An Artificial Neutral Network Chip Based on Two-Dimensional Semiconductor

Shunli Ma
Fudan University
Tianxiang Wu
Fudan University
Xinyu Chen
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
Yin Wang
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
Hongwei Tang
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433
Yuting Yao
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
Yan Wang
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
Jianan Deng
Fudan University
Jing Wan
State Key Laboratory of ASIC and System, School of Information Science and Engineering, Fudan University, Shanghai 200433
Zhengzong Sun
Fudan University
Zihan Xu
Shenzhen 6 Carbon Technology, Shenzhen, China
Antoine Riaud
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433
Chenjian Wu
School of Electronic and Information Engineering, Soochow University, Suzhou, China

Yang Chai
Hong Kong Polytechnic University https://orcid.org/0000-0002-8943-0861
Peng Zhou
Fudan University https://orcid.org/0000-0002-7301-1013
Junyan Ren
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China
Wenzhong Bao (baowz@fudan.edu.cn)
State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433 https://orcid.org/0000-0002-3871-467X
Wei Zhang
Fudan University

Article

Keywords: Integrated Circuits, Field Effect Transistors, Convolutional and Activation Function Circuits, Wafer-scale Processing

Posted Date: November 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-101486/v1

An Artificial Neutral Network Chip Based on Two-Dimensional Semiconductor

Shunli Ma[1]†, Tianxiang Wu[1]†, Xinyu Chen[1]†, Yin Wang[1], Hongwei Tang[1], Yuting Yao[1], Yan Wang[1], Jianan Deng[2], Jing Wan[2], Zhengzong Sun[1], Zihan Xu[3], Antoine Riaud[1], Chenjian Wu[4], Yang Chai[5], Peng Zhou[1]*, Junyan Ren[1]*, Wenzhong Bao[1]*, David Wei Zhang[1]

[1] *State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China*

[2] *State Key Laboratory of ASIC and System, School of Information Science and Technology, Fudan University, Shanghai 200433, China*

[3] *Shenzhen Sixcarbon Technology, 188 Jiangshi Road, Shenzhen 518106, China*

[4] *School of Electronic and Information Engineering, Soochow University, Suzhou, China*

[5] *Department of Applied Physics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, P. R. China*

† These authors contributed equally to this work.

* Correspondence to: pengzhou@fudan.edu.cn; junyanren@fudan.edu.cn; baowz@fudan.edu.cn

Abstract

Two-dimensional semiconductors can be used to build integrated circuits for running artificial neural networks (ANN) with higher energy efficiency. The implementation of an ANN with 2D semiconductors has been held back by the large-scale and high-quality transistors required for running machine learning algorithms. Here we demonstrate the first functional MoS_2 analog ANN integrated circuit, including memory, multiply-and-accumulate (MAC), activation function, and weight update circuits. The ANN integrated circuit is realized through 818 field effect transistors (FETs) with wafer-scale and high-homogeneity MoS_2 film. The large current on/off ratio and output linearity of these MoS_2 FETs allow the realization of convolutional and activation function circuits with a few number of transistors. This ANN can be used for recognizing tactile digit, showing the recognition rate exceeding 97%. Our work demonstrates wafer-scale processing of a 2D semiconductor for building integrated circuits with the functions of AI computation.

Two-dimensional (2D) layered semiconducting transition metal chalcogenides (TMDs) have become a promising material family for building circuitry owing to their outstanding electronic transport properties, sensing capabilities and mechanical compliance. In particular, monolayer MoS_2 has been widely investigated as a channel material for use in field-effect transistors (FETs) [1-3], mainly because of its 1.8 eV direct bandgap that gives rise to reasonably high mobility and large on/off ratio. These characteristics make it suitable for logic circuits [4-7], analog circuits [8,9], memories[10], and photodetectors [11-39]. Moreover, the environment-sensitive and flexible characteristics[40,41] make it a promising material for light [36-39], temperature [42], gas [43-46], and pressure sensing [47-49] for flexible electronics, such as wearable devices [50] and electronic skins [42,51]. Meanwhile, the weak interlayer van der Waals forces allow different 2DLM films to be vertically transferred and stacked. Therefore, it is possible to fabricate three-dimensional (3D) monolithic integrated circuits by stacking multilayer 2DLMs with different functions [42,52,53], which may enable energy efficient smart prosthesis (*e.g.*, artificial retina, nose, and skin, as shown in Fig.1a). Unlike sensors, the development of such sensing-computing devices is very challenging due to the large number of required transistors, accurate models, and various types of circuits[54] needed for signal amplification and data processing.

Natural sensing-computing is provided by sensing neurons and the neurons located next to them. Similarly, integrating preprocessing directly on a sensor integrated circuit can significantly improve the speed and efficiency of the irregular signal processing in various types of sensors [55]. Integrating these functions into a single package also ensures power consumption is lower than would be seen in discrete components. Given the tremendous success of artificial neural networks (ANN) for image processing and pattern recognition [14,22,29,34,56], ANNs are foreseen to be key models used in individual sensors and an overall system (Fig. 1b). Compared with digital ANN, analog ANN integrated circuits are more suitable for preprocessing tasks due to their lower power consumption and higher throughout.

2D semiconductors are a potential material for building ANNs at the hardware level. First, semiconducting TMDs grown on insulating substrates have a similar structure as silicon on insulator (SOI), which provides low gate-induced drain leakage (GIDL) and substrate leakage current. Thus, for TMDs with a relatively large bandgap like MoS_2, a device with low leakage current can have a lower data refresh rate, and less energy

would be required for information storage than in CMOS devices, such as in dynamic random-access memory (DRAM). Experimental results also suggest that MoS_2 transistors show less 1/f noise than Si devices fabricated with the CMOS process[57-59]. Under the same signal energy, the signal-to-noise ratio is sufficiently large that analog convolution calculations can be performed. Moreover, by choosing different voltage ranges, a desirable relationship between the measured current and applied voltage can be obtained when implementing analog matrix computation[12]. Therefore, TMDs are ideal transistor candidates for implementing ANNs in application-specific integrated circuits[60], as they offer an advantageous combination of established sensing capabilities and outstanding transistor performance.

However, there are still obstacles to overcome before the use of 2DLMs becomes practical. (*1)* Most of the previously demonstrated devices are implemented using exfoliated 2DLM sheets with micrometer size, which limits the circuit scale. Therefore, reliable wafer-scale material growth is under investigation[61]. *(2)* Processing wafer-scale 2D materials integrated circuits is still in the early stage, especially gate-last technology with accurate control over the threshold voltage V_T. Additionally, an accurate device model for large-scale circuits still requires development. *(3)* An ANN requires a higher noise budget and linearity than in a digital circuit, thus a more accurate device model is needed. Analog convolutional neural network (CNN) design also requires additional circuits. A complete functional CNN requires various circuits modules, such as circuits for activation functions, memory, and the multiply-and-accumulate (MAC) operation [14,22,34,56,62].

In this work, we develop a gate-last processing technique for wafer-level MoS_2 films and build a chip for general ANN processing (Fig. 1c), which can be used for future smart sensing applications. The analog MAC can be realized on a dual-gate MoS_2 FET. A simple charge storage structure is formed by one MoS_2 transistor and one capacitor (1T-1C), which acts as an analog random-access memory (a-RAM). This chip provides all fundamental functions required in AI computation: convolution calculation, memory, activation function, and integrated weight updating (Fig. 1d), which are all formed by 2D MoS_2. A trained ANN is was used for tactile letter recognition with the accuracy greater than 97%. Thus, a 2DLM-based ANN makes it possible to fabricate a sensing-computation system beyond silicon.

Circuit Architecture

An ANN model requires complex peripheral circuits. Previous research into ANNs implemented in integrated circuits focused on devices built from memristors or flash memories [12,62] that can perform analog MAC operations[11,14,22,29]. However, most results lacked activation function and weight update circuits or were complemented by peripheral silicon circuits, without which a neural network is merely a linear regression model. In this paper, we show that an ANN system that implements on convolutional in a MoS_2-FET can also integrate MoS_2-a-RAM to implement both activation and weight update functions.

Taking biological neurons as an example, neurons have multiple synapses which transmit sensed signals to soma. Each synapse can store and change the corresponding weight of the sensed signal, and a multiplication operation between the sensed signal and weight can be achieved. A soma then can realize an accumulation of multiplication results from all synapses. When the summation is larger than the threshold, the soma will be activated and generate an output through an axon. A mathematical model of a biological neuron is shown in Fig. 2a, where w_0, $w_1 \ldots w_{10}$ are synaptic weights, and x_0, $x_1 \ \ldots \ x_{10}$ are bio-electric signals. Σ is the integration operation in the soma and f indicates the activation function that gives the output. The corresponding device model of one neuron is shown in Fig. 2b**.** Its input signals come from various sensors, the MoS_2 FETs perform a MAC operation, whose result is then transmitted into an activation function circuit. The weight values are stored in MoS_2 a-RAM circuits. A complex ANN with multiple neurons is designed as shown in Fig. 2c. The proposed MoS_2 ANN circuit consists of an input layer with 6 neurons, a fully-connected layer, and an output layer with 10 neurons. In each neuron, the MAC result is sent to a nonlinear activation function circuit that maps the output results and normalizes them to within a specified range (0 to 1). The fully-connected layer connects the input and output layers. Finally, an off-chip logical function classifier is used to label sensed targets. Weight update and optimization are performed using the back propagation (BP) algorithm[63] and implemented with a DAC whose digital inputs come from off-chip control. The weight update circuits, a-RAM circuits, and activation function circuits are integrated into our MoS_2 chip. Compared to previously reported RRAM devices, our proposed circuit integrates weight storage and update in local memory, which significantly simplified the system design for ANN.

Wafer-scale MoS_2 film synthesis and device processing

Wafer-scale ANN circuits require high-quality 2DLM thin films. While wafer-scale 2DLM synthesis by CVD has been recently demonstrated [5,6,64], other challenges impede practical application, including processing of electrical contacts and dielectric layer deposition due to atomic thickness and complex interface conditions. Here, based on the high quality continuous MoS_2 film we previously reported [65-67] (Fig. 3b, also see Fig. S1), a gate-last technology is developed for MoS_2 FETs with the manufacturing processing flow shown in Fig. 3b. For the subsequent simulations, the device parameters were fitted to a Gaussian distribution. The average mobility from 380 MoS_2 TG-FETs was found to be 27.2 $cm^2 \cdot V^{-1} \cdot s^{-1}$ with an average threshold voltage of V_T = 1.17 V, as shown in Fig. 3c-d. Compared with the previously reported gate-first processing-based MoS_2-FETs [68-71], our MoS_2 TG-FETs exhibit E-mode characteristics and are more compatible with traditional device processing, which facilitates reduction in size and integration of the large-scale circuits.

Moreover, the implementation of high-performance analog circuits requires device modeling with higher accuracy [72-74]. For our MoS_2 FETs, we established a device model with analytic equations and a level-62 SPICE model, whose results are consistent with the experimental results (also see SI). HSPICE software can be used to simulate and optimize the analog circuits in ANNs, which reduces the excessive burden of repeated tests and tape-outs, as illustrated in Fig. 3g. In the following section in this paper, we show how MoS_2 can be used to build complex digital and analog circuits for AI computation, which is a major step forward beyond laboratory-level 2DLM devices [1,72,73].

MoS_2 Dendrite and Synapse Structure for ANN Computing

Here, we examine how a MAC operation is implemented in MoS_2 circuits. As shown in Fig. 4a, a single-branch structure consisting of a dual-gate transistor acts as a multiplication module, which can be used for multiplication operations between the weight w and an input signal x on dendrite j. *The input data is connected to TG_1 and the synapse weight is transmitted to TG_2.* Such a dual-gate circuit is widely used for mixer or multiplier circuits [75]. In our circuit, the channel under G_1 works as a source-follower FET, and the G_2 drain voltage is $V_{d\text{-}G2} = V_{G1}\text{-}V_{th\text{-}G1}$, where $V_{th\text{-}G1}$ is the threshold voltage of gate G_1. G_2 works in the linear region because both V_{G1} and V_{G2} of T_2 can be independently adjusted to modify output current I_d. Multiplier between V_{G1} and V_{G2} can

be performed with only one MoS_2 FET, i.e., $I_d \propto w_{ij}^{(l)} x_j^{(l)}$ (see SI for more details). To better express the multiplication operation, Fig. 4b shows a 3-D plot of the measured current I_d as a function of V_{G1} and V_{G2}. The section cut above the gray plane corresponds to the more linear region (relatively large V_{G1} and V_{G2} values), which is apparently similar to the simulation results for an ideal multiplication operation between V_{G1} and $V_{G2,}$ as shown in the inset. Fig. 4c outlines a multi-branch parallel MoS_2 device and its corresponding circuit structure. In such a structure, multiple MoS_2 FET branches are connected in parallel by connecting all the drains together, which is similar to a synapse-neuron structure. Therefore, the result of an accumulation operation corresponds to the total output current, expressed $I_{out} = \Sigma I_{d.}$ Fig. 4d shows how the total current is a linear function of branch number at different values of V_{G1} and V_{G2}. The complete matrix computation process is shown in Fig. 4e. $G1$ and $G2$ are connected to an input sensor signal x_j and weight signal w_{ij}, where i is the neuron, and j corresponds to the synapse in neuron i. Thus, the total output current from the neuron i is proportional to $h_i = \Sigma w_{ij} x_j$.

After the MAC operation is implemented, as expressed by $h_i = \Sigma w_{ij} x_j$, the result is input to the activation function. It is noteworthy that our proposed circuit unit is now ready to perform convolution operations, which can be implemented by multiple MAC operations depending on the size of the input matrix. Out MoS_2 chip shown in Fig. 1cconnect 10 input signals for convolution operations. Furthermore, compared with a digital MAC operation, which requires many more transistors and multiple types of blocks, our analog convolution transistor network requires much lower computation time and chip area.

Peripheral circuits: a-RAM, activation function and digital-to-analog converter (DAC)

In addition to performing a MAC operation, the weight update function is also key for training an ANN, and dynamic weight writing and holding can be implemented by connecting the circuit in Fig. 4a with our a-RAM circuit, whose schematic diagram and electrical connections are shown in Fig. 5a. This consists of a MoS_2 FET T_1 with a high on/off ratio for switching and a capacitor C, which is able to store the weight value of dendrite j. During training, the control voltage V_C on the transistor is at a high level, and T_1 is turned on as a low-impedance path. We then test its memory functions. During a write operation, the drain voltage on T_1 acts as a bit line (BL) and is raised from 0 to

1 V. $V_{g\text{-}T1}$ acts as a word line (WL) and is raised from −3 to 3 V, as shown in Fig. 5b. Measured results are plotted as shown in the red curve, whose first positive current pulse (in red) indicates the capacitor has been recharged. During the hold state, T_1 is turned off to provide a high-impedance path with low current leakage for long-term retention. During a read operation, the BL increases from 0 to 0.5 V and the WL shifts from −3 to 3 V. The negative measured current indicates charge remains stored after the hold time.

Due to ultralow current leakage through monolayer MoS_2, the charge saved in the capacitor may persist in the neuronal layer over many processing cycles. Consequently, Fig. 5c shows that different output current levels can be maintained on the order of 10 s, which is sufficiently long for successive computation cycles in the convolutional layer. It also confirms that the input signal V_{in} is successfully stored in the capacitor, and thus it can provide a high-resolution analog voltage control signal for $V_{G1\text{-}T2}$, unlike an SRAM structure that can only maintain binary 0 and 1 levels.

In the von-Neumann architecture, the computing and memory units are separate. The computing unit needs to read and write memory at high frequency, and memory speed is a computing bottleneck. Based on the aforementioned electrical characteristics, it is noticed that in our MoS_2 circuit the a-RAM storage units and computing (MAC) modules are located close to each other (as shown in Fig. 5d); thus, a computation-in-memory structure can be implemented in our proposed ANN network.

Besides the a-RAM module, activation function circuits were also designed and fabricated on our MoS_2 chip. The activation function circuit is used to process the convolutional output, which simply consists of two MoS_2 FETs M_1 and M_2. The corresponding measurement results are shown in Fig. 5e, where the gate of M_2 is kept at a constant voltage and M_2 can be treated as a resistance R_{M2}. R_{M2} changes accordingly to different control voltages V_c. The output voltage V_{out} is:

$$V_{out}(V_{in}) = \frac{R_{M1}(V_{in})}{R_{M1}(V_{in}) + R_{M2}} V_{DD} \tag{4}$$

When the input voltage V_{in} is small, we have $V_{out} \sim V_{DD}$. By increasing V_{in}, R_{M1} approaches an appropriate range for Eq. (4) to act as an activation function whose gradient is $g_{m1}R_{M2}$, where g_{m1} is the transconductance of M_1. When V_{in} continues increasing, V_{out} approaches 0, which provides the required capability of an activation function.

During implementation of the ANN algorithm, it is necessary to efficiently update and store weight signals in a-RAM for recognizing samples. To achieve this, a weight update circuit is also designed and fabricated from MoS_2. The specific implementation is shown in Fig. 5e, in which M_{b0} is the bias current source, M_{cal1} and M_{cal2} are the calibration transistors. Eight different MoS_2 FETs M_1-M_8 act as current sources to form an 8-bit digital-to-analog converter (DAC) with W/L = 1, 2, 4... 128. V_{DD1} is a 4 V power source that outputs a fixed current I_0. When I_0 is constant, $V_{gs} = V_{ds}$ for M_{b0} also remains constant. Then, the V_{gs} values for M_{b0}, M_1, M_2 … M_8 are all equal. The according current values are then proportional to W/L for M_n, and each current source can be switched on and off individually based on the digital codes set for switches S_1-S_8 (i.e. 8-bit) [76,77]. Therefore, due to the current-mirror function, the total output current I_{OUT} can be expressed as follows:

$$I_{out} = \sum_{n=1}^{8} I_n = \sum_{n=1}^{8} I_{ns} \times S_n$$

where S_n is the state of the switch, which can be 0 or 1 based on whether the switch is off or on. There are 256 combinations that correspond to output states with 256 DAC steps, and the total output current I_{out} ranges from 0-255 I_0. I_{out} flows through resistor R_1 and the output voltage $V_{out} = V_{DD} - R_1 \times I_{out}$, thus the corresponding V_{out} value also has 256 levels. The experimental results show that M_{cal1} and M_{cal2} can be tuned for mismatch calibration of output levels (details see SI). Fig. 5f-g show the measured output current with 256 levels. Due to various mismatches in the circuit, the output current becomes less linear, which is primarily characterized by differential nonlinearity (DNL) and integral nonlinearity (INL). Upon turning on the calibration by tuning M_{cal1} and M_{cal2}, such DNL and INL can be significantly lowered (details see SI).

Braille character recognition with a MoS_2 ANN classifier

The required synapse weights in the device are not known, so they must be determined by training the network. Starting from small random weights, the network is fed with an input matrix $\boldsymbol{I}$ composed of sensed signals for ANN training, whose output matrix $\boldsymbol{T}$ is known. According to the current weights, the network computes an output $\boldsymbol{O}$ that we want to be as close as possible to $\boldsymbol{T}$. Our proposed multilayer ANN takes inputs with 10 features and classifies its output into 10 categories (Fig. 6a). Therefore, for any given training example S, $\boldsymbol{I}_S$, $\boldsymbol{T}_S$, and $\boldsymbol{O}_S$ are of size 10×1. During training, it is common to create batches of training examples, where each feature is a

column and each line is a training example. Taking m as the number of training examples, $\boldsymbol{I}$, $\boldsymbol{T}$, and $\boldsymbol{O}$ are matrices of size $10\times m$. We define the following loss function E to minimize the distance between $\boldsymbol{O}$ and $\boldsymbol{T}$ for all training examples:

$$E\left(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}\right) = \frac{1}{2m}\sum_{S=1}^{m}\sum_{i=1}^{10}(T_{iS} - O_{iS})^2\,, \tag{5}$$

where $\boldsymbol{w}^{(1)}$and $\boldsymbol{w}^{(2)}$ are the synapse weights between the input and hidden layer, and between the hidden and output layers, respectively. The synapse weights are then optimized using gradient descent starting from small random values. Similar to most machine learning applications, the weight optimization in our chip is realized off-chip (on a separate computer) using BP (details see SI).

To conclude, the above illustrated proof-of-concept ANN system integrates the essential functions required for practical ANN-preprocessing applications, and definitely has room for further optimization since this is only an early stage exploration. In addition, due to the atomically thin and air stable nature of 2DLMs, our MoS_2-based ANNs can be fabricated on flexible substrates for biocompatible and implantable application, such as perception and processing functions of biological nerves combined with energy-efficient sensing.

Finally, we demonstrate a tactile Braille classifier algorithm that combines all the ANN-MoS_2 circuits described above. The tactile sensation of skin is an important part of human perception, including perception of temperature, pain, vibration, and shapes by receptor neurons. The skin receptors can also be divided into slow and fast adaptive types, mainly depending on the location and distribution of the receptor in the skin (Fig. 5b, also see SI more details). The skin perceives the shape of an object because the generated pressure is similar to the shape of the object, which is how blind individuals can read Braille.

The chip shown in Fig 1a takes a 3 × 3 array input representing tactile sensation of touching a Braille character. The value of each tactile pixel is discretized into 256 pressure levels. Although the pressure sensing function is not included in this work, it can be easily realized by taking advantage of the piezoelectric properties of MoS_2, which converts pressure into an electric signal and mimics skin receptors. The inputs (I_1, I_2…I_9) into the ANN correspond to 9 voltage values in the flattened input array, and I_{10} is the bias current (see Fig. 6a). Therefore, the signals can be displayed as a voltage intensity mapping graph. Typical current distributions for letters ‘N’, ‘V’, and ‘Z’ with

different pressure levels are shown in Fig. 6c. In this work, we generate pressure distribution data by adding noise to ideal samples, and three different distribution sets are designed for each letter, and each set has 50 samples for further ANN testing (Fig. 6c). In our system, the input voltage range is 0 to 3 V, while the bias input V_{10} is fixed based on the algorithm. After optimizing the weights for 23 training epochs, the final weights of the input and output layers are shown in Fig. 6d. These weights are subsequently updated by the DAC in the ANN chip, ultimately yielding a classification accuracy of 97%, as shown in Fig. 6e-f.

In this paper, we used high-quality uniform MoS_2 thin films to produce the first analog-ANN circuit based on MoS_2. These analog circuits were designed and optimized using high-fidelity Level-62 SPICE models for MoS_2 FET transistors. The chip includes 16 artificial neurons organized in two layers. Each artificial neuron combines MAC units acting as soma, an A-RAM memory to store the synapse weights, a DAC to update weights, and a nonlinear activation circuit. After optimizing the weights off-chip, the device provides a Braille letter recognition rate above 97%, which provides a way for smarter and flexible artificial sensing tissues.

Methods

A continuous MoS_2 monolayer film is grown on a 2-inch sapphire substrate. Detailed synthesis of wafer-scale MoS_2 and the corresponding device fabrication processes are described in the supplementary information. For the convenience of a circuit-level electrical test, most of the circuits in the work were fabricated on a 1×1 cm^2 wafer cut from the 2-inch MoS_2 wafer. In Fig. 3d, 380 MoS_2 TG-FETs are fabricated and measured on such 1×1 cm^2 wafer to test the device performance uniformity, and all TFTs had the same channel size. Agilent b1500a with 6 SMU channels is used to measure simple circuits such as logic invertor. For a more complex circuitry, the wafer is mounted and wire-bonded to a custom designed PCB board for electrical measurement.

Acknowledgment:

This work was supported by National Key Research and Development Program (2016YFA0203900, 2018YFB2202500), Shanghai Municipal Science and Technology Commission (18JC1410300), Research Grant Council of Hong Kong (15205619) and National Natural Science Foundation of China (61925402, 61934008). S. Ma

acknowledges financial support from the Fudan Project for Scientific Research of Introduced Talents in 2020 (JIH1233034).

References：

1 Verreck, D. *et al.* The Role of Nonidealities in the Scaling of MoS_2 FETs. *IEEE Trans. Electron Dev.* **65**, 4635-4640 (2018).

2 Desai, S. B. *et al.* MoS_2 transistors with 1-nanometer gate lengths. *Science* **354**, 99-102 (2016).

3 Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nat. Nanotechnol.* **6**, 147-150 (2011).

4 Wachter, S., Polyushkin, D. K., Bethge, O. & Mueller, T. A microprocessor based on a two-dimensional semiconductor. *Nat. Commun.* **8**, 1-6 (2017).

5 Lin, Z. *et al.* Solution-processable 2D semiconductors for high-performance large-area electronics. *Nature* **562**, 254-258 (2018).

6 Zhang, Q. *et al.* Simultaneous synthesis and integration of two-dimensional electronic components. *Nat. Electron.* **2**, 164-170 (2019).

7 Fiori, G. *et al.* Electronics based on two-dimensional materials. *Nat. Nanotechnol.* **9**, 768 (2014).

8 Gao, Q. *et al.* Scalable high performance radio frequency electronics based on large domain bilayer MoS_2. *Nat. Commun.* **9**, 1-8 (2018).

9 Polyushkin, D. K. *et al.* Analogue two-dimensional semiconductor electronics. *Nat. Electron.* **3**, 486-491 (2020).

10 Hou, X., Chen, H., Zhang, Z., Wang, S. & Zhou, P. 2D Atomic Crystals: A Promising Solution for Next-Generation Data Storage. *Adv. Electron. Mater.* **5**, 1800944 (2019).

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

12 Chen, W.-H. *et al.* CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. *Nat. Electron.* **2**, 420-428 (2019).

13 Yao, J. *et al.* Highly transparent nonvolatile resistive memory devices from silicon oxide and graphene. *Nat. Commun.* **3**, 1-8 (2012).

14 Graves, A. *et al.* Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471-476 (2016).

15 Ielmini, D. & Wong, H.-S. P. J. N. E. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333-343 (2018).

16 Yang, J. J. *et al.* Memristive switching mechanism for metal/oxide/metal nanodevices. *Nat. Nanotechnol.* **3**, 429 (2008).

17 Wang, Z. *et al.* Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101-108 (2017).

18 Le Gallo, M. *et al.* Mixed-precision in-memory computing. *Nat. Electron.* **1**, 246-253 (2018).

19 Waser, R. & Aono, M. in *Nanoscience And Technology: A Collection of Reviews from Nature Journals* 158-165 (World Scientific, 2010).

20 Jang, J. E. *et al.* Nanoscale memory cell based on a nanoelectromechanical switched capacitor. *Nat. Nanotechnol.* **3**, 26 (2008).

21 Rozenberg, M., Inoue, I. & Sanchez, M. Nonvolatile memory with multilevel switching: a basic model. *Phys. Rev. Lett.* **92**, 178302 (2004).

22 Wang, Z. *et al.* Reinforcement learning with analogue memristor arrays. *Nat. Electron.*

2, 115-124 (2019).

23 Lastras-Montaño, M. A. & Cheng, K.-T. Resistive random-access memory based on ratioed memristors. *Nat. Electron.* **1**, 466-472 (2018).

24 Wang, M. *et al.* Robust memristors based on layered two-dimensional materials. *Nat. Electron.* **1**, 130-136 (2018).

25 Yang, R. *et al.* Ternary content-addressable memory with MoS_2 transistors for massively parallel data search. *Nat. Electron.* **2**, 108-114 (2019).

26 Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22-29 (2018).

27 Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. J. n. The missing memristor found. *Nature* **453**, 80-83 (2008).

28 Shulaker, M. M. *et al.* Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **547**, 74-78 (2017).

29 Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64 (2015).

30 Xiang, D. *et al.* Two-dimensional multibit optoelectronic memory with broadband spectrum distinction. *Nat. Commun.* **9**, 1-8 (2018).

31 Roy, K. *et al.* Graphene–MoS_2 hybrid structures for multifunctional photoresponsive memory devices. *Nat. Nanotechnol.* **8**, 826-830 (2013).

32 Liu, C. *et al.* A semi-floating gate memory based on van der Waals heterostructures for quasi-non-volatile applications. *Nat. Nanotechnol.* **13**, 404-410 (2018).

33 Liu, C. *et al.* Small footprint transistor architecture for photoswitching logic and in situ memory. *Nat. Nanotechnol.* **14**, 662-667 (2019).

34 Kim, K.-H. *et al.* A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. *Nano Lett.* **12**, 389-395 (2012).

35 Bertolazzi, S., Krasnozhon, D. & Kis, A. Nonvolatile memory cells based on MoS_2/graphene heterostructures. *ACS nano* **7**, 3246-3252 (2013).

36 Huo, N. & Konstantatos, G. Ultrasensitive all-2D MoS_2 phototransistors enabled by an out-of-plane MoS_2 PN homojunction. *Nat. Commun.* **8**, 1-6 (2017).

37 Lopez-Sanchez, O., Lembke, D., Kayci, M., Radenovic, A. & Kis, A. Ultrasensitive photodetectors based on monolayer MoS_2. *Nat. Nanotechnol.* **8**, 497-501 (2013).

38 Kim, K. S. *et al.* Ultrasensitive MoS_2 photodetector by serial nano-bridge multi-heterojunction. *Nat. Commun.* **10**, 1-10 (2019).

39 Goossens, S. *et al.* Broadband image sensor array based on graphene–CMOS integration. *Nat. Photonics* **11**, 366-371 (2017).

40 Lee, G.-H. *et al.* Flexible and Transparent MoS_2 Field-Effect Transistors on Hexagonal Boron Nitride-Graphene Heterostructures. *ACS Nano* **7**, 7931-7936 (2013).

41 Cheng, R. *et al.* Few-layer molybdenum disulfide transistors and circuits for high-speed flexible electronics. *Nat. Commun.* **5**, 5143 (2014).

42 Hua, Q. *et al.* Skin-inspired highly stretchable and conformable matrix networks for multifunctional sensing. *Nat. Commun.* **9**, 1-11 (2018).

43 Schedin, F. *et al.* Detection of individual gas molecules adsorbed on graphene. *Nat. Mater.* **6**, 652-655 (2007).

44 Perkins, F. K. *et al.* Chemical vapor sensing with monolayer MoS_2. *Nano Lett.* **13**, 668-673 (2013).

45 Cui, S., Wen, Z., Huang, X., Chang, J. & Chen, J. Stabilizing MoS_2 nanosheets through SnO_2 nanocrystal decoration for high-performance gas sensing in air. *Small* **11**, 2305-2313 (2015).

46 Late, D. J. *et al.* Sensing behavior of atomically thin-layered MoS_2 transistors. **7**,

4879-4891 (2013).

47 Schwartz, G. *et al.* Flexible polymer transistors with high pressure sensitivity for application in electronic skin and health monitoring. *Nat. Commun.* **4**, 1-8 (2013).

48 Yu, F. *et al.* Ultrasensitive Pressure Detection of Few-Layer MoS_2. *Adv. Mater.* **29**, 1603266 (2017).

49 Gong, S. *et al.* A wearable and highly sensitive pressure sensor with ultrathin gold nanowires. *Nat. Commun.* **5**, 1-8 (2014).

50 Sugiyama, M. *et al.* An ultraflexible organic differential amplifier for recording electrocardiograms. *Nat. Electron.* **2**, 351-360 (2019).

51 Kim, J. *et al.* Stretchable silicon nanoribbon electronics for skin prosthesis. *Nat. Commun.* **5**, 1-11 (2014).

52 Liu, Y. *et al.* Van der Waals heterostructures and devices. *Nature Reviews Materials* **1**, 16042 (2016).

53 Jiang, J., Parto, K., Cao, W. & Banerjee, K. Monolithic-3D Integration with 2D Materials: Toward Ultimate Vertically-Scaled 3D-ICs. *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)* **1-3** (2018).

54 Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52 (2018).

55 Zhou, F., Zhou, Z., Chen, J., Choy, T. H. & Chai, Y. Optoelectronic resistive random access memory for neuromorphic vision sensors. *Nat. Nanotechnol.* **14**, 776-782 (2019).

56 Ambrogio, S. *et al.* Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60-67 (2018).

57 P. Srinivasan, O. O., Y. S. Choi, A. Marshall. Context Dependent Effects on LF (1/f) Noise in Advanced CMOS Devices. *IEDM* **10-612** (2010).

58 Xuejun Xie, D. S., Wei Liu, Jiahao Kang, Ognian Marinov, M. Jamal Deen, and Kaustav Banerjee. Low-Frequency Noise in Bilayer MoS_2 Transistor. *ACS Nano* **8, 6, 5633–5640**, (2014).

59 Vinod K. Sangwan, H. N. A., Deep Jariwala, Tobin J. Marks, Lincoln J. Lauhon, and Mark C. Hersam. Low-Frequency Electronic Noise in Single-Layer MoS_2 Transistors. *Nano Lett* **13, 4351−4355** (2013).

60 Mennel, L. *et al.* Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62-66 (2020).

61 Kang, K. *et al.* High-mobility three-atom-thick semiconducting films with wafer-scale homogeneity. *Nature* **520**, 656-660 (2015).

62 Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

63 Fitzsimonds, R. M., Song, H. J. & Poo, M. M. Propagation of activity-dependent synaptic depression in simple neural networks. *Nature* **388**, 439-448 (1997).

64 Yang, P. *et al.* Batch production of 6-inch uniform monolayer molybdenum disulfide catalyzed by sodium in glass. *Nat. Commun.* **9**, 1-10 (2018).

65 Zhang, S. *et al.* Wafer-scale transferred multilayer MoS_2 for high performance field effect transistors. *Nanotechnology* **30**, 174002, (2019).

66 Xu, H. *et al.* High-Performance Wafer-Scale MoS_2 Transistors toward Practical Application. *Small* **14**, e1803465, (2018).

67 Tang, H. *et al.* Realizing Wafer-Scale and Low-Voltage Operation MoS_2 Transistors via Electrolyte Gating. *Advanced Electronic Materials* **6**, (2019).

68 Yu, L. *et al.* Design, Modeling, and Fabrication of Chemical Vapor Deposition Grown MoS_2 Circuits with E-Mode FETs for Large-Area Electronics. *Nano Lett* **16**, 6349-6356, (2016).

69 Wachter, S., Polyushkin, D. K., Bethge, O. & Mueller, T. A microprocessor based on a two-dimensional semiconductor. *Nat Commun* **8**, 14948, (2017).

70 L. Yu , D. E.-D., S. Ha, X. Ling, Y. Lin, A. Zubair, Y. Zhang, Y.-H. Lee, J. Kong, A. Chandrakasan and T. Palacios. Enhancement-Mode Single-layer CVD MoS_2 FET Technology for Digital Electronics. *IEDM* **15-835** (2015).

71 Dmitry K. Polyushkin, S. W., Lukas Mennel, Maksym Paliy, Giuseppe Iannaccone, Gianluca Fiori, Daniel Neumaier, Barbara Canto and Thomas Mueller. Analogue two-dimensional semiconductor electronics. *Nature Electronics* **3, 486–491** (2020).

72 Kim, D. H. *et al.* Physical parameter-based SPICE models for InGaZnO thin-film transistors applicable to process optimization and robust circuit design. *IEEE Electron Device Lett.* **33**, 59-61 (2011).

73 You, W.-X. & Su, P. A compact subthreshold model for short-channel monolayer transition metal dichalcogenide field-effect transistors. *IEEE Trans. Electron Dev.* **63**, 2971-2974 (2016).

74 Perumal, C. *et al.* A Compact a-IGZO TFT Model Based on MOSFET SPICE Level=3 Template for Analog/RF Circuit Designs. *IEEE Electron Device Lett.* **34**, 1391-1393 (2013).

75 Razavi, B. & Behzad, R. *RF microelectronics*. Vol. 2 (Prentice Hall New Jersey, 1998).

76 Bastos, J., Marques, A. M., Steyaert, M. S. & Sansen, W. A 12-bit intrinsic accuracy high-speed CMOS DAC. *IEEE J. Solid-State Circuits* **33**, 1959-1969 (1998).

77 Marques, A. *et al.* A 12 b accuracy 300 Msample/s update rate CMOS DAC. *1998 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, ISSCC. First Edition*, 216-217 (1998).

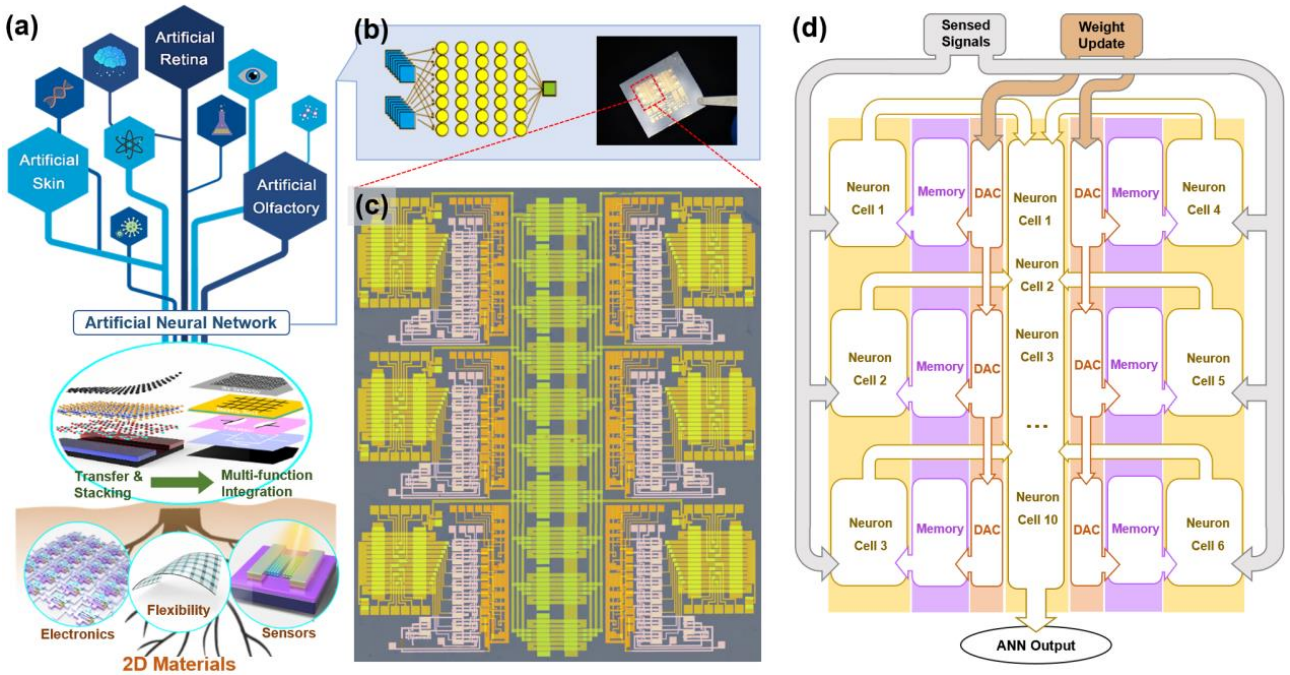

Figure 1. (a) Schematic diagram showing applications based on sensors, computing, and memory modules fabricated from flexible 2DLMs. Different modules can be stacked vertically. (b) Schematic algorithm of the ANN module and a photograph of the integrated circuit. (c) A false-color optical microscopic image of an as-fabricated ANN integrated circuit. (d) Schematic circuit architecture of our 2DLM based ANN.

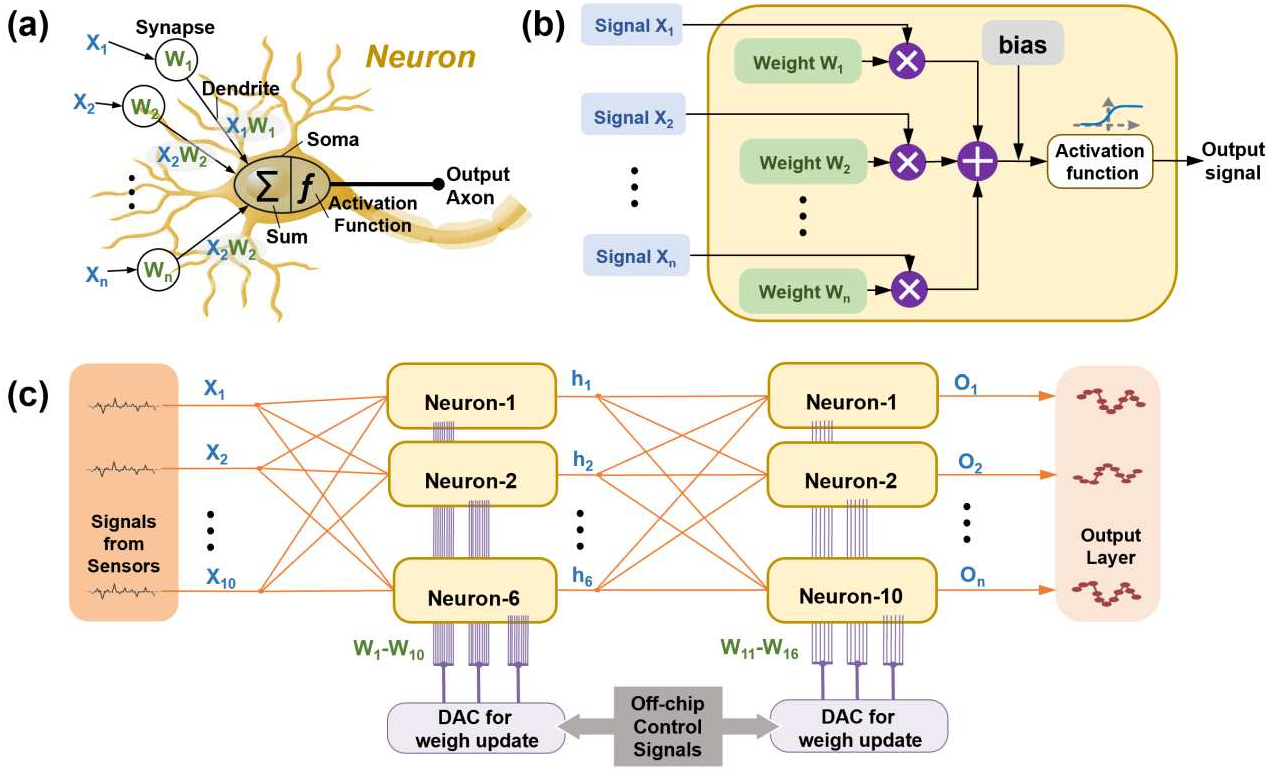

Fig. 2 (a) A computational model of a CNN neuron. (b) An MoS_2-based artificial neuron cell including soma (MAC operation), dendrite (data input from sensor), and synapse (the a-RAM is to store the weight value). (c) Schematic diagram of the proposed ANN, which consists of an activation function, MAC, a-RAM, and timing control circuits.

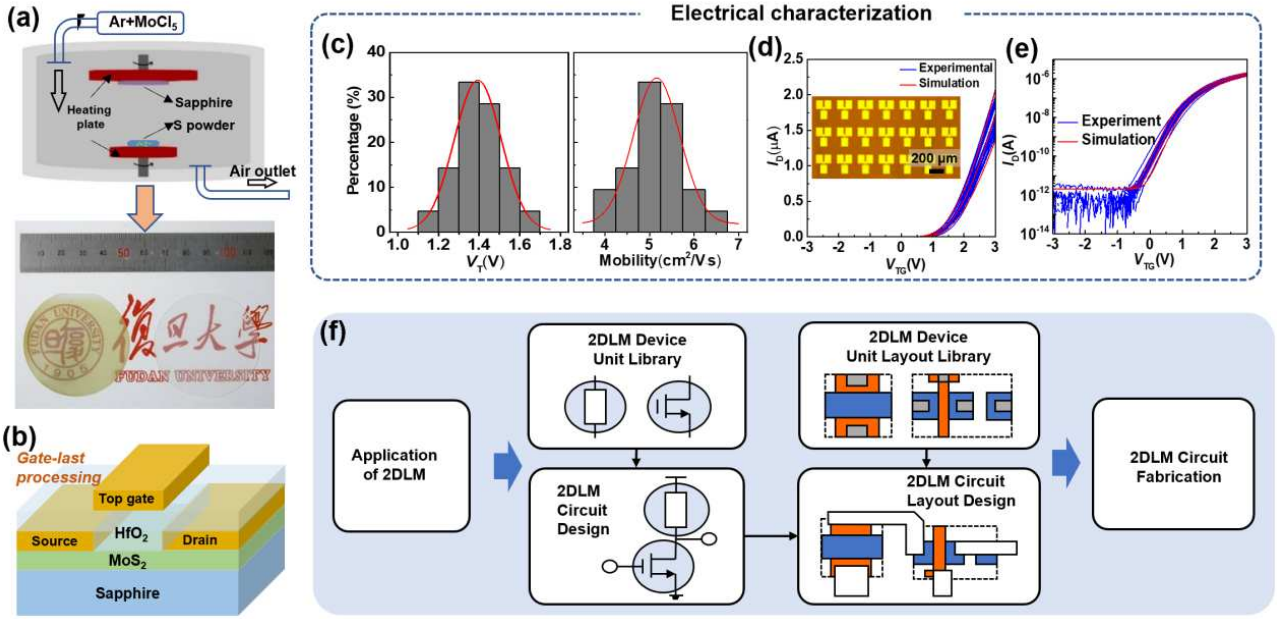

Fig 3. (a) Schematic diagram of the material synthesis setup. The right photograph shows two 2 in. diameter sapphire wafers. The dark yellow wafer is uniformly covered with MoS_2 grown via CVD, and the other wafer is a bare sapphire wafer. (b) Schematic of a MoS_2 TG-FET structure. (c) Histogram of the threshold voltage and mobility analyzed after measuring the transfer curves of more than 20 MoS_2 TG-FETs on a 1 × 1 cm^2 wafer. (d-e) Transfer characteristics of 20 FETs located at the center region of a wafer in linear and logarithmic scale, respectively. The red curves show simulation results. The insert shows a photograph of the MoS_2 TG-FET arrays, in which the scale bar is 150 μm. (f) A flow diagram of the integrated circuit design process and customized design process for fabricating 2DLMs.

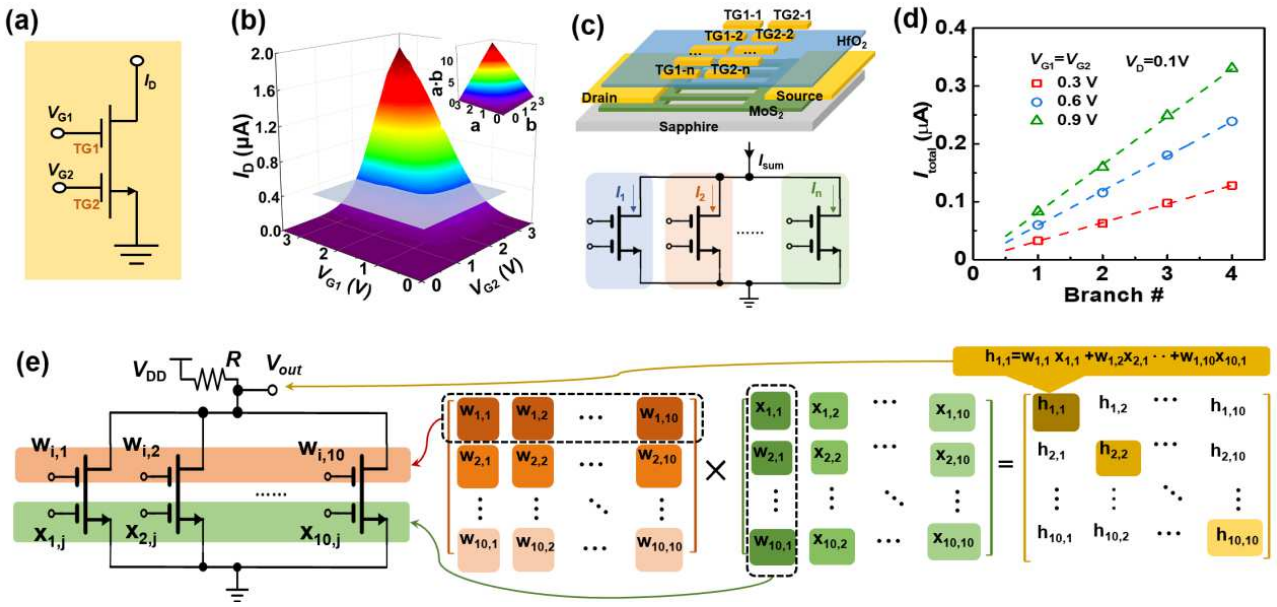

Fig 4. (a) Schematic diagram of the multiplier unit (one branch) consisting of a MoS_2 FET with dual top gate TG_1 and TG_2. (b) 3-D plot showing measured I_D data as a function of V_{G1} and V_{G2} from the multiplier unit. At a relatively large voltage (the region cut above the gray plane), the result is close to the ideal multiplication correlation between V_{G1} and V_{G2}, and the inset shows simulation results for $a \cdot b$ as a function of a and b. (c) The upper graph shows results for a parallel MoS_2 device with multi-branch dual-gate FETs, and the lower graph shows the corresponding schematic. (d) Total current as a function of the number of multiplier branches, where the result shows a summation operation.

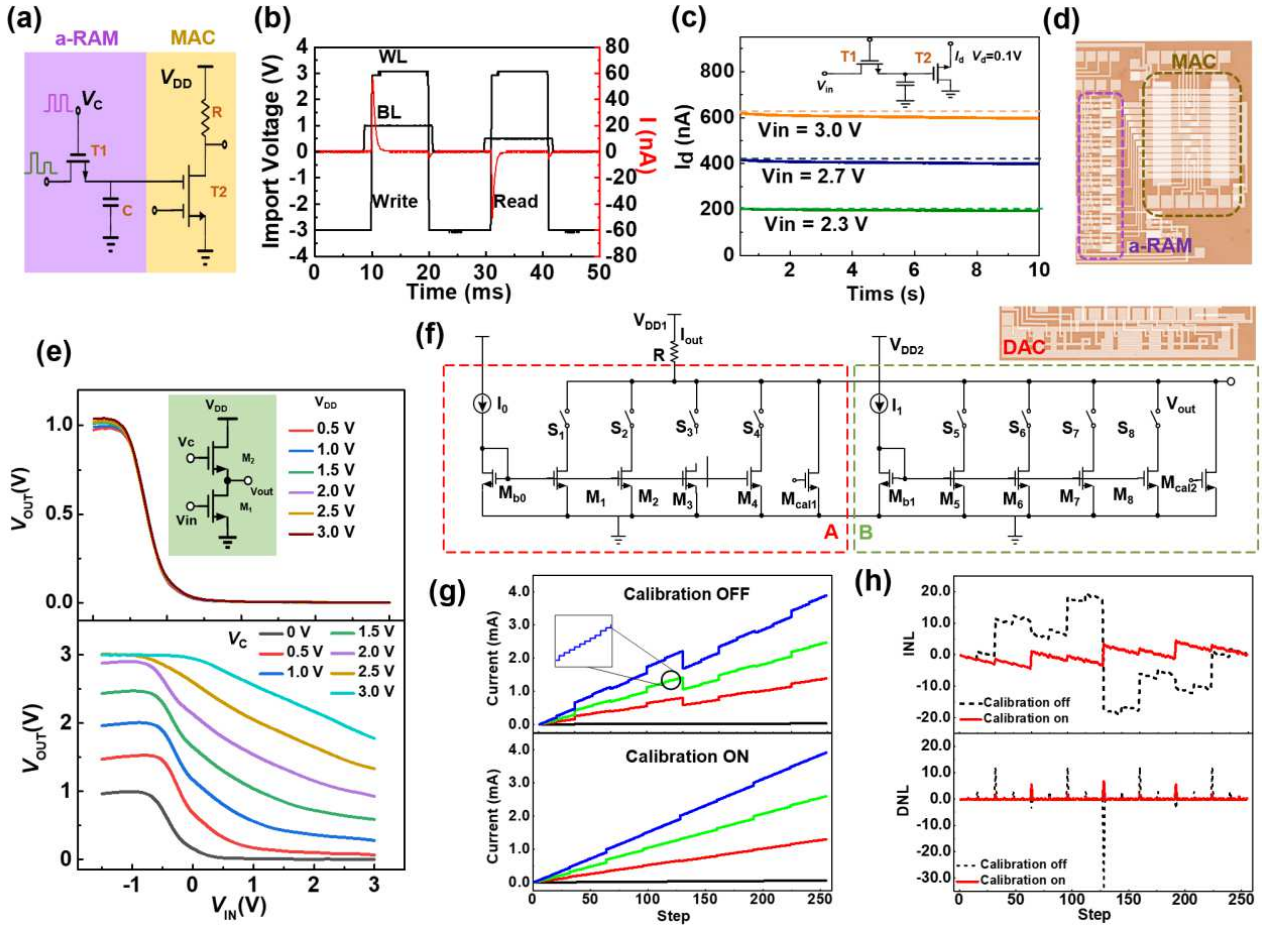

Fig 5. Peripheral circuits used to implement ANN calculations. (a) Schematic MoS_2 circuit diagram resembling a neuron with one synapse, where the left a-RAM circuit is used to store and update the weight. (b) V_{out} vs. V_{in} in different control voltages V_c. (c) Measured I_D value in T_2 controlled by the voltage stored in the capacitor during one test cycle (10 s). Different values of V_{in} is used to charge the capacitance before turning off T_1. (d) An optical microscopic image of a multi-branch MoS_2 MAC circuit with each branch connected an a-RAM. (e) Measured V_{out} vs. V_{in} for the activation function circuit with different V_{DD} values (upper graph) and control voltages V_c (lower graph). The inset shows a schematic circuit model. (f) Schematic diagram of the DAC converter used for weight update. The upper right optical image shows the layout of the proposed DAC. (g) Measured current output from the DAC with 256 levels. Different colors correspond to different bias currents. The top and bottom graphs correspond to calibration off and on, respectively. (h) DNL and INL as a function of 256 levels. The step error can be minimized when the calibration circuits are turned on.

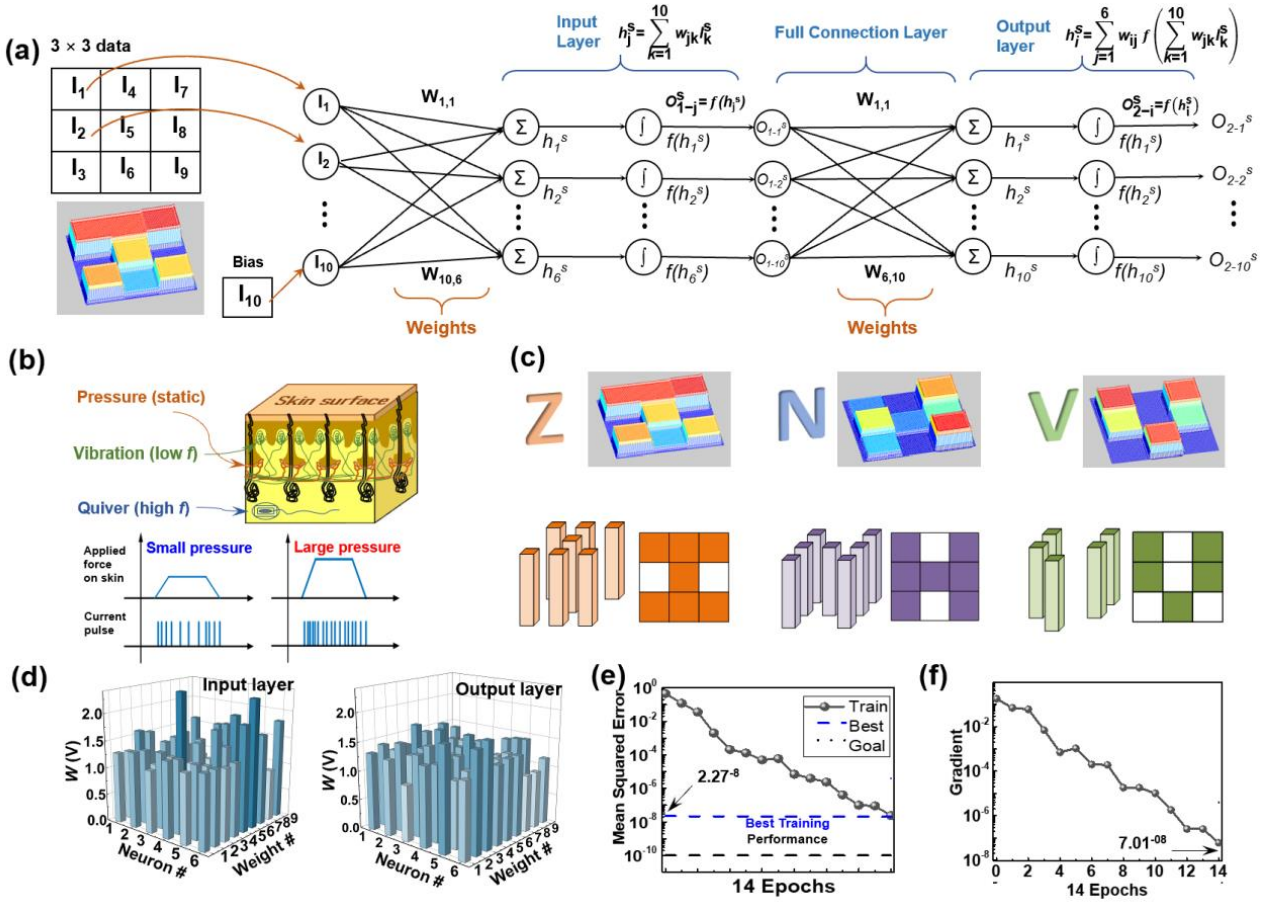

Fig 6. Summary of chip evaluation and modeling. (a) A schematic structure of the proposed CNN, which includes the input layer, fully-connected layer, and output neurons. I_1-I_9 correspond to the input signals from the sensor. (b) Schematic showing three types of sensors in different locations on human skin corresponding to three different vibration modes. The bottom graph shows generated current pulses with small and large pressures. (c) Current distribution for letters 'N', 'V', and 'Z' with different pressure levels and classification results after successive training epochs. (d) The upper graph shows the mean squared error of the ANN network as a function of training epochs. The lower graph shows the gradient of weight convergence as a function of the number of training epochs.

Figures

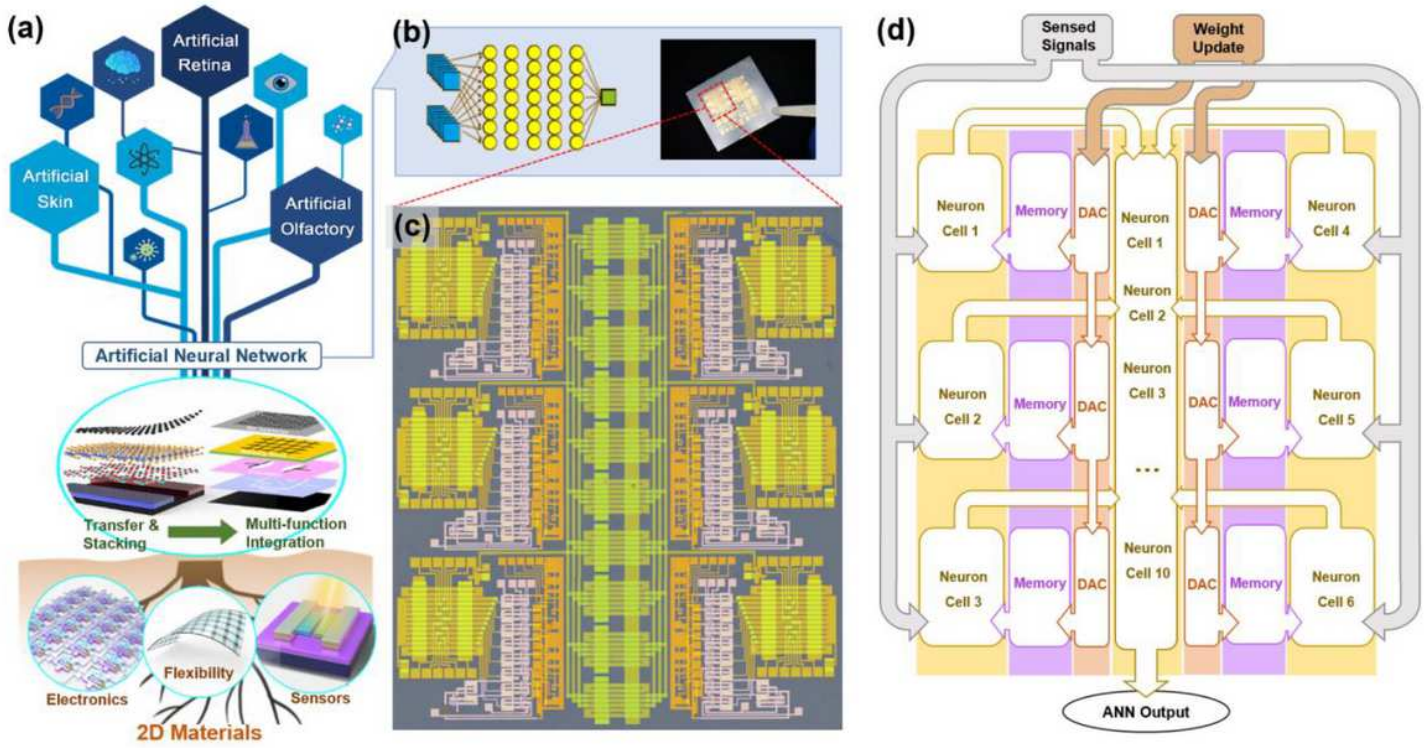

Figure 1

(a) Schematic diagram showing applications based on sensors, computing, and memory modules fabricated from flexible 2DLMs. Different modules can be stacked vertically. (b) Schematic algorithm of the ANN module and a photograph of the integrated circuit. (c) A false-color optical microscopic image of an as-fabricated ANN integrated circuit. (d) Schematic circuit architecture of our 2DLM based ANN.

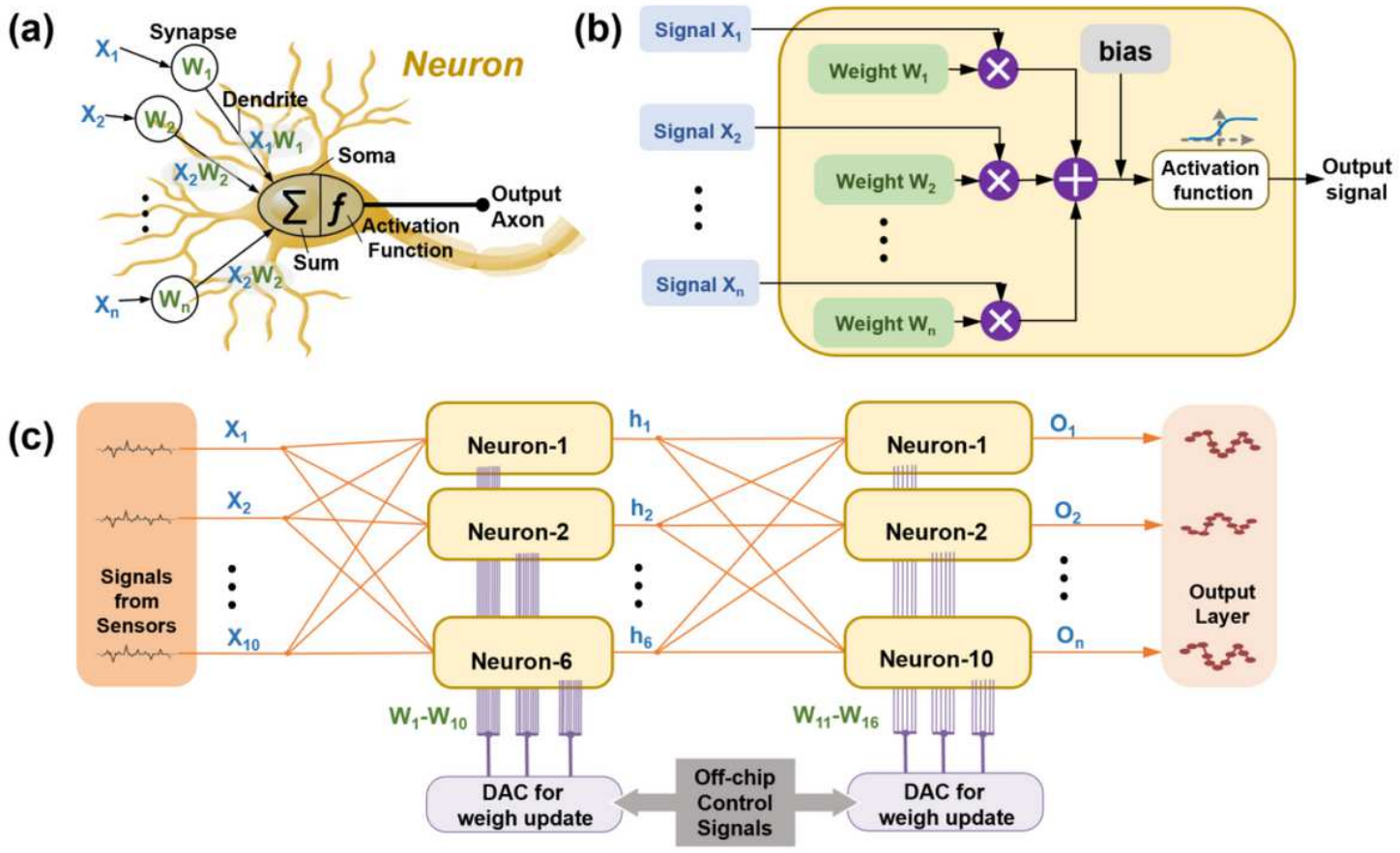

Figure 2

(a) A computational model of a CNN neuron. (b) An MoS2-based artificial neuron cell including soma (MAC operation), dendrite (data input from sensor), and synapse (the a-RAM is to store the weight value). (c) Schematic diagram of the proposed ANN, which consists of an activation function, MAC, a-RAM, and timing control circuits.

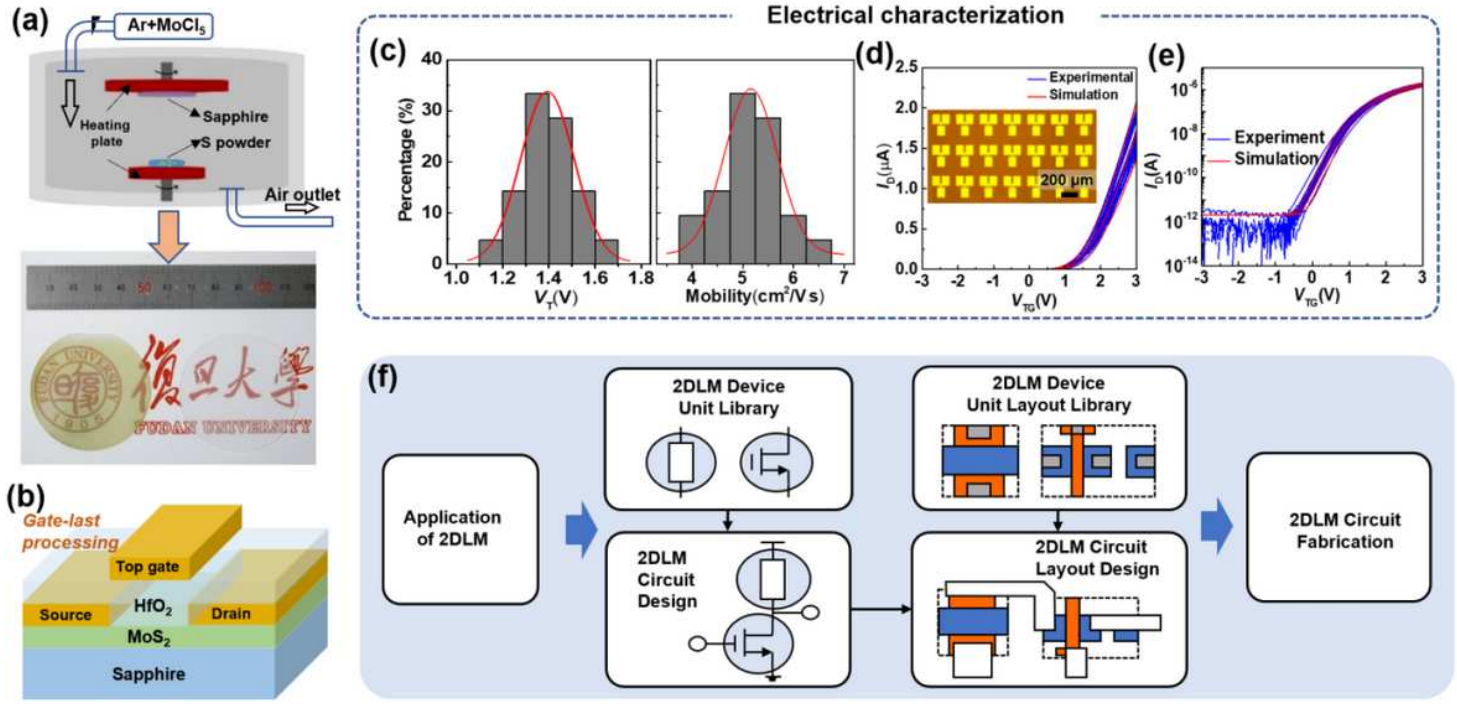

Figure 3

(a) Schematic diagram of the material synthesis setup. The right photograph shows two 2 in. diameter sapphire wafers. The dark yellow wafer is uniformly covered with MoS2 grown via CVD, and the other wafer is a bare sapphire wafer. (b) Schematic of a MoS2 TG-FET structure. (c) Histogram of the threshold voltage and mobility analyzed after measuring the transfer curves of more than 20 MoS2 TG-FETs on a 1 × 1 cm2 wafer. (d-e) Transfer characteristics of 20 FETs located at the center region of a wafer in linear and logarithmic scale, respectively. The red curves show simulation results. The insert shows a photograph of the MoS2 TG-FET arrays, in which the scale bar is 150 μm. (f) A flow diagram of the integrated circuit design process and customized design process for fabricating 2DLMs.

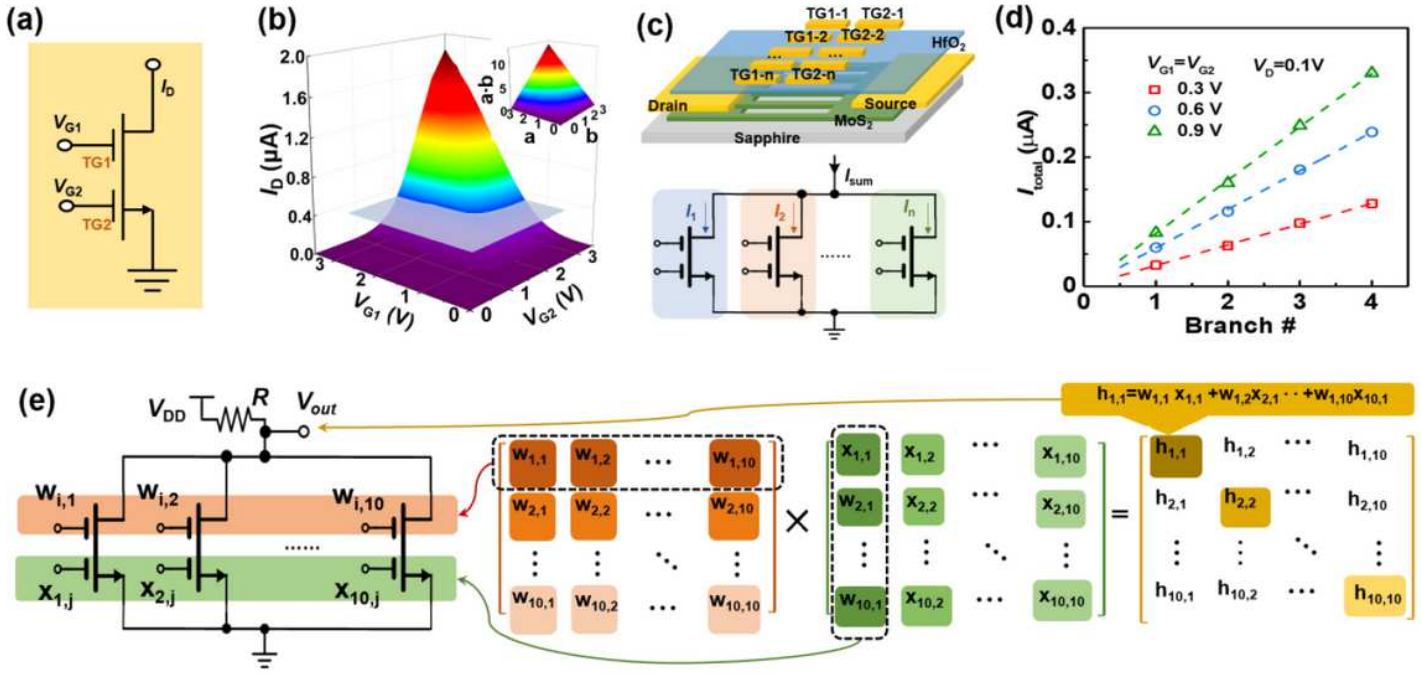

Figure 4

(a) Schematic diagram of the multiplier unit (one branch) consisting of a MoS2 FET with dual top gate TG1 and TG2. (b) 3-D plot showing measured ID data as a function of VG1 and VG2 from the multiplier unit. At a relatively large voltage (the region cut above the gray plane), the result is close to the ideal multiplication correlation between VG1 and VG2, and the inset shows simulation results for a⊠b as a function of a and b. (c) The upper graph shows results for a parallel MoS2 device with multi-branch dual-gate FETs, and the lower graph shows the corresponding schematic. (d) Total current as a function of the number of multiplier branches, where the result shows a summation operation.

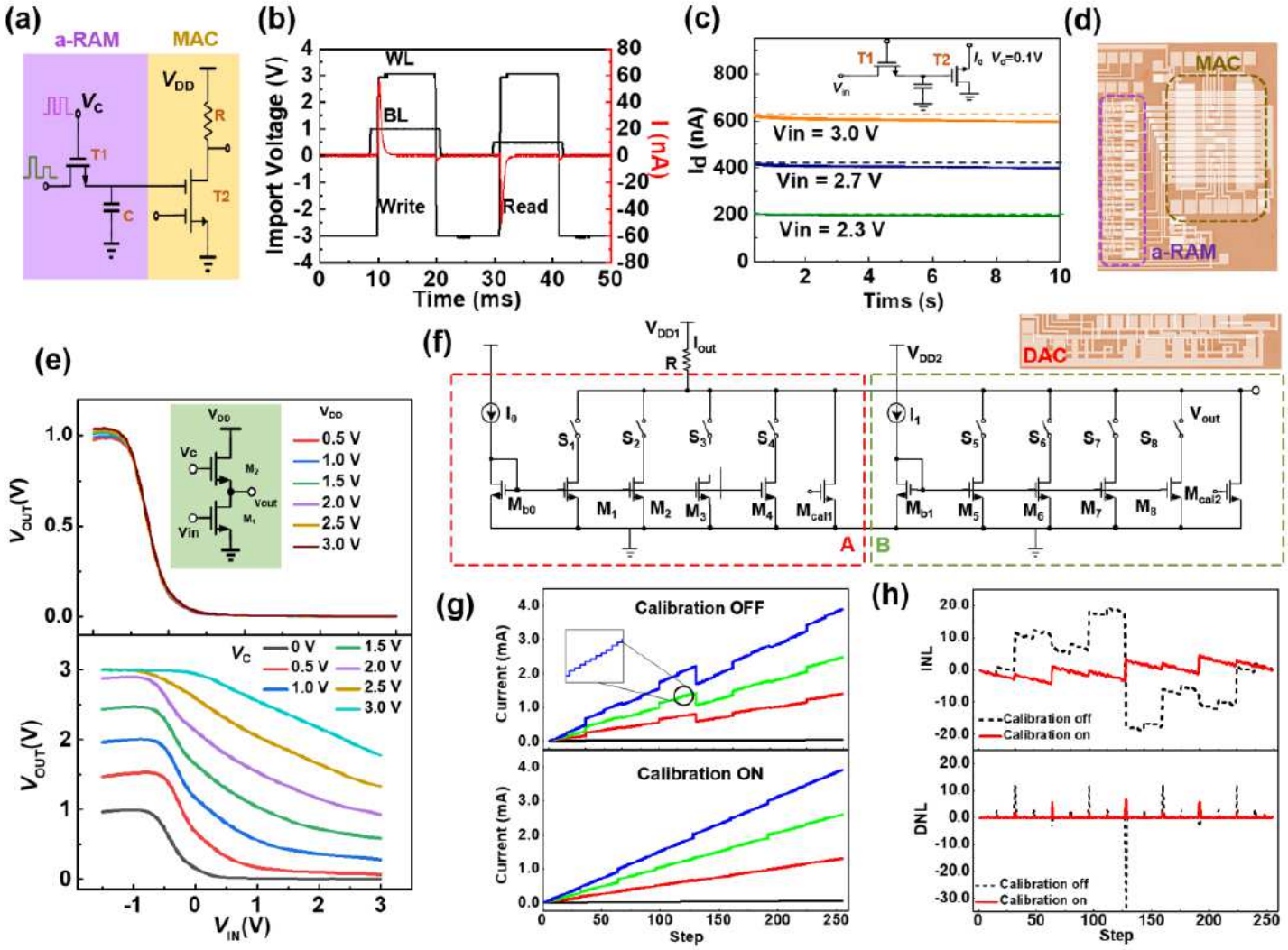

Figure 5

Peripheral circuits used to implement ANN calculations. (a) Schematic MoS2 circuit diagram resembling a neuron with one synapse, where the left a-RAM circuit is used to store and update the weight. (b) Vout vs. Vin in different control voltages Vc. (c) Measured ID value in T2 controlled by the voltage stored in the capacitor during one test cycle (10 s). Different values of Vin is used to charge the capacitance before turning off T1. (d) An optical microscopic image of a multi-branch MoS2 MAC circuit with each branch connected an a-RAM. (e) Measured Vout vs. Vin for the activation function circuit with different VDD values (upper graph) and control voltages Vc (lower graph). The inset shows a schematic circuit model. (f) Schematic diagram of the DAC converter used for weight update. The upper right optical image shows the layout of the proposed DAC. (g) Measured current output from the DAC with 256 levels. Different colors correspond to different bias currents. The top and bottom graphs correspond to calibration off and on, respectively. (h) DNL and INL as a function of 256 levels. The step error can be minimized when the calibration circuits are turned on.

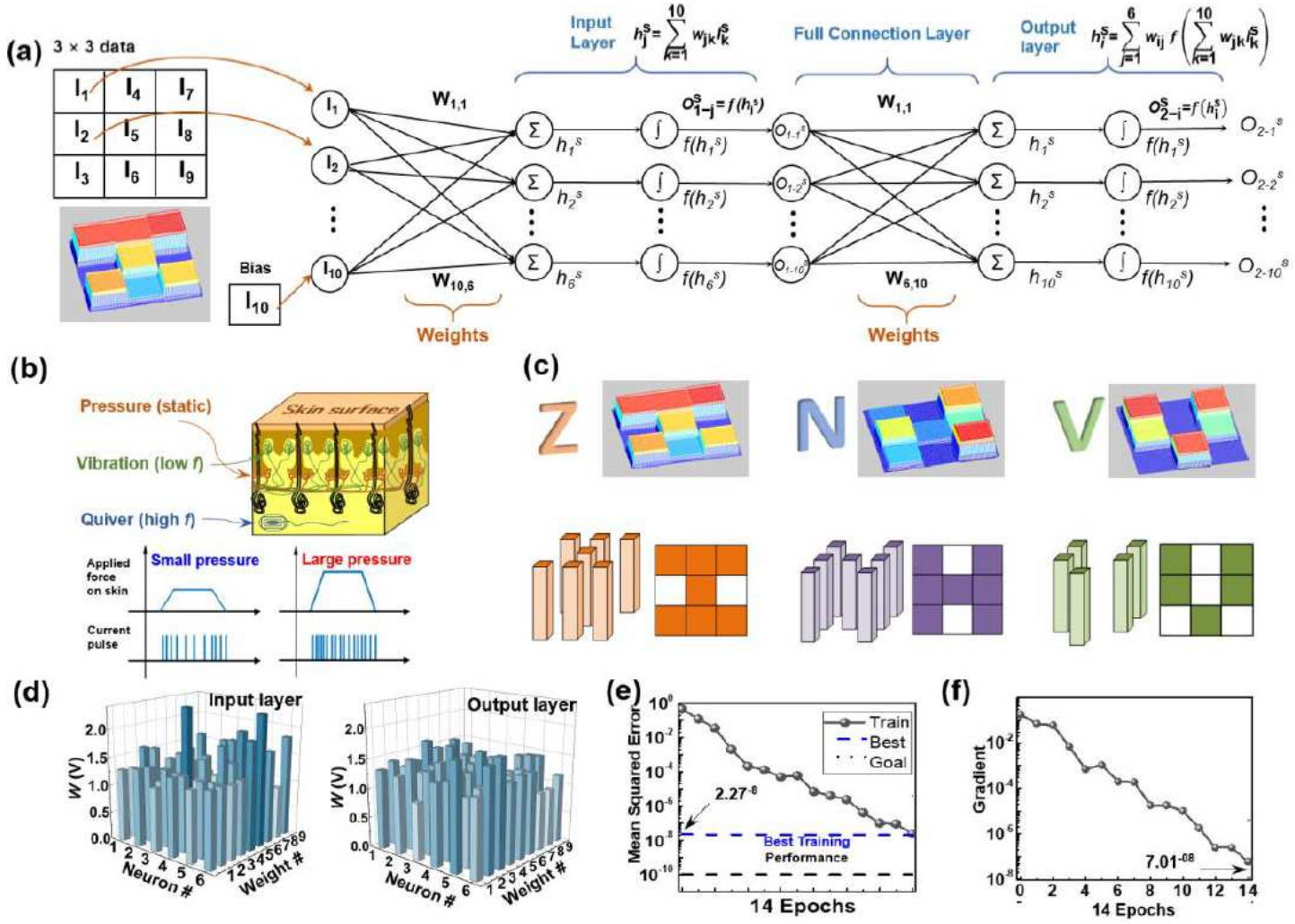

Figure 6

Summary of chip evaluation and modeling. (a) A schematic structure of the proposed CNN, which includes the input layer, fully-connected layer, and output neurons. I1-I9 correspond to the input signals from the sensor. (b) Schematic showing three types of sensors in different locations on human skin corresponding to three different vibration modes. The bottom graph shows generated current pulses with small and large pressures. (c) Current distribution for letters ‘N’, ‘V’, and ‘Z’ with different pressure levels and classification results after successive training epochs. (d) The upper graph shows the mean squared error of the ANN network as a function of training epochs. The lower graph shows the gradient of weight convergence as a function of the number of training epochs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf