Process of finding the shortest path from N1 to N9 with the mCBA-based pathfinding algorithm. Pathfinding consists of two steps: 1. Search neighbor nodes (NNs) and the actual distance to the neighbor node with

the MGM, and calculate the distance from the neighbor node to the target node (TN) as the reciprocal of the SGM. 2. Go to the adjacent node with the lower sum of the cumulative sum of the actual distance (source node to present node) and the estimated distance (NN to TN).

This part describes how the MGM and SGM in mCBA can solve the path-finding problem. For this purpose, a non-Euclidean graph of Figure 3-13 is mapped onto the mCBA, and 1.0 V of reading voltage was used in the MGM and SGM. This method aims to find the minimum value of the F score that can be calculated as follows:

$$F_{i,j}^{t} = \sum_{t \to 1} f(MGM_i) + \frac{a}{SGM_{i,j}} \cdot k \,,$$

$$f(x) = \begin{cases} 1, & I_{degree1} < x \\ 2, & I_{degree2} \le x \le I_{degree1} \\ 3, & I_{degree3} \le x < I_{degree2} \end{cases} \quad (1)$$

, where F, t, i, j, MGM_i, f(x), $SGM_{i,j}$, a, and k are the F score, the number of attempts, source node, target node, MGM current for node i, step function according to MGM output current, SGM current from node i to j, scaling constant, and heuristic scale factor, respectively. In addition, $I_{degree1}$, $I_{degree2}$, $I_{degree3}$, a, and k were set to 10 pA, 5 pA, 1 pA, $5x10^{-13}$, and 1.5, respectively. A more detailed explanation of Eq (1) is included in the discussion related to Figure 3-14, 15.

The first step is to find the nodes connected to the starting node 1, which can

be accomplished by finding the MGM current from node 1 (or WL_1, Figure 3-15a). As a result, it was identified that nodes 2, 4, and 5 are connected to node 1, as the BL_2, BL_4, and BL_5 currents were detected. The next step is identifying which one should be chosen among the three connected nodes (Figure 3-15b-d). MGM and SGM at the three nodes allow for determining the shortest path.

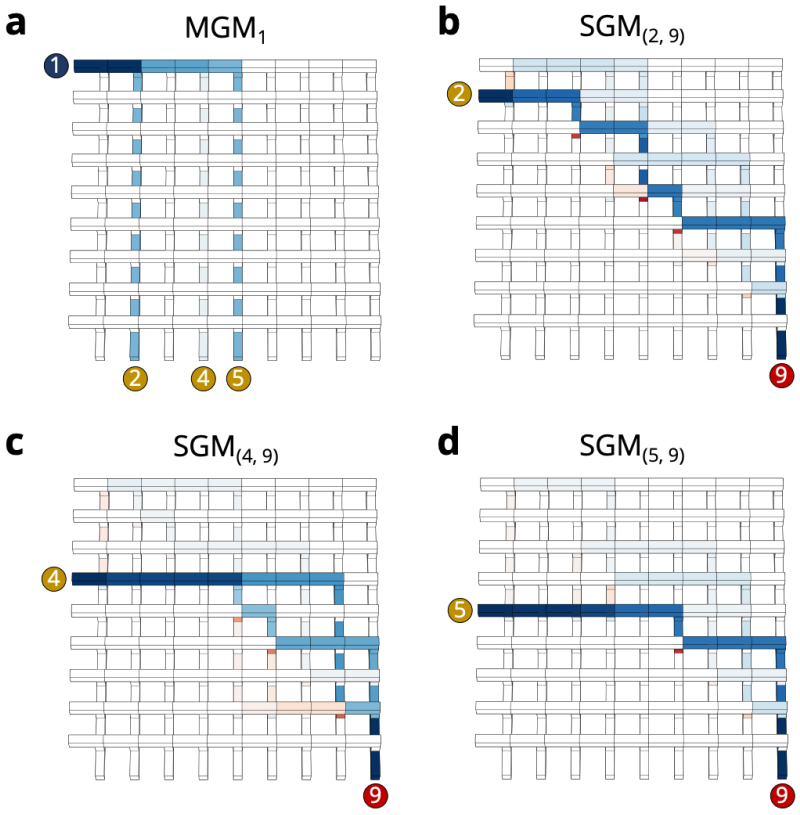


Figure 3-15. MGM and SGM current path at N1. At the source node (N1), the neighboring nodes, N2, N4, and N5, are searched for by the MGM. (left upper panel) From the SGM of the neighbor node of N1 to the target node, it can be seen that N5 is closest to the target node.

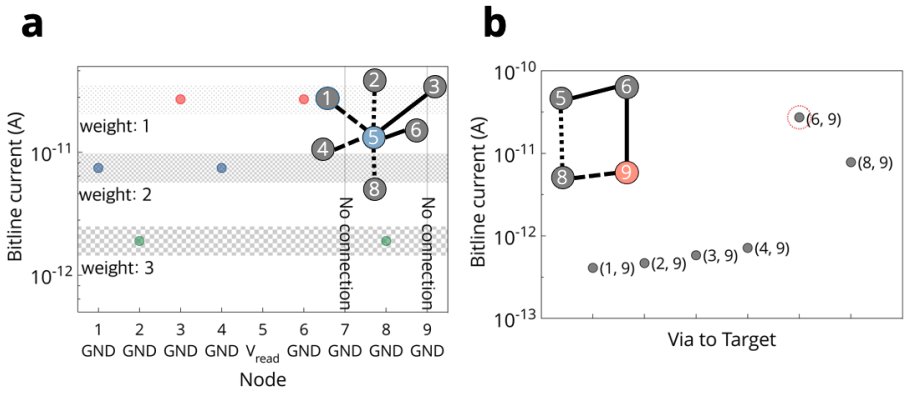


Figure 3-16. MGM and SGM current value at N1. a, MGM result at node 5. Based on the current level, the weights of adjacent nodes of node 5 can be identified, which coincides with the inset figure. **b**, SGM results from the neighbor node of node 5 to the target node.

For example, Figure 3-16a shows the results of MGM operation at node 5, which is a part of the path-finding process from node 1 to node 9. The BLs with currents above $I_{degree3}$ are 1, 2, 3, 4, 6, and 8, which correspond to the node numbers directly connected to node 5 of the inset figure. The grey box in Figure 3-16a indicates the current level of each weight in discrete quantities. MGM current level can determine the distance to the neighboring node. Figure 3-16b shows the SGM results from the neighboring nodes of node 5 to the target node. The node pair (6, 9) shows a higher BL current than other node pairs because the distance from node 6 to node 9 is the shortest (highest conductance at (6,9) node of the mCBA), as shown in the inset figure. Besides, since node 6 also had the highest MGM current in Figure 3-16a, it is determined that node 6 is the next node to go to from node 5. Although node 3 had the same highest current as node 6, the SGM current of the (3,9) pair was lower in Figure 3-16b, suggesting that the path involving node 3 is not the shortest. When a similar analysis was performed for nodes 2 and 4, all the output BL current was lower than the optimal one (1 → 5 → 6 → 9). Therefore, MGM and SGM can be used as a method to obtain information on adjacent nodes and as a method for estimating an approximate distance, respectively. The minimum value of the F score is obtained by updating the node where the cumulative f(MGM) is the minimum, and the SGM is the maximum in each trial t.

In MGM, the sneak current is suppressed, allowing accurate information to be achieved. However, only approximate information is obtained in SGM

because the sneak current is used.

The heuristic scale factor, k, in Equation (1) gives flexibility to the path-finding algorithm by adjusting the weight of the SGM. As shown in Figure 3-17, as k increases, the average number of attempts decreases, but the number of incorrect attempts increases. Conversely, as k decreases, the average number of trials increases, but an optimal solution with zero incorrect is guaranteed. By appropriately adjusting k, it is possible to implement a path-finding algorithm according to the desired accuracy or efficiency. Generally, the SGM score includes a higher error, especially when the attempt is made at a location far from the target node. Therefore, the score value becomes more accurate as the attempt number increases.

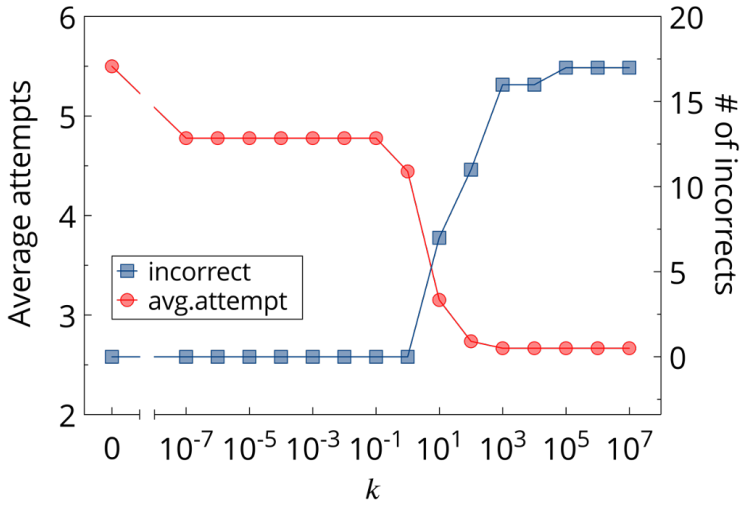


Figure 3-17. The path-finding result for all 72 paths of the graph in Figure 3-13. The average number of attempts (red) and incorrect results (blue) according to the heuristic scale factor k were plotted.

The SGM-based estimated distance was compared with the Landmark embedding distance, widely used for vectorizing non-Euclidean graphs (Figure 3-18) [24]. From the Figure 3-13 graph, 2~5 landmarks were randomly designated and embedded in multidimensional space to obtain Euclidean distance and Manhattan distance.

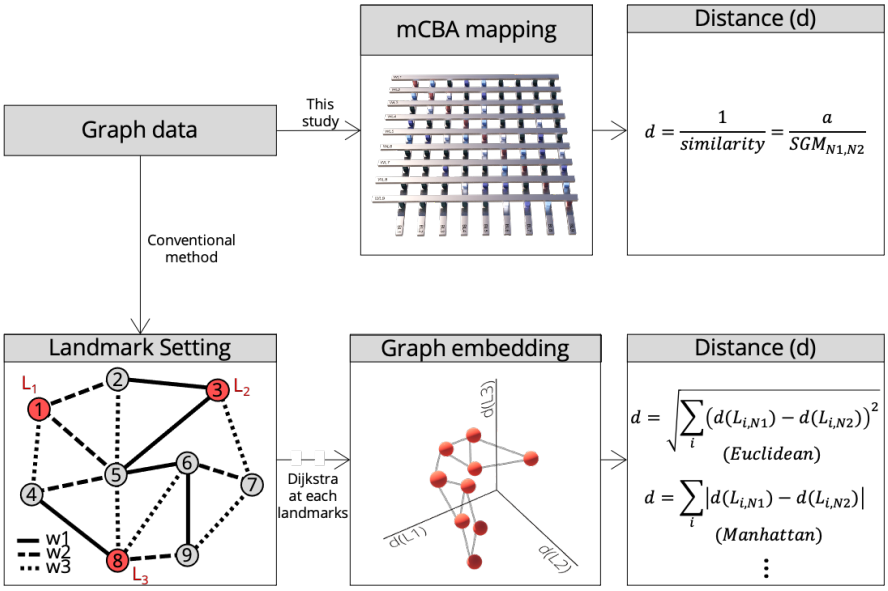


Figure 3-18. Distance calculation method in non-Euclidean graph based on mCBA and software algorithm.

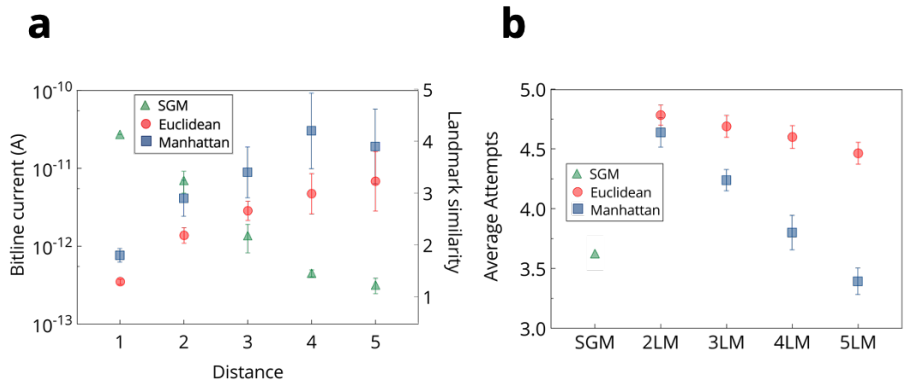


Figure 3-19. Comparison of the mCBA and Landmark algorithm for the pathfinding results. a, Comparison of SGM currents of mCBA and software algorithm-based distance estimation. Euclidean distance and Manhattan distance were obtained using a landmark algorithm (2 nodes were set as landmarks), and the bit line current obtained using SGM was plotted according to the actual distance. **b**, Average attempts of landmark algorithm and mCBA-based algorithm.

Figure 3-19a compares the Euclidean distance obtained using two landmarks, the Manhattan distance and the BL current obtained using the SGM, and the actual distance. In the case of landmarks, the deviation according to the randomly extracted landmarks is enormous, so the landmark distance does not effectively represent the actual distance. In contrast, each SGM current level expresses the actual distance well (the smaller the current, the longer the distance), showing the excellence of the SGM embedding method. This high performance is because the non-Euclidean graph is directly mapped to mCBA, so there is no loss due to the data pre-processing. Next, the average number of trials according to the number of landmarks was compared with the average number of trials of the proposed MGM + SGM embedding method (Figure 3-19b). The proposed method of this study was superior to the result of using four landmarks. Each time a landmark increases, the cost to be performed in pre-processing increases. When embedding a graph consisting of N nodes by setting L landmarks, the time complexity of embedding the graph is $O(L \times N^2)$ [25]. On the other hand, the time complexity of similarity calculation in the mCBA is O(1) since only one SGM is required for the graph data stored in the mCBA. Therefore, the MGM+SGM method shows excellent embedding performance without pre-processing for the path-finding tasks, even in non-Euclidean space.

3.3.3. LINK PREDICTION

The combined operation of the MGM and SGM can also be used to predict the evolution of graphic networks, such as friend recommendations in social network service (SNS) [26], [27] and product recommendations in e-commerce [28], [29]. These are generally regarded as link prediction problems. The mCBA can efficiently implement the link prediction using the MGM- and SGM-based similarity indices. For link prediction, the similarity indices can be used as a similarity score $S(i,j)$, defined as Eq. (2)

$$S(i,j) = MGM_i \cdot MGM_j \cdot SGM_{i,j} \qquad (2)$$

, where MGM_i and MGM_j indicate the number of edges (degree, d) connected to node i and j, respectively, which can be calculated by the MGM current. $SGM_{i,j}$ indicates the SGM current between the node i and j.

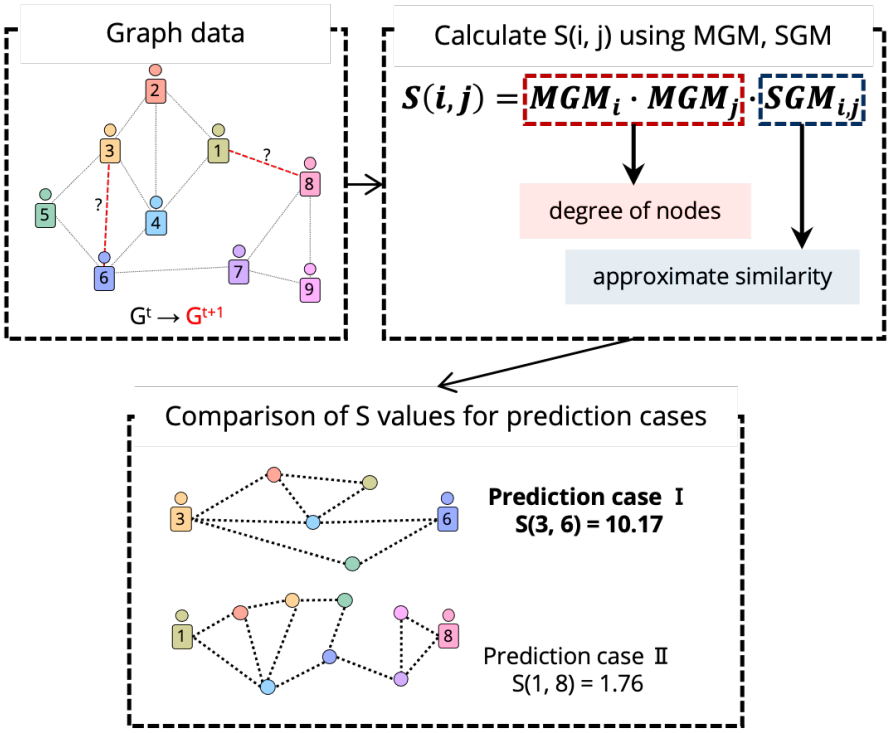


Figure 3-20. Schematic diagrams of link prediction algorithm and community detection algorithm using similarity index based on SGM and MGM.

For the social network graph of Figure 3-20, the MGM+SGM-based link prediction system predicts which non-edge among (3,6) and (1,8) will change to edge at time t+1 step in G^{t+1} graph. Similarity scores are assigned to each unconnected pair; the higher the score, the higher the probability that the pair will be connected. If there are many low-hop connections between two nodes and the degree of each node is high, a high similarity index is assigned in this MGM+SGM-based link prediction algorithm. For example, in prediction case

1, there are two 2-hop connections between nodes 3 and 6, and the degree of nodes 3 and 6 is high, so prediction case 1 will be assigned a higher similarity than prediction case 2, which has 4-hop connections and a low degree. To make this prediction, MGM_i, MGM_j and $SGM_{i,j}$ values (1.0 V of reading voltage) for each source-target combination are calculated (Figure 3-21), which eventually generate the S(3,6) = 10.17 and S(1,8) = 1.76 (Figure 3-20). Therefore, the link between persons 3 and 6 will be made, but the link between persons 1 and 8 will not be at G^{t+1} stage. When seeing the sneak current paths for SGM in Figure 3-21, the short paths containing 3→4→6 and 3→5→6 (2-hop) comprise the main current path. In contrast, there are many more paths between 1 and 8, but none are 2-hop paths, so the effective current is lower. This circumstance represents the connection configuration of G^t precisely, and thus, the link prediction must be accurate.

Graph sampling is used to evaluate the performance of scores for static graphs. Figure 3-22 shows the S value distribution of non-edges and sampled non-edges. Sampled non-edges had the highest score, indicating that the S value predicts links reflecting the graphical structure. The performance of this MGM+SGM method is compared with other software-based algorithms for various datasets. The receiver operating characteristics (ROC) curve and the area under the ROC curve (AUC) were used as the evaluation metrics.

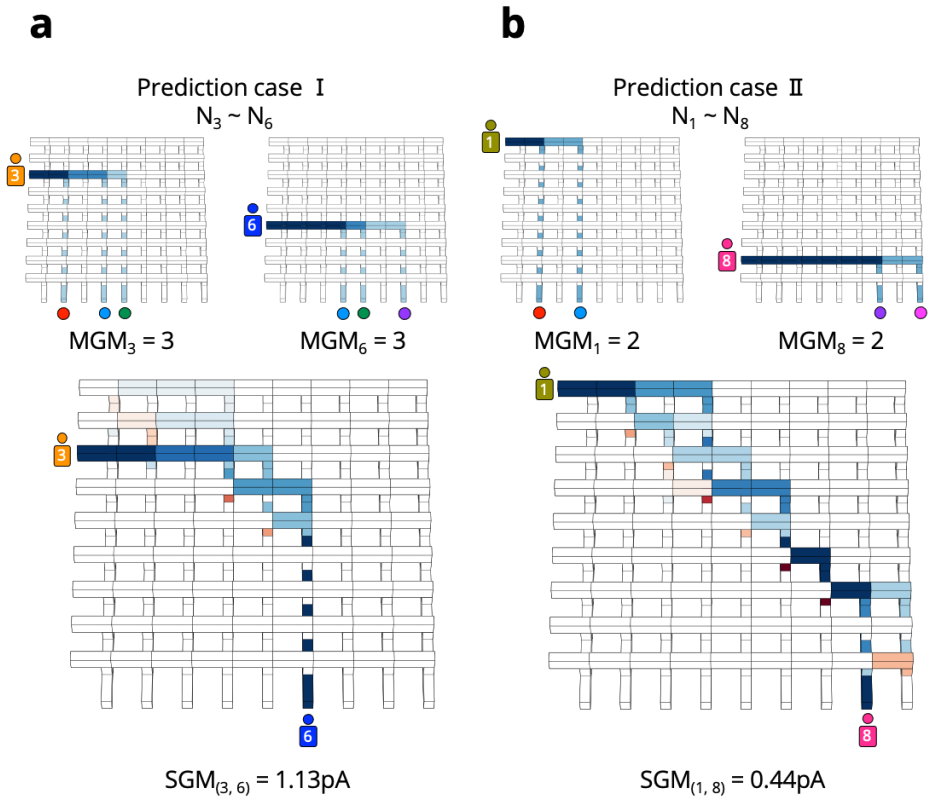


Figure 3-21. MGM+SGM similarity score. a and b, MGM and SGM results in case 1 (node 3, 6) and case 2 (node 1, 8). Calculation procedures of S(3, 6) and S(1, 8). For the non-edge (3, 6), MGM_3 = 3, MGM_6 = 3, $SGM_{(3,\,6)}$ = 1.13 pA and S(3, 6) = 10.17. For the non-edge (1, 8), MGM_1 = 2, MGM_8 = 2, $SGM_{(1,\,8)}$ = 0.44 pA and S(1, 8) = 1.76.

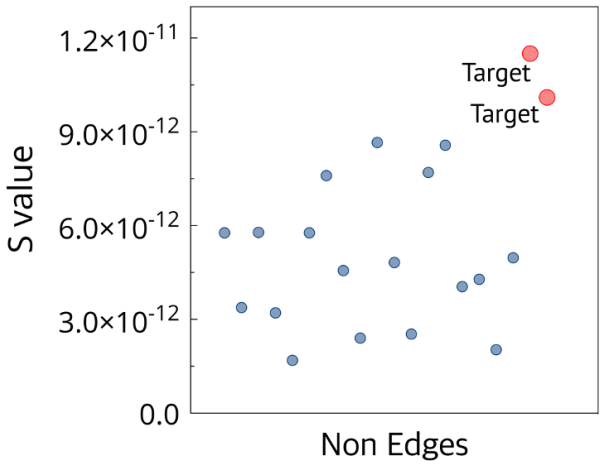


Figure 3-22. Similarity values assigned to non-edges and sampled non-edges after 20% sampling in the example graph of Figure 3-20. Since sampled non-edges are created by cutting the original edges, high S values are assigned due to peripheral connections.

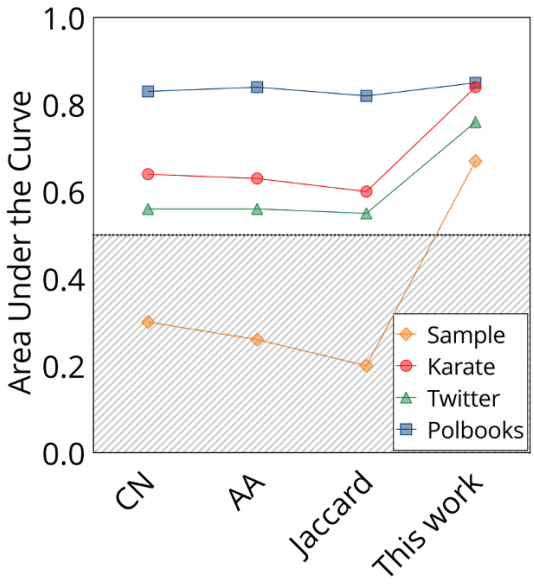


Figure 3-23. Performance results (area under ROC curve) for Zachary's karate club, Books about US politics, and Twitter retweet network datasets of SGM+MGM, CN, AA, and Jaccard indices. SGM+MGM showed the highest and most consistent performance in the four datasets.

Figure 3-23 shows that for all the tasks, the MGM+SGM method outperforms all other competitors, demonstrating the superiority of the suggested approach.

Unlike existing methods based on counting the number of nodes that satisfy a specific condition, mCBA-based link prediction utilizes SGM. Therefore, MGM+SGM metrics can score inter-node connectivity precisely and continuously.

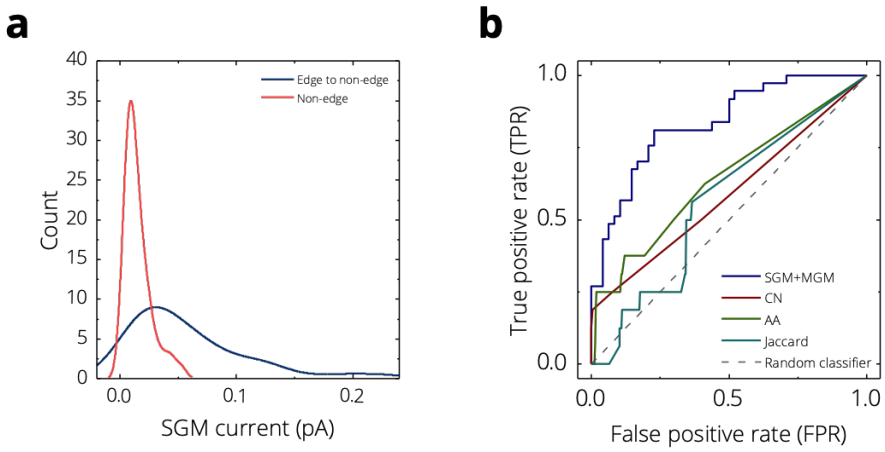


Figure 3-24. SGM distribution and ROC curves of each algorithm for the Zachary's karate club dataset. a, Distribution plot of the SGM+MGM index values. **b**, Receiver operating characteristic (ROC) curve of the SGM+MGM index values.

Several rising points in the ROC curve of Figure 3-24 show how this characteristic brings high performance in Zachary's karate club dataset. Many rising points in the ROC curve mean that similarities between 'non-edge' and 'edge to non-edge' are well separated, and the ROC curve is located above the dashed diagonal. There are many rising points in the ROC curve of MGM+SGM because continuous scoring can distinguish subtle connectivity differences that node-counting-based algorithms cannot. In Zachary's karate club dataset, the number of rising points of MGM+SGM, CN, AA, and Jaccard are 15, 6, 8, and 8, respectively.

3.3.4. COMMUNITY DETECTION

Another application of mCBA-based hardware is community detection, which binds dense groups within a given community based on similarity.

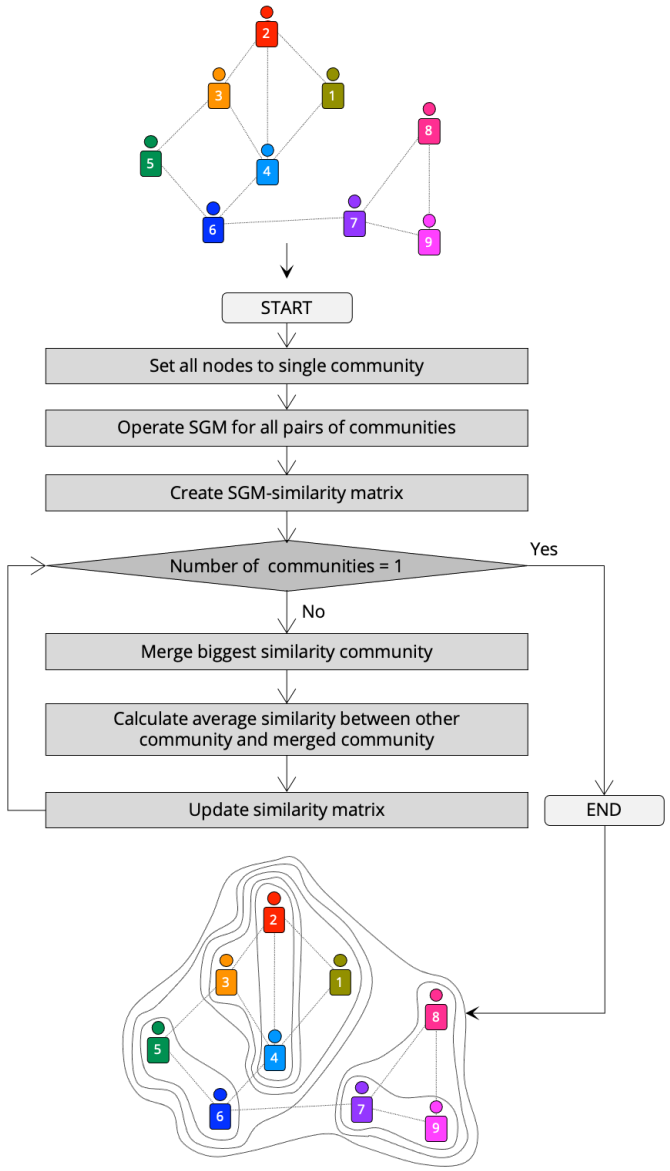


Figure 3-25. The flow chart that describes the community detection algorithm using SGM-similarity in a small social network composed of 9 people.

Figure 3-25 shows a schematic diagram that describes the community detection algorithm using SGM-similarity matrix S, defined as Eq. (3) [30].

$$S = \begin{bmatrix} SGM_{1,1} & \cdots & SGM_{1,j} \\ \vdots & \ddots & \vdots \\ SGM_{i,1} & \cdots & SGM_{i,j} \end{bmatrix} \qquad (3)$$

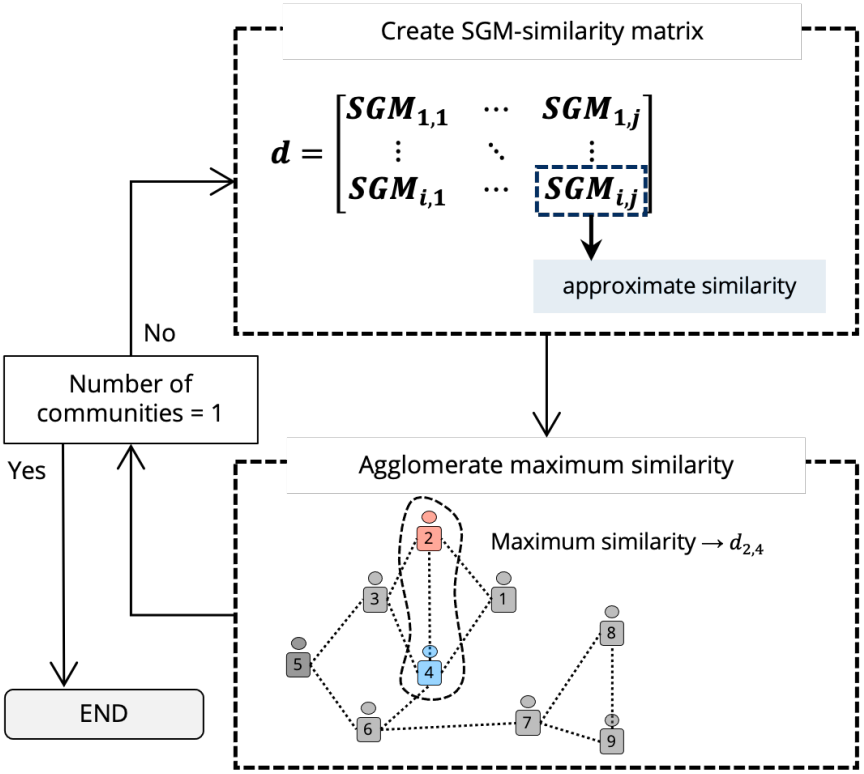


Figure 3-26. Schematic diagrams of community detection algorithm using similarity index based on SGM.

Figure 3-26 shows the schematic diagram of the mCBA-based community detection algorithm. In the SGM-based community detection algorithm, a node (or community) pair having a high SGM current (similarity) forms a community first. Unlike link prediction, the community detection algorithm proceeds by comparing edges, not non-edges. The more the low-hop bypasses in addition to the direct connection (edge), the greater the similarity and the higher the connectivity. After repeating community formation until the entire graph becomes one community, the step with the most increased clustering is identified, and the algorithm is terminated.

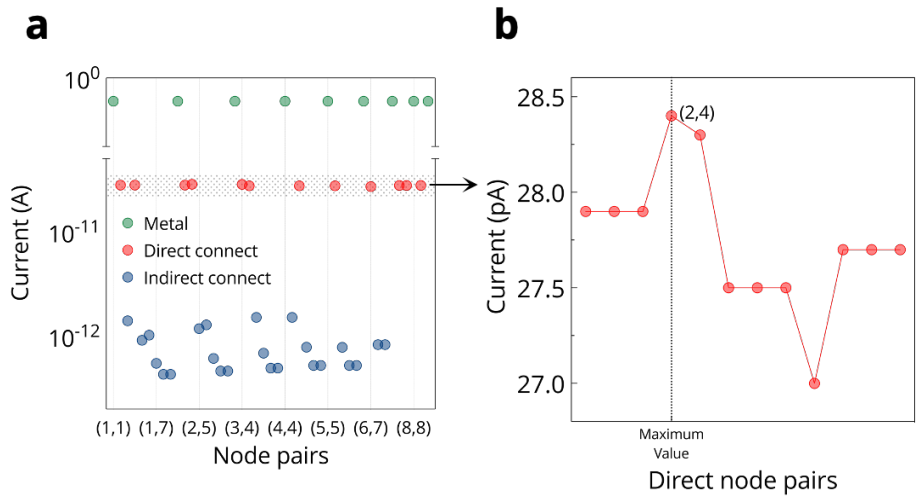


Figure 3-27. SGM-similarity for community detection. a, SGM currents in total 45 node pairs. **b**, SGM currents in 1-hop pairs.

All the estimated SGM-similarity values (1.0 V of reading voltage) between the two nodes in the graph of Figure 3-26 are shown in Figure 3-27a. The numbers in the figure indicate the estimated current values, representing the similarity. The highest current (28.4 pA) value is achieved for $SGM_{2,4}$ (Figure 3-27b), indicating that the connection between nodes 2 and 4 is strongest among others. This strong connection is due to a direct (1-hop) connection between nodes 2 and 4 and the additional 2-hop connection through nodes 1 and 3. Therefore, the first community is formed in pairs (2, 4). In the next step, the similarity between the just-formed community and other nodes is calculated by averaging the similarity between individual nodes within the community and the node [31]. After that, the similarity matrix is updated and repeated until it is grouped into a single community.

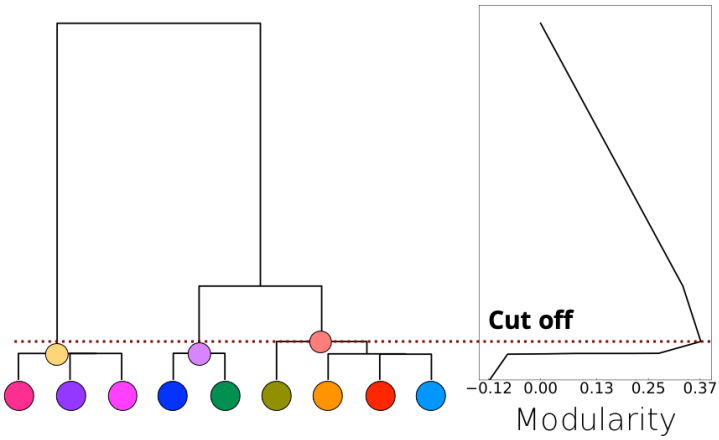


Figure 3-28. A schematic of the dendrogram. The dendrogram can confirm the results of community aggregation according to the progress of the algorithm. (left panel) Modularity changes according to community agglomeration. (right panel).

However, as readily anticipated, the single community containing all members shown in Figure 3-25 does not bear sufficient meaning, so the communities must be cut off at an appropriate branch in the dendrogram that can confirm community agglomeration at each step. This cut-off can be accomplished by estimating the modularity (Figure 3-28) [32]. While no definite value for the optimum modularity is known, a value above 0.3 is considered an appropriate criterion.

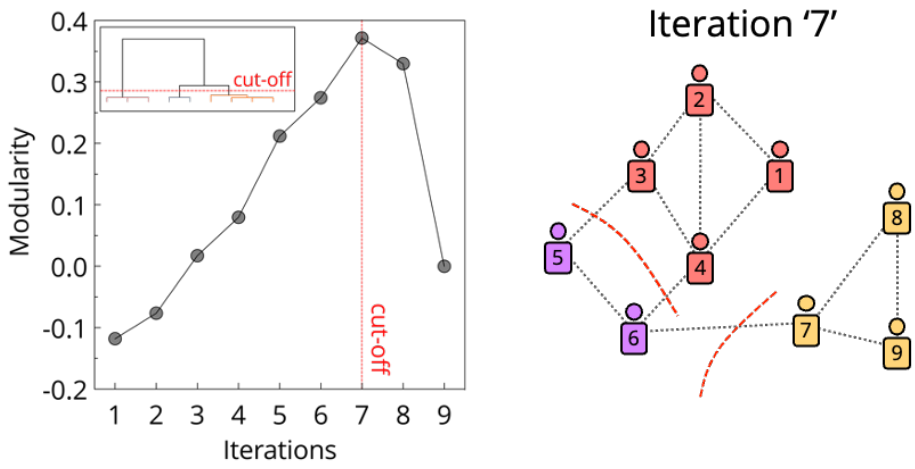


Figure 3-29. The modularity change in each iteration and a schematic of the dendrogram. This result can confirm the results of community aggregation according to the algorithm's progress. (inset) Modularity changes according to community agglomeration. After obtaining the modularity according to the branch formation of the dendrogram, the branch is cut-off at the point corresponding to the highest value (≈ 0.37, at iteration 7). In the inset dendrogram, each bar from right to left corresponds to nodes 1 to 9.

The optimum community cut-off can be found when the modularity reaches the maximum, 0.372 (Figure 3-29). It can be understood that this modularity coincides with the case where the three communities (1, 2, 3, 4), (5, 6), and (7, 8, 9) are formed.

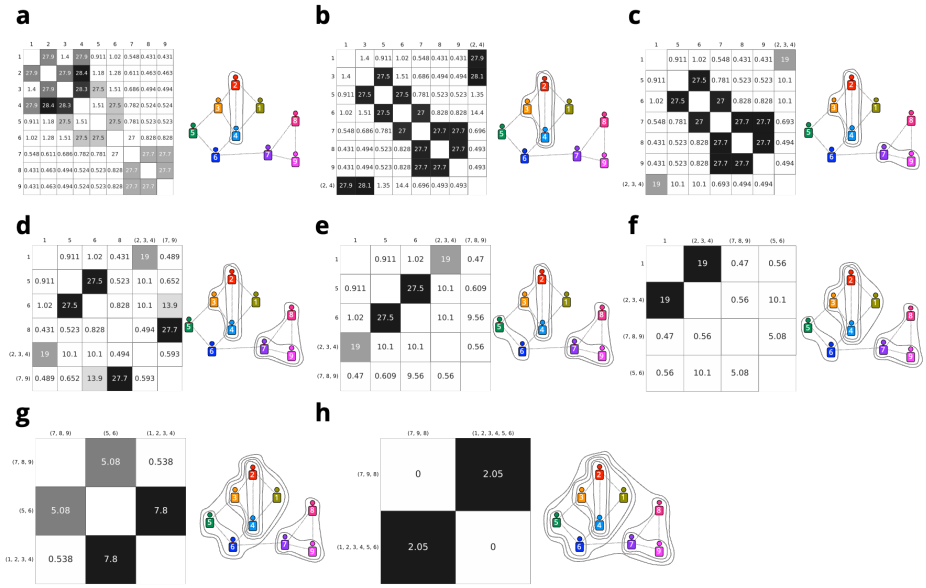


Figure 3-30. The whole process of the SGM-based community detection algorithm. a-f, Similarity matrices and community formation at each iteration step. After initially creating the SGM-similarity matrix, the aggregation is shown in the schematic diagram in the pair with the highest value in the matrix. After the aggregation process, the SGM-similarity matrix is updated by calculating a new similarity between nodes and communities, and between communities according to the UPGMA linkage criteria. Finally, the algorithm is repeated until a single community remains (**h**).

Figure 3-30 shows the community formation process and the resulting similarity matrix. The state with the maximum modularity corresponds to Figure 3-30h.

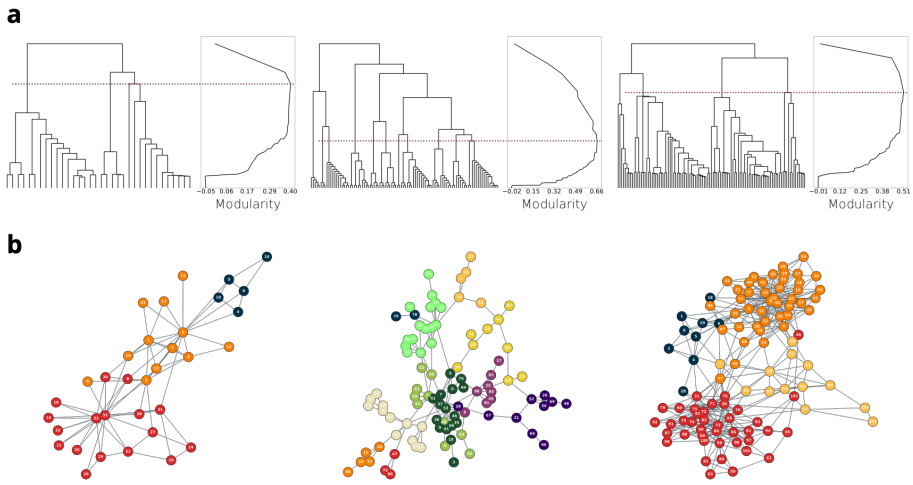


Figure 3-31. Algorithm performance evaluation results using various graph data. a, The dendrogram plot according to the sequential community agglomeration in Zachary's karate club, Twitter retweet network, and Books about US politics dataset, and the modularity calculated at each branch of the dendrogram. **b,** Schematics of community detection results at points with maximum modularity.

Figure 3-31 shows the results of SGM matrix-based community formation in several datasets. Each graph data is displayed in a different color for each community and shows that the cluster is well detected.

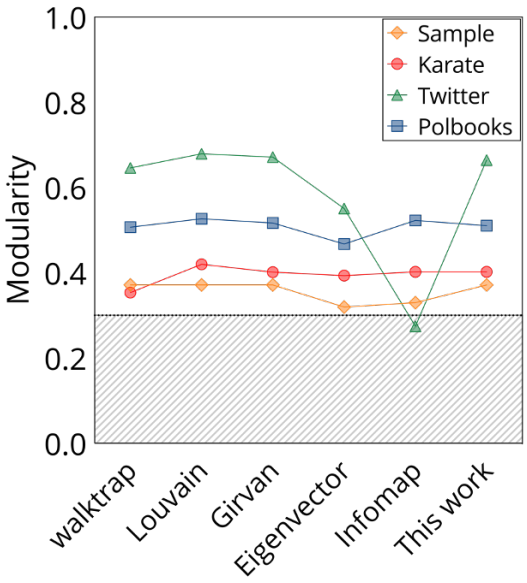


Figure 3-32. The maximum modularity of the SGM-based method was compared with conventional community detection algorithms.

The performance of the SGM-based suggested method was compared with conventional community detection algorithms for various datasets (Figure 3-32). In general, the SGM-based method always belongs to the group with the highest modularity, demonstrating the higher performance of the suggested method.

3.3.5. BRAIN NETWORK-BASED ADHD CLASSIFICATION

To further highlight the strength of mCBA proposed in this study, mCBA-based brain network (connectome) analysis and attention-deficit/hyperactivity disorder (ADHD) diagnosis are performed. A connectome is a brain map that comprehensively expresses the connections of neurons in the brain. Each region of the human brain interacts structurally and functionally at multiple levels and modes [33], [34]. The connectome analysis is essential because it provides information about the brain and psychiatric disorders. However, human connectomes are highly complex and vast, making it challenging to use conventional image analysis techniques. This study generates connectomes from functional magnetic resonance imaging (fMRI) scan data from the subjects with ADHD and the healthy subjects (neurotypical controls, NC), and mapped them to mCBA. For this purpose, using only SGM was sufficient.

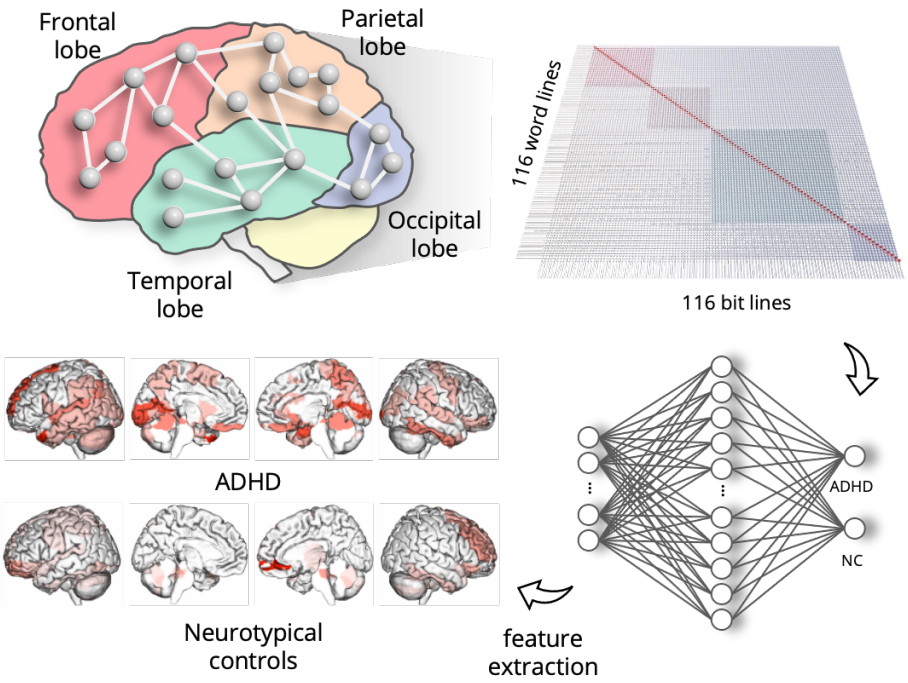


Figure 3-35. A schematic diagram of ADHD classification and identifying ADHD determining brain region based on the brain network analysis using mCBA. The intracortical connections of the brain region are mapped to square areas symmetrical to the main diagonal of the mCBA, and the intercortical connections are mapped between each square. SGM extracts features from the brain network of each subject, and a 2-layer readout network is trained with the SGM vector. Based on the classification result, brain regions where the difference in neural activity was prominent were mapped to the brain figure.

ADHD is diagnosed by effectively extracting the features of each connectome using the SGM method and training the readout network based on SGM current vector. (Figure 3-35)

The SGM current in the mCBA can quickly identify the link in the connectome and classify the connection's hop number (distance) according to the current level. In this study, SGM current data are obtained from the mapped mCBA to confirm connectivities of the connectomes in the ADHD and NC subjects.

Among the distributions of ADHD and NC subjects for various pairs, pairs with the most separated (high AUC) data from the two groups are selected as determining pairs. Determining pairs are classified into 1, 2, and 3-hop according to the SGM current level and sorted based on the AUC calculated from each distribution.

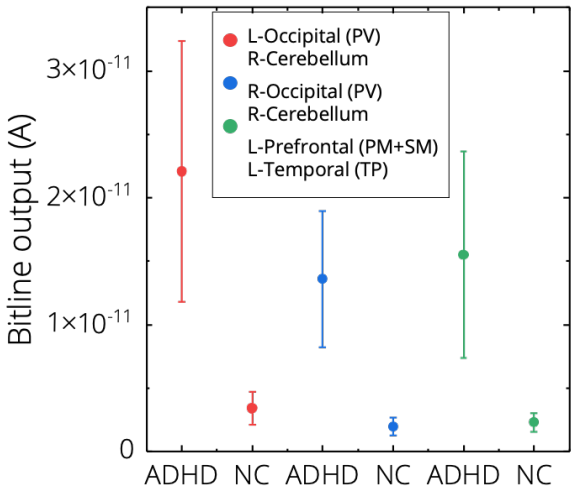


Figure 3-34. SGM current distribution of ADHD and NC subjects in three determining pairs with AUC greater than 0.8.

Figure 3-34 shows the SGM current (1.0 V of reading voltage) distribution in pairs of L-Occipital-PrimVisual (17) – R-Cerebellum-Cerebellum, R-Occipital-PrimVisual (17) – R-Cerebellum-Cerebellum, and L-Prefrontal-PreMot+SuppMot (6) – L-Temporal-Temporalpole (38), which are the 2-hop pairs among the determining pairs. The SGM currents from ADHD and NC subjects are significantly separated in these three pairs. However, several other pairs showed notable differences between the two subjects, which can also be used in classifying ADHD and NC subjects (Figure 3-35 for more details about the network formation).

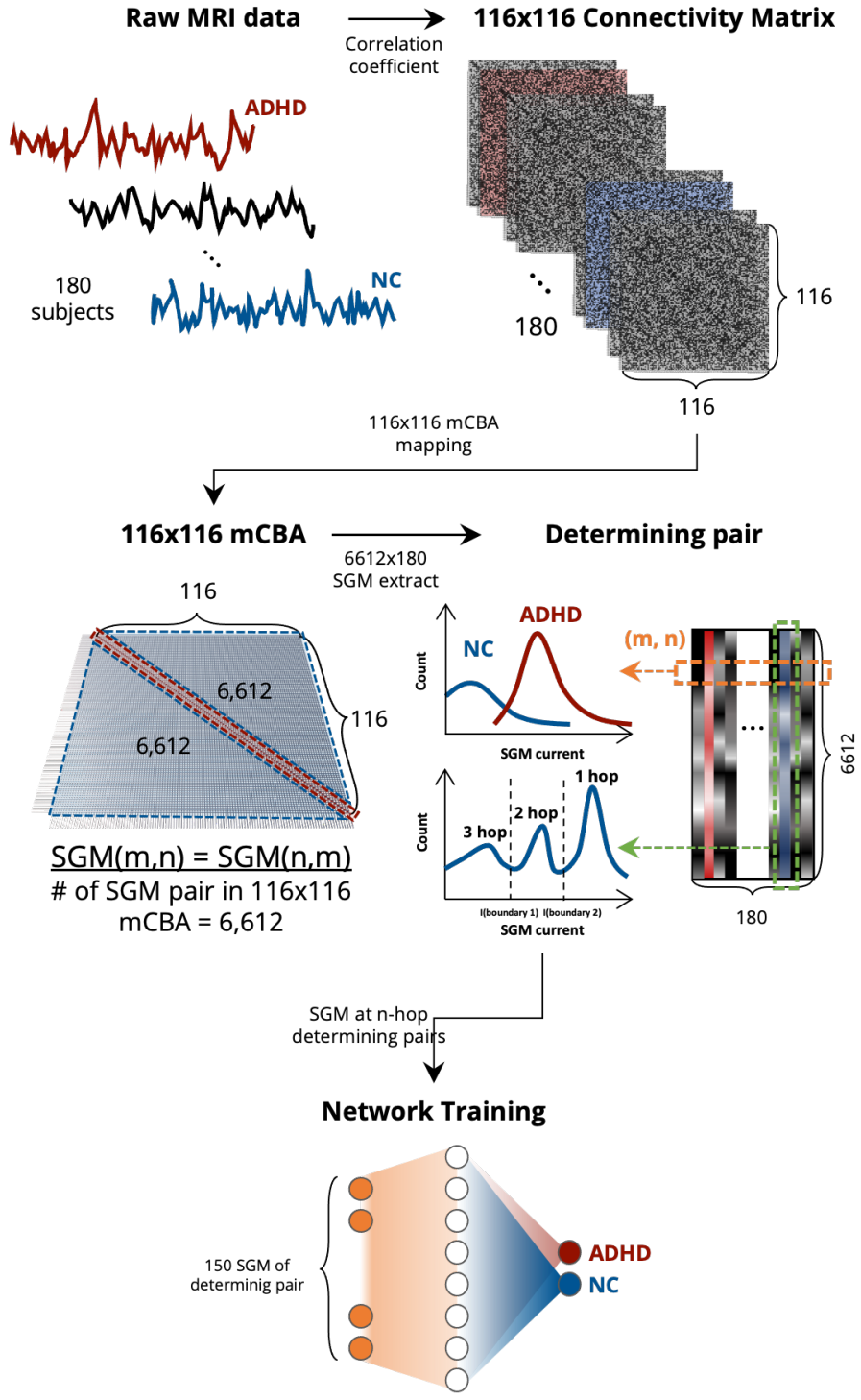


Figure 3-35. Flow chart of the entire process of ADHD classification using

mCBA. Connectivity matrices are obtained by calculating

correlation coefficients after the parcellation of raw fMRI data. The connectivity matrix is mapped to mCBA, and 6612x1 SGM current vector is generated in each brain network. Among the 6,612 components in the given SGM vectors of the training sets (180 subjects), the 150 determining pairs that distinguish ADHD and NC were selected and used as the input vector to train the feedforward network. The hop number can be identified according to the current level.

The pairs containing the crucial information were selected, and their SGM current value vectors (150 Components) were used as the input vectors to train the fully-connected feedforward readout network. The network could be trained well when multi-hop pairs with high AUC were used as inputs. The accuracy was low when the 1-hop, 2-hop, and 3-hop pairs of the determining pair were individually trained.

However, when they were trained together, the accuracy was 77.5% (Figure 3-36a). Therefore, this result indicates that both the intracortical and intercortical connections are different in ADHD and NC subjects. It outperforms many existing algorithms regarding accuracy and AUC. (Figure 3-36b) This is because the information of the multi-hop pair contributed significantly to the accurate diagnosis, which can be extracted efficiently in mCBA. For example, to check the 3-hop connection, matrix multiplication should be performed twice. In the conventional GPU method, it is necessary to

perform MAC operations ~$2.7x10^6$ times for the matrix size used in this task. However, in this mCBA, it is possible to check whether a connection is made by applying a voltage once, which significantly simplifies the computation. Therefore, training the network becomes feasible even for the connectomes with deep links. This connectome processing method can ensure accuracy while inducing significant energy savings. In the mCBA, a deep connection can be identified with one SGM, in which 0.39 pW of power is consumed with a reading voltage of 1 V. In the brain network-based ADHD classification, a maximum of 150 connections were used for training the readout network, and the required power consumption corresponds to 58.5 pW. Meanwhile, in the memristor-CMOS system [35], designed for efficient MAC operation, a power of 1.9 mW is consumed for the MAC operations to check the 3-hop deep connection.

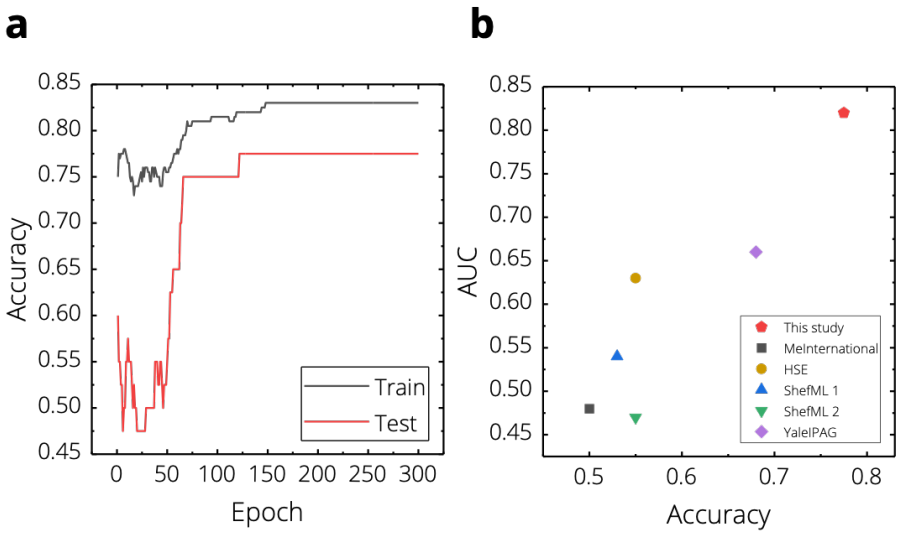


Figure 3-36. Performance of the mCBA-based ADHD classification. a, Train and test accuracy per epoch when SGM current vector of 1-hop, 2-hop, and 3-hop pairs were all used as inputs in ADHD classification. **b,** Accuracy and AUC of SGM-based method and existing studies in ADHD classification.

3.4. Conclusion

Analog computing based on physical means has been an appealing contender to solve several computationally hard problems, such as NP-hard problems. Those problems may not have an algorithmically appropriate solution or take an excessively long time to reach a reasonable answer. The issues dealt with in the above sections correspond to these problems. Memristors have been exploited to apply to specific hardware that may physically solve these problems. Nonetheless, the stochastic nature of the switching mechanism generally hinders the reliable operation of the hardware. The neuromorphic inference machine based on the CBA of memristors, which rapidly processes the vector-matrix multiplication (VMM), is a typical system that suffers from these issues. In contrast, there could be other applications for physical computation, which less or do not suffer from the non-uniformity and repeatability issues.

The mCBA structure in this work performs the physical calculation to realize the similarity and adjacency search functions in the graphic data structures. The graphic data structure can be non-Euclidean, which may not necessarily be transformed into the Euclidean one by even the most complicated method in the vector space. In this case, the known algorithmic solution using the similarity function may not work. However, the SGM in mCBA can extract the similarity, the hidden information, between the nodes in any graph using the

non-ideality of the array structure. This process does not require any pre-processing of the graphic data, even if they are of the non-Euclidean form.

Also, the similarity function is only used for the relative comparison, not the deterministic calculation, and thus, suffers far less from the random variation than the VMM. It is mainly determined by the number of connecting nodes, not by the resistance of each cell. Besides, MGM provides a physical mean for the adjacency search function, which searches the nearby (or directly connected) nodes. Therefore, mCBA can be used for both the similarity and adjacency search functions, corresponding to the process and memory functions, respectively. In other words, the mCBA is an optimized process-in-memory device performing both data process and memory functions.

The mCBA can efficiently identify deep connections in a huge graph, which grows in the real world. Checking multi-hop connections in a large graph requires considerable computation using existing hardware. mCBA can easily extract hidden information (deep connection) in the graph and reduce the analytical complexity of the real-world graph.

This work demonstrates that the various graphic structures can be mapped onto the mCBA, and the physical calculations using the suggested mCBA outperform the previous software-based algorithms. The mCBA can be used for any type of graph, e.g., directed or weighted. Stacking the two-dimensional mCBA or even vertical integration of the

mCBA in three-dimensional space will allow applying it to a multidimensional graphic network.

3.5. References

[1] Kim, K. M. *et al.* Low-Power, Self-Rectifying, and Forming-Free Memristor with an Asymmetric Programing Voltage for a High-Density Crossbar Application. *Nano Lett.* **16**, 6724–6732 (2016).

[2] Kim, G. H. *et al.* 32 × 32 Crossbar Array Resistive Memory Composed of a Stacked Schottky Diode and Unipolar Resistive Memory. *Adv. Funct. Mater.* **23**, 1440–1449 (2013).

[3] Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).

[4] Jang, Y. H. *et al.* Time-varying data processing with nonvolatile memristor-based temporal kernel. *Nat. Commun.* (2021) doi:10.1038/s41467-021-25925-5.

[5] Kim, Y. *et al.* Kernel Application of the Stacked Crossbar Array Composed of Self-Rectifying Resistive Switching Memory for Convolutional Neural Networks. *Adv. Intell. Syst.* **2**, 1900116 (2020).

[6] Yoon, J. H. *et al.* Highly uniform, electroforming-free, and self-rectifying resistive memory in the Pt/Ta2O5/HfO2-x/TiN structure. *Adv. Funct. Mater.* **24**, 5086–5095 (2014).

[7] Kim, J., Woo, H. C., Jeong, T., Choi, J.-H. & Hwang, C. S. In-Depth Analysis of One Selector–One Resistor Crossbar Array for Its Writing and Reading Operations for Hardware Neural Network with Finite Wire

Resistance. *Adv. Intell. Syst.* (2021) doi:10.1002/aisy.202100174.

[8] Flood, M. M. The Traveling-Salesman Problem. *Oper. Res.* (1956) doi:10.1287/opre.4.1.61.

[9] S.Lin & B.W. Kernighan. An effective heuristic algorithm for the traveling-salesman problem. *Oper. Res.* (1973).

[10] Myers, S. A., Sharma, A., Gupta, P. & Lin, J. Information network or social network? The structure of the twitter follow graph. in *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web* (2014). doi:10.1145/2567948.2576939.

[11] Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nat. Commun.* (2019) doi:10.1038/s41467-019-09177-y.

[12] Albert, R., Jeong, H. & Barabási, A. L. Diameter of the world-wide web. *Nature* (1999) doi:10.1038/43601.

[13] Kim, K. H. Graph theory and its applications to problems of society. *Math. Soc. Sci.* (1981) doi:10.1016/0165-4896(81)90012-3.

[14] Barnes, J. A. & Harary, F. Graph theory in network analysis. *Soc. Networks* (1983) doi:10.1016/0378-8733(83)90026-6.

[15] Qiao, L., Zhang, L., Chen, S. & Shen, D. Data-driven graph construction and graph learning: A review. *Neurocomputing* **312**, 336–351 (2018).

[16] Mason, O. & Verwoerd, M. Graph theory and networks in biology. *IET Systems Biology* at https://doi.org/10.1049/iet-syb:20060038 (2007).

[17] Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing

networks in cell biology. *Briefings in Bioinformatics* at https://doi.org/10.1093/bib/bbl022 (2006).

[18] Pohl, I. Heuristic search viewed as path finding in a graph. *Artif. Intell.* (1970) doi:10.1016/0004-3702(70)90007-X.

[19] Dabaghi Zarandi, F. & Kuchaki Rafsanjani, M. Community detection in complex networks using structural similarity. *Phys. A Stat. Mech. its Appl.* (2018) doi:10.1016/j.physa.2018.02.212.

[20] Lü, L., Jin, C. H. & Zhou, T. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* (2009) doi:10.1103/PhysRevE.80.046122.

[21] Goyal, P. & Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Syst.* (2018) doi:10.1016/j.knosys.2018.03.022.

[22] Cai, H., Zheng, V. W. & Chang, K. C. C. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Trans. Knowl. Data Eng.* (2018) doi:10.1109/TKDE.2018.2807452.

[23] Rossiter, P. L. & Bass, J. The Electrical Resistivity of Metals and Alloys . *Phys. Today* (1988) doi:10.1063/1.2811462.

[24] Goldberg, A. & Harrelson, C. *Computing the Shortest Path: A* Search Meets Graph Theory*. https://www.microsoft.com/en-

us/research/publication/computing-the-shortest-path-a-search-meets-graph-theory/ (2004).

[25] Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* (1959) doi:10.1007/BF01386390.

[26] Gupta, P. *et al.* WTF: The Who to Follow service at Twitter. in *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web* (2013).

[27] Shahmohammadi, A., Khadangi, E. & Bagheri, A. Presenting new collaborative link prediction methods for activity recommendation in Facebook. *Neurocomputing* (2016) doi:10.1016/j.neucom.2016.06.024.

[28] Lü, L. *et al.* Recommender systems. *Physics Reports* at https://doi.org/10.1016/j.physrep.2012.02.006 (2012).

[29] Shaikh, S., Rathi, S. & Janrao, P. Recommendation system in E-Commerce Websites: A graph based approached. in *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017* (2017). doi:10.1109/IACC.2017.0189.

[30] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* (1967) doi:10.1007/BF02289588.

[31] Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* (1958).

[32] Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E - Stat. Physics, Plasmas,*

Fluids, Relat. Interdiscip. Top. (2004) doi:10.1103/PhysRevE.70.066111.

[33] Sporns, O. The human connectome: A complex network. *Annals of the New York Academy of Sciences* at https://doi.org/10.1111/j.1749-6632.2010.05888.x (2011).

[34] Fornito, A., Zalesky, A. & Breakspear, M. The connectomics of brain disorders. *Nature Reviews Neuroscience* at https://doi.org/10.1038/nrn3901 (2015).

[35] Cai, F. *et al.* A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. *Nat. Electron.* **2**, 290–299 (2019).

4. Conclusion

In this dissertation, complex data processing with memristor-based physical computing was established using intrinsic physical properties (R-C delay, I-V nonlinearity, sneak current) of the memristive hardware.

First, a new method of sequential data processing using a nonvolatile memristor-based temporal kernel with time constants controllability was proposed. A temporal kernel was constructed using memristors (M), resistors (R), and capacitors (C) for effective sequential data processing. The unit cell has a 1M1R1C structure in which a memristor is connected in series with a resistor and a capacitor, and the resistor and capacitor are connected in parallel with each other. The 1M1R1C kernel has the advantage of being applicable to various situations as it can have various time constants through R and C control. 1M1R1C-based MNIST recognition showed high accuracy (90%) with high energy efficiency and fast processing speed. In addition, the 1M1R1C kernel was applied to ultrasound and electrocardiogram-based medical diagnosis with very different time constants (frequency range of 1 to 10 MHz).

Second, a method for processing non-Euclidean graphs using self-rectifying memristor arrays was proposed. Non-Euclidean graphs were represented using a metal-cell-at-diagonal crossbar-array (mCBA), made up of self-rectifying memristors. The mCBA's sneak current, a natural physical property, can be used to determine similarity. The sneak current-based similarity function can

be used to measure the distance between nodes, connections between communities and nodes, and the likelihood of unconnected nodes becoming connected in the future. This research demonstrates the practical use of memristor-based physical calculations for solving various types of graph-related problems.

This dissertation presents a new breakthrough for the next-generation physical computing using memristor-based novel hardware for complex data, such as time-varying data and graph data. The results in this thesis could shed light on this novel data processing field by suggesting a new pathway that is a step forward from the conventional approach.

Curriculum Vitae

Yoon Ho Jang

Department of Materials Science and Engineering

College of Engineering

Seoul National University

1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

E-mail: dbsgh0147@snu.ac.kr

Tel.: +82-10-5651-8839

Fax.:+82-2-880-8643

I. Educations

2014. 03. - 2018. 08. B.S.

Department of Materials Science and Engineering

Seoul National University, Seoul, Korea

2018. 09. – 2023. 02 Integrated M. S. & Ph. D.

Department of Materials Science and Engineering

Seoul National University, Seoul, Korea

II. Research Areas

1. Analysis of the switching mechanism of the ReRAM

- Characterization of electronic properties of ReRAM with MIM structure
- Characterization of electronic properties of ReRAM CBA using multi-point probing system
- Time-dependent measurement of ReRAM with parasitic components
- Finite-element method simulation of ReRAM structure using physical attributes (Electric field, Joule heating)

2. Compact modeling of ReRAM devices for simulation

- Extraction of ReRAM modeling parameters based on the measurement results
- Designing ReRAM compact model for analog-circuit simulators

3. Resistive switching memory device fabrication and electrical measurements

- Deposition / etching / lithography / electrical characterizations

- Studies on the switching mechanism of resistive switching memory device

4. Data processing with the memristor-based physical computing system

- Multi-layer perceptron and conventional neural network simulation using Graphical Processing Unit
- Temporal data processing with memristive temporal kernel
- Chaotic time-series prediction
- Graph data processing with self-rectifying memristor crossbar array

III. Experimental Skills

1. Deposition methods

- Atomic layer deposition for oxide (Hafnium oxide, Aluminium oxide, Tantalum oxide, Titanium oxide)
- DC & RF magnetron sputtering and E-beam evaporation for electrode materials

2. Annealing methods

- Rapid thermal process

3. Analysis methods

- X-ray photoelectron spectroscopy (XPS, UK VG, Sigma Probe) for analysis of the chemical states of the elements

- Transmission Electron Microscopy (TEM, JEOL, JEM-2100F, JEM-3000F, JEM-200CX) for microstructure analysis of thin film

- Energy Dispersive Spectroscopy (EDS, Oxford Instrument, AZtec) incorporated by Scanning Transmission Electron Microscopy (STEM, JEOL, JEM-2100F, JEM-3000F) for elemental analysis.

- Auger electron spectroscopy (AES, ULVAC-PHI, PHI-700) for analysis of impurity.

- Scanning electron microscopy (SEM, Hitachi, S-4800) and Atomic Force Microscopy (AFM, JEOL, JSPM-5200) for analysis of the topography

- Spectroscopic Ellipsometer (SE, J.A. Woollam, M-2000) for analysis of optical properties and thicknesses of thin films

- Pulse/pattern generator (Agilent, 81110A/81111A) and digital oscilloscope for pulse switching measurements / B1500A with WGFMU and RSU units for faster measurements

4. Programs apprentice

- HSPICE (Synopsys) for analog circuit simulation
- Pytorch with Python for neural network simulation
- MATLAB (Mathworks) for general calculation and analysis
- COMSOL (COMSOL) for simulation with the finite-element method

IV. Academic Honors

1. Excellent Paper Award, BK21 4th phase, Department of Materials Science and Engineering, College of Engineering, Seoul National University (June 2022)
2. Excellent Graduate Students Award, BK21 4th phase, Seoul National University (June 2022)
3. Excellence Award, AI Model Presentation, Korea Research Institute of Standards and Science (August 2022)
4. 17th Semiconductor Scholarship, Korea Semiconductor Industry Association (October 2022)

List of publications

1. Refereed Journal Articles (SCI)

1.1 Domestic

1.2. International

1. **Yoon Ho Jang,** Woohyun Kim, Jihun Kim, Kyung Seok Woo, Hyun Jae Lee, Jeong Woo Jeon, Sung Keun Shim, Janguk Han, and Cheol Seong Hwang*, "Time-varying data processing with nonvolatile memristor-based temporal kernel", Nature Communications, 12, 5727 (2021)
2. Seung Kyu Ryoo, Kyung Do Kim, Hyeon Woo Park, Yong Bin Lee, Suk Hyun Lee, In Soo Lee, Seungyong Byun, Doosup Shim, Jae Hoon Lee, Hani Kim, **Yoon Ho Jang**, Min Hyuk Park, and Cheol Seong Hwang*, "Investigation of Optimum Deposition Conditions of Radio Frequency Reactive Magnetron Sputtering of Al0.7Sc0.3N Film with Thickness down to 20 nm"**,** Advanced Electronic Materials, 2200726 (2022)
3. Kyung Seok Woo, Jaehyun Kim, Janguk Han, Woohyun Kim, **Yoon Ho Jang** & Cheol Seong Hwang, "Probabilistic computing using $Cu_{0.1}Te_{0.9}/HfO_2/Pt$ diffusive memristors"**,** Nature Communications, 13, 5762 (2022)
4. Hyeon Woo Park, Minsik Oh, In Soo Lee, Seungyong Byun, **Yoon Ho Jang**, Yong Bin Lee, Beom Yong Kim, Suk Hyun Lee, Seung Kyu Ryoo, Doosup Shim,

Jae Hoon Lee, Hani Kim, Kyung Do Kim, and Cheol Seong Hwang*, “Double S-Shaped Polarization – Voltage Curve and Negative Capacitance from Al_2O_3-$Hf_{0.5}Zr_{0.5}O_2$-Al_2O_3 Triple-Layer Structure”**,** Adv. Funct. Mater. 2206637, (2022)

5. **Yoon Ho Jang**, Janguk Han, Jihun Kim, Woohyun Kim, Kyung Seok Woo, Jaehyun Kim, and Cheol Seong Hwang, “Graph analysis with multi-functional self-rectifying memristor array”, Advanced Materials, 2209503 (2022)

2. CONFERENCES

2.1 Domestic

1. **Yoon Ho Jang**, Jihun Kim, Jaehyun Kim, and Cheol Seong Hwang, “Analysis Of Multi-bit Resistive Switching Of W/HfO2/TiN Memristor Based On Electronic Bipolar Resistive Switching Mechanism”, The 27th Korean Conference on Semiconductors (February 2020), Poster
2. Woohyun Kim, Manick Ha, Chanyoung Yoo, Jeong Woo Jeon, Wonho Choi, Byongwoo Park, Gil Seop Kim, Kyung Seok Woo, Jihun Kim, **Yoon Ho Jang**, Eui-Sang Park, Yoon Kyeung Lee, and Cheol Seong Hwang, Atomic Layer Deposited N-doped GeSe for Leaky-Integrate-and-Fire Neuron Application, The 28th Korean Conference on Semiconductors (January 2021), Oral
3. **Yoon Ho Jang**, Ji Hun Kim, Jeong Woo Jeon, Woo Hyun Kim, and Cheol Seong Hwang, Memristive Reservoir Computing for Medical Diagnosis, The 28th Korean Conference on Semiconductors (January 2021), Oral

4. Jang Uk Han, **Yoon Ho Jang**, and Cheol Seong Hwang, Analysis of the Role of the Al2O3 Layers in Self-Rectifying and FormingFree Pt/Al2O3/HfO2/Al2O3/TiN Memristor, The 29th Korean Conference on Semiconductors (January 2022), Poster
5. Sung Keun Shim, **Yoon Ho Jang**, and Cheol Seong Hwang, Demonstration of Adaptable Artificial Nerve Using 2Memristor-1Capacitor Structure, The 29th Korean Conference on Semiconductors (January 2022), Poster
6. **Yoon Ho Jang**, Janguk Han, and Cheol Seong Hwang, Demonstration of Sneak Current-Based A* Pathfinding Algorithm, The 29th Korean Conference on Semiconductors (January 2022), Oral
7. Jang Uk Han, **Yoon Ho Jang**, and Cheol Seong Hwang, Analysis of sneak current through n cell in the memristive-crossbar array, The 30th Korean Conference on Semiconductors (February 2023), Poster
8. Sung Keun Shim, **Yoon Ho Jang**, and Cheol Seong Hwang, Time-Series Data Processing using 2Memristor-1Capacitor Integrated Temporal Kernel, The 30th Korean Conference on Semiconductors (February 2023), Poster
9. **Yoon Ho Jang**, Janguk Han, and Cheol Seong Hwang, Graph analysis using self-rectifying memristor crossbar array, The 30th Korean Conference on Semiconductors (February 2023), Poster

2.2 International

1. Woohyun Kim, Manick Ha, Chanyoung Yoo, Jeong Woo Jeon, Wonho Choi, Byongwoo Park, Gil Seop Kim, Kyung Seok Woo, Jihun Kim, **Yoon Ho Jang**, Eui-Sang Park, Yoon Kyeung Lee, and Cheol Seong Hwang, "High-Reliable Atomic Layer Deposited N-doped GeSe and Its Leaky-Integrate-and-Fire Neuron Application", AVS 21st International Conference on Atomic Layer Deposition, Virtual Meeting, June 27-30, 2021, Poster
2. Jang Uk Han, **Yoon Ho Jang**, Ji Hun Kim, Woo Hyun Kim, and Cheol Seong Hwang, "Demonstration of a diagonal shorted self-rectifying memristive crossbar array for performing graph algorithms", The 2022 E-MRS Fall Meeting, Warsaw University of Technology, Sep 19 to 22, Poster
3. **Yoon Ho Jang**, Sung Keun Shim, Janguk Han, Jihun Kim, Woohyun Kim, and Cheol Seong Hwang, "Time series data processing using non-volatile memristor-based temporal kernel", The 2022 E-MRS Fall Meeting, Warsaw University of Technology, Sep 19 to 22, Poster
4. Sung Keun Shim, **Yoon Ho Jang,** Janguk Han, Jeong Woo Jeon, and Cheol Seong Hwang, "Energy-Efficient Time-Series Data Processing Using HfO_2-Based 2Memristor-1Capacitor Integrated Temporal Kernel", 2022 MRS Fall Meeting & Exhibit, Boston, Nov 27 to Dec 2, Poster

Abstract (in Korean)

최근 deep learning의 대두로 다양한 data들이 축적되고 학습에 사용되었다. Big data는 그 구조가 더욱 다양화되고 복잡해지면서 기존 하드웨어로 처리하기 힘든 complex data 가 등장했다. Complex data의 예시로는 Sequential data, graph data 가 있다. Sequential data는 현재 스테이트가 인풋 히스토리를 반영하면서 그 패턴이 일정하지 않고 예측하기 힘든 특성이 있다. 그래프 타입의 데이터는 주체와 주체간의 연결성들을 다루기에 vector 형태로 표현되기 어려워, 기존 하드웨어 구조에서 처리하기 힘들다는 문제가 있다. 이런 복잡한 data를 처리하기 위해서는 novel data processing technique이 요구된다.

본 연구의 첫번째 파트에서는, 효과적인 시퀀셜 데이터 처리를 위해서 멤리스터, 리지스터, 캐패시터를 이용해 temporal kernel 을 구성하였다. 전체적인 컴퓨팅 스킴은 conventional reservoir system 과 동일하여 input 이 temporal kernel 에서 처리된 데이터가 멤리스터에 저장된다. 이후 이러한 멤리스터 컨덕턴스 벡터를 인풋으로 readout network 를 학습시킨다. 유닛셀은 멤리스터가 리지스터, 캐패시터와 직렬연결되어 있고 리지스터와 캐패시터는 서로 병렬 연결되어 있는 1M1R1C 구조를 가진다. 1M1R1C kernel은 R, C 조절을 통해 다양한 time constant 를 가질 수 있어 다양한 상황에 적용가능하다는 장점이 있다. 본 연구에서는 1M1R1C 기반 MNIST recognition에서 높은 에너지 효율과 빠른 처리속도로 높은 정확도(90 %)를 보였다. 한편 1M1R1C kernel은 시간 상수가 매우 다른 ultrasound, electrocardiogram 기반 medical diagnosis에도 적용되어 1 ~ 10 MHz 의 넓은 주파수 영역에서 성공적으로 task를 수행하였다.

본 연구의 두번째 파트에서는, 자가정류 멤리스터 어레이를 이용해

비유클리드 그래프를 처리하는 방법이 다뤄진다. 비유클리드 그래프에서는 유사도를 구할 수 없어 그래프 임베딩 등의 복잡한 전처리 과정이 요구되며 그 과정에서 정보의 손실도 발생한다. 본 연구에서는 비유클리드 그래프를 벡터화하지 않고 그 본래 데이터 그대로 맵핑하고 분석하는 방법을 제안한다.

주요어: 저항변화 메모리, ReRAM, 메모리, 하프늄 옥사이드, HfO_2, 자가정류 멤리스터, 컴플렉스 데이터, 커널, 시간 커널, 시계열 데이터, 의료 진단, 크로스바 어레이, 누설전류, 그래프 알고리즘, 프로세스 인 메모리

학번: 2018 – 24630

장 윤 호