



OPEN

In-sensor neural network for high energy efficiency analog-to-information conversion

Sudarsan Sadasivuni¹, Sumukh Prashant Bhanushali², Imon Banerjee³ & Arindam Sanyal²✉

This work presents an on-chip analog-to-information conversion technique that utilizes analog hyper-dimensional computing based on reservoir-computing paradigm to process electrocardiograph (ECG) signals locally in-sensor and reduce radio frequency transmission by more than three orders-of-magnitude. Instead of transmitting the naturally sparse ECG signal or extracted features, the on-chip analog-to-information converter analyzes the ECG signal through a nonlinear reservoir kernel followed by an artificial neural network, and transmits the prediction results. The proposed technique is demonstrated for detection of sepsis onset and achieves state-of-the-art accuracy and energy efficiency while reducing sensor power by 159× with test-chips prototyped in 65 nm CMOS.

Radio frequency (RF) transmission is the largest contributor of wireless sensor power, and hence, local in-sensor signal processing is preferred to continuous RF transmission¹, especially for bio-medical sensors since bio-medical signals, such as ECG, are naturally sparse and are usually sampled at a rate far exceeding their information rate. As an example, a low-power MedRadio transmitter consumes 67μW power² which is significantly higher than on-chip feature extraction^{1,3}. While there are several techniques to compress RF transmission^{1,4-9}, the compression ratio is typically limited to < 20×. Approaches to compress RF transmission so far have been sparsity-based data compression algorithms⁴⁻⁶, derivative-based adaptive sampling⁷, level-crossing sampling⁸, and adaptive resolution digitization^{1,9}. The aforementioned techniques have reported compression of transmission data by 2 ~ 16×. In contrast to the prior techniques, we propose to embed AI in the sensor itself to analyze each ECG segment and transmit only prediction score instead of ECG data or extracted features to reduce RF transmission by > 5000×.

It is challenging to design in-sensor neural networks with low energy consumption since artificial intelligence (AI) algorithms are computationally intensive. The majority of attempts¹⁰⁻¹⁷ to reduce energy consumption of AI circuits use (a) in-memory/near-memory computing (b) reduced precision computations. To address this energy bottleneck in wireless bio-medical sensors, we propose an analog signal processing neural network that directly processes analog ECG samples. The key contribution of this work is design and demonstration of an on-chip analog classifier comprising of a reservoir-computer (RC) followed by a 3-layer artificial neural network (ANN) that process analog ECG segments while reducing energy consumption by 13× compared to digital baseline (front-end ADC followed by digital ANN) and reduces overall sensor energy by 159× compared to direct transmission of digitized ECG segment. In-sensor processing AI circuits have been demonstrated primarily for CMOS image sensing applications by performing on-chip feature extraction¹⁸⁻²⁰ to reduce the amount of data that needs to be transmitted off-sensor. To the best of our knowledge, our work presents the first in-sensor AI circuits for analog-to-information conversion for significantly reducing transmission energy in wearable physiological sensors and extend the battery life of such sensors.

Figure 1 summarizes the analog-to-information methodology of this work. In a conventional ECG sensor, all ECG samples are digitized by analog-to-digital converter (ADC) and transmitted wirelessly. In contrast, we propose in-sensor AI (RC+ANN) that analyzes non-overlapping ECG segments and prediction score of the model is digitized and transmitted wirelessly. Thus, the RF transmission volume can be reduced by a large factor which is proportional to the number of samples in the ECG segments. To ensure that the in-sensor AI does not consume large energy and offset the advantages of reduced RF transmission, we propose hyper-dimensional computing using reservoir-computing paradigm that nonlinearly projects the input sensor data to a different hyper-plane for easy separation of distinct classes in the input signal by a read-out layer. The input and reservoir layers are not trained, and nonlinearity in analog computing is leveraged to create the nonlinear kernel in the

¹Electrical Engineering, University at Buffalo, Buffalo 14260, USA. ²School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe 85287, USA. ³Mayo Clinic, Phoenix 85054, USA. ✉email: arindam.sanyal@asu.edu

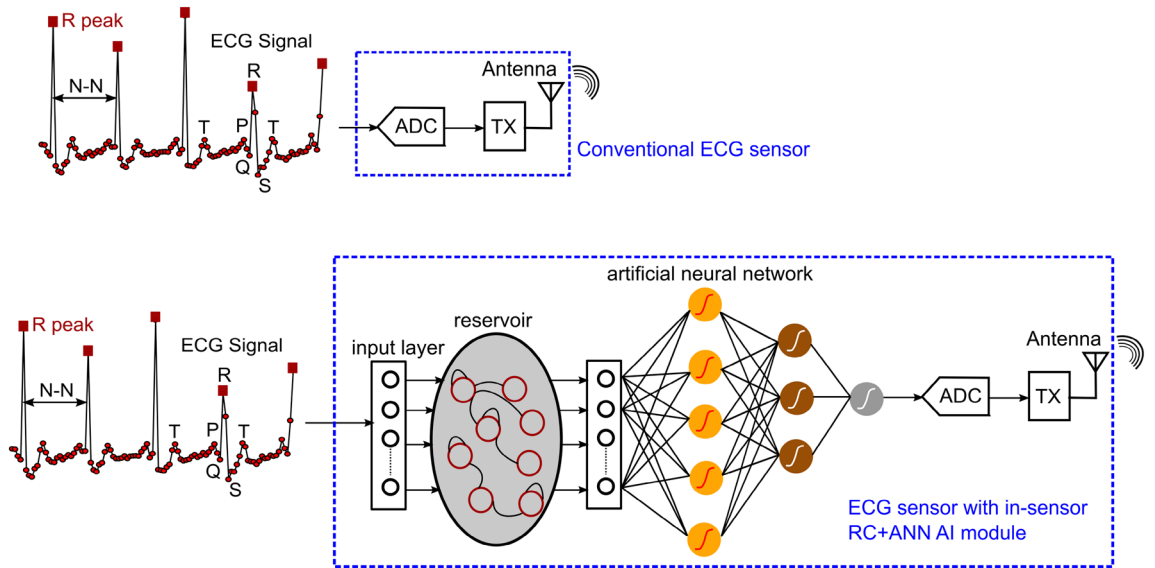
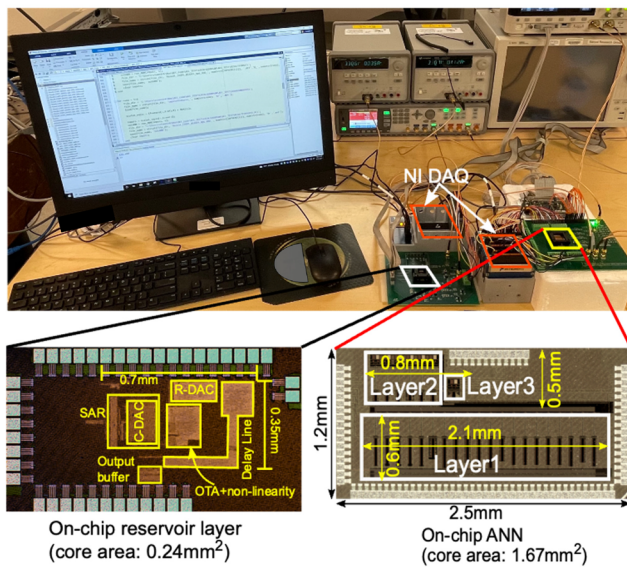


Figure 1. Overview of conventional ECG sensor compared to the proposed ECG sensor with in-sensor RC+ANN AI module for analyzing ECG segments and transmitting prediction score instead of all ECG samples.



Performance Summary	
Process (nm)	65
Computation type	RC+ANN
Supply voltage	1.2V
Input precision (bit)	12
Weight precision (bit)	1(RC)/4-4-6(ANN)
Energy/inference	17.4nJ
Efficiency (TOPS/W)	43.6
RF transmission reduction	5400x ¹
Sepsis detection accuracy	80.8% ²

¹for ECG segments with 6000 samples; ²mean of 4 test-chips at 4 hours before onset

Figure 2. Lab measurement setup with die microphotograph of reservoir and ANN test-chips and performance summary.

reservoir layer. In conventional analog computing nonlinearity is typically a limiting factor necessitating the analog circuits to run from high supply voltage and consume large power, thus trading off power efficiency for linearity. In contrast, analog circuits in the reservoir computer can be nonlinear and tolerate errors from incomplete settling, low gain and bandwidth. Thus, the in-sensor AI circuits can achieve high power efficiency as demonstrated in the Results section.

Results

The ANN and reservoir test-chips are fabricated separately in 65 nm CMOS, and integrated on printed circuit board level for lab measurement. National Instrument data acquisition (DAQ) module is used to load input data from a computer, and output of the ANN is captured using an oscilloscope and sent to the computer. A Matlab interface is used for communication between the computer, test chips and the NI DAQ. Figure 2 shows the measurement setup for the test-chips and a summary of the chip performance. The ANN has a core area of 1.67 mm² and the RC has a core area of 0.24 mm². The on-chip reservoir layer consumes 2 nJ/inference and the ANN consumes 7 nJ/inference while the off-chip reservoir input matrix multiplier consumes 8.4 nJ/inference at

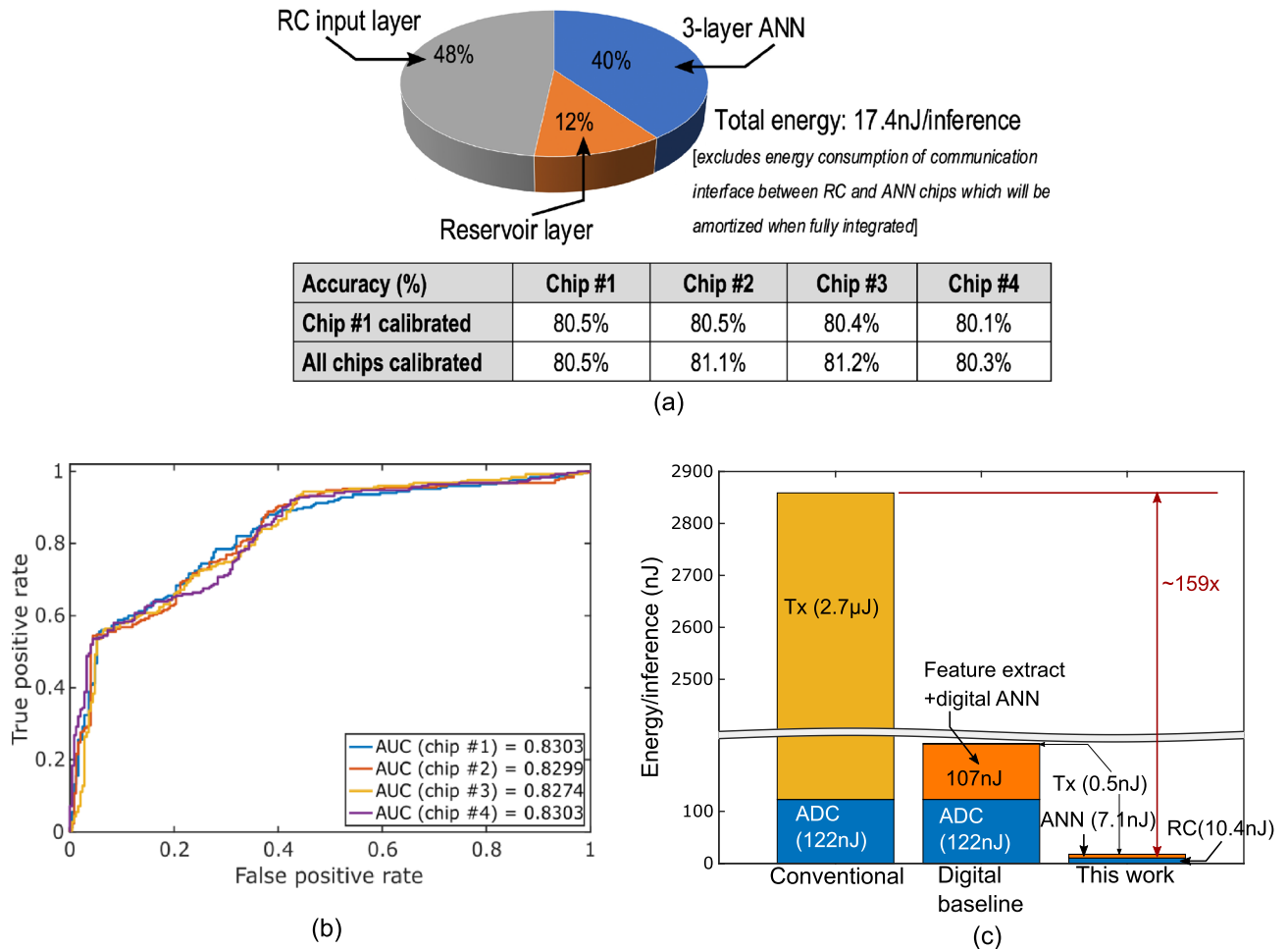


Figure 3. (a) Energy breakdown and measured accuracy of 4 test chips; (b) AUROC plots for 4 test-chips and; (c) comparison of energy/inference with conventional sensor technique of transmitting all digitized data and digital baseline.

1.2 V supply and operating at $F_s = 1$ kHz. The energy for communication between the test chips is not included since this will be amortized once the two chips are integrated on the same die.

The proposed technique is demonstrated on the publicly available MIMIC-III dataset which comprises of de-identified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The MIMIC-III dataset has 4559 patients with 40.2% sepsis patients and 59.8% non-sepsis patients, male/female split for sepsis and non-sepsis patients of 53.8%/46.2% and 55.8%/44.2% respectively. The dataset is randomly partitioned into 80% training samples and 20% test samples. ECG segments with 6000 samples are used for detection of sepsis 4 h before onset. All methods were carried out in accordance with relevant guidelines and regulations (Declaration of Helsinki).

Figure 3a shows energy breakdown and measured performance metrics for 4 test-chips. Due to offset induced by random mismatches introduced during chip fabrication, the decision threshold for each chip is shifted after fabrication. To correct for the offset, each chip is calibrated by applying the training samples to the test-chip and setting the decision threshold voltage to maximize prediction accuracy on training samples. The proposed RC+ANN combine detects sepsis with mean accuracy of 80.8% 4 h before onset. Figure 3b plots the receiver operating characteristic curves for 4 test-chips. The test-chips all show high area-under-the-curve (AUC) value indicating the ability to distinguish between sepsis and non-sepsis cases with high accuracy.

Figure 3c compares energy consumption of proposed RC+ANN with estimated energy of conventional technique of digitizing ECG segments and transmitting the digitized data, and digital baseline which performs feature extraction on digitized ECG segment followed by digital ANN before transmission of prediction scores. Transmission energy is assumed to be state-of-the-art 38 pJ/bit²¹, and the ADC for digitizing ECG segment is assumed to consume 5 fJ/conversion-step at 1 kHz and 12-bit resolution²². RC+ANN reduces energy/inference by 13× compared to digital baseline at 2% loss in accuracy, and by 159× compared to conventional technique. The proposed analog-to-information conversion technique reduces RF transmission by close to 5400×. This significant reduction in RF transmission is due to the fact that for every ECG segment the on-chip RC+ANN transmits only prediction score instead of transmitting all the digitized ECG samples in that segment. In this work, we consider non-overlapping ECG segments with 6000 samples for analysis. If the entire ECG segment is

	23	24	25	26	27	
	Nature Sc. Reports'19	ICHP'21	Plos One'21	Crit. Care Med. '18	Plos One'22	This work
AI model	RNN ¹	CNN ²	DL-ATT ³	Cox	LSTM	RC+ANN
Time-to-onset	7 h	6 h	4 h	4 h	1 h	4 h
Accuracy	–	84.7%	–	64%	–	80.8% ⁴
Sensitivity	0.88	0.87	0.49	0.89	0.85	0.84 ⁴
Specificity	0.84	0.86	–	0.90	0.64	0.75 ⁴
Vitals	10	0	7	10	6	1
Lab tests	6	13	17	30	27	0
EMR ⁵	0	0	3	19	3	0

Table 1. Comparison with state-of-the-art software AI models for sepsis prediction. ¹Long-short term memory. ²Recurrent neural network. ³Attention-based deep-learning model. ⁴Average of 4 test-chips. ⁵ Demographics and co-morbidities.

	10	11	12	13	14	15	28	16	17	This
	VLSI'18	JSSC'20	ISSCC'19	JSSC'18	JSSC'20	ISSCC'19	TCAS1'19	ISSCC'21	ISSCC'22	work
Compute type	SRAM					Analog				
	10T1C	12T	8T	6T	8T1C	ReRAM	ELM	3T DRAM	TD+SC	RC+ANN
Process	65 nm	65 nm	55 nm	65 nm	65 nm	55 nm	65 nm	65 nm	28 nm	65 nm
Weight precision	1	1	2	8	1	1	1	4	5	1(RC) 4–4–6 (ANN) ¹
Input precision	1	1	1	8	1	3	8	4	5	12(RC) 10–8–8 (ANN) ¹
Efficiency TOPS/W	658 ²	403 ²	18.4 ²	6.25 ²	671.5 ²	53.2 ²	2.9 ²	32.5 ²	13.3 ²	43.6 ³
Norm. eff. (TOPS/W) ⁴	658	403	36.8	400	671.5	159.5	23.2	520	332.7	528 ³
Throughput (GOPS)	9438	–	–	4.1	1638	–	–	–	100.8	0.8 ³
Area eff. (TOPS/W/mm ²)	104.4	6307	5808	1143	8290	787.6	13.8	157.6	264	277.9 ³

Table 2. Comparison with AI hardware accelerator. ¹Precision for 2 hidden layers and output layer; ²One MAC is considered as 2 OPS (multiplication and addition) and does not include energy for data movement and output activations; ³Nonlinearity, ADC and DAC of RC are considered as 1 operation each; ⁴Normalized efficiency is given by efficiency (TOPS/W) × input precision × weight precision.

digitized and transmitted, the transmission energy needed is 2.73 μJ at 38 pJ/bit for 6000 samples digitized with 12-bit resolution. Instead of transmitting each ECG samples in the segment, transmission of RC+ANN prediction requires digitization of only the RC+ANN output. Since the RC+ANN output is differential, the prediction score is digitized to 13-bit precision, and its transmission consumes only 0.5 μJ. Thus, the 5400× reduction in RF transmission is due to the reason that for each ECG segment, the proposed analog-to-information conversion techniques requires transmission of 13 bits instead of 72000 bits with naive transmission of all ECG samples.

Discussion

Comparison with state-of-the-art. Table 1 compares the proposed reservoir-computing model with state-of-the-art software AI models for MIMIC-III dataset. The proposed technique compares favorably with state-of-the-art using single modality sensor data source and no laboratory test results, which demonstrates feasibility of the proposed technique for at-home monitoring and is a key differentiation from state-of-the-art which requires multiple modality sensor data and/or laboratory test results. Table 2 compares efficiency in terms of tera-operations/watt (TOPS/W) of the RC+ANN with state-of-the-art in-memory computing AI accelerator macros. The proposed RC+ANN achieves competitive power efficiency as state-of-the-art AI accelerators and matrix multiplier macros even after including energy for data movement and output activations. Compared to the state-of-the-art hardware accelerators, the proposed RC+ANN circuit achieves competitive area efficiency (in terms of TOPS/W/mm²), but in general SRAM based macros have an order-of-magnitude better area efficiency. This is expected since SRAM cells are fully digital and have a much higher density than the analog switched-capacitor circuits used in our design.

Analog in-memory computation using switched-capacitor circuits. Energy efficiency of traditional AI computing systems are limited by communication costs of bringing together many input activations, and neuron weights, and distributing output activations in von-neumann architectures with separate memory and computing units. Wearable sensors have low energy budget which cannot accommodate conventional AI computing systems. In-memory computing (IMC) can break the von-neumann bottleneck by massively parallelizing computations and drastically reducing communication costs by performing computations using memory units. In this work, we perform analog IMC by using switched-capacitor circuits that store the ANN weights as capacitor values and intermediate results as charge on the circuit nodes. The complete vector matrix multiplications across all the neurons in each layer are performed simultaneously, and the results are stored locally in charge-domain on the shared top-plate of the capacitors. Apart from charge-domain IMC, SRAM based IMC is another popular technique for vector matrix multiplications using CMOS circuits. Compared to SRAM based IMC, switched-capacitor IMC has better linearity for vector matrix multiplications. This is because arithmetic computation through passive charge sharing/redistribution in switched-capacitor IMC is more linear and less sensitive to random variations introduced during chip fabrication than SRAM array since switched-capacitor circuits use ratios of capacitors for computation. In contrast, linearity of vector matrix multiplication using SRAM cells is fundamentally limited by nonlinear relationship between current and voltage on the bitlines where the accumulation happens, and the vector matrix multiplication results are not linear over the full dynamic range²⁹. In addition, matching capacitors in switched-capacitor circuits is easier than matching transistors in SRAM cells across large IMC array. However, the trade-off of using switched-capacitor IMC is the lack of re-configurability of ANN weights encoded as capacitor values in switched-capacitor IMC circuits, whereas the ANN weights can be easily re-configured in SRAM array by re-writing new weights into the array.

Limitations and future work. In this work, cleaned ECG signal from MIMIC-III dataset has been used to demonstrate the proposed technique of analog-to-information conversion through in-sensor AI. However, in practical at-home monitoring applications, the acquired ECG signal is likely to contain artifacts and will require analog front-end (AFE) with band-pass filtering before the ECG signal is sent to the RC+ANN combination. The AFE will consume additional power that will reduce the energy advantage of the proposed technique over the conventional method of transmitting all the sensor data. As an example, state-of-the-art AFEs for ECG sensor typically consume $1 - 8 \times$ power of ADC^{1,3,7,30}, and a similar AFE in front of our RC+ANN will result in $3.8 - 21 \times$ reduction in energy compared to naive transmission. Hence, the in-sensor AI technique shifts the design burden for high energy efficiency from transmitter to the AFE. The AFE energy efficiency can be potentially improved through inverter based amplifier design and inverter stacking^{31,32} to reduce overall energy consumption of the sensor.

A circuit level limitation of our work is the lower area efficiency our circuit than state-of-the-art SRAM based AI circuits. This area limitation is fundamentally due to the use of switched-capacitor circuit as building blocks for matrix multiplication which has lower area efficiency than SRAM cell. The area density (in terms of fF/ μm^2) of metal-on-metal capacitors used in this design does not scale as much as transistor area density, thus the area efficiency advantage of SRAM circuits over RC+ANN is likely to increase with CMOS technology scaling. Since computations in the input and reservoir layer can be nonlinear, a potential solution to improve area efficiency of the RC+ANN is to adopt SRAM arrays for matrix multiplications in the input and reservoir layers, and use switched-capacitor circuits for the ANN based read-out only where higher linearity is needed.

Methods

Reservoir-computer design. RC is a well-known computing paradigm that uses static nonlinearity to project the input signal to high-dimensional space, thus allowing easier separation of different input classes. No training is performed in the input or reservoir layers, and the weights are drawn from random distribution. While reservoir computing was invented almost two decades earlier and has been extensively used in the machine-learning literature, hardware implementation of reservoir computing have been mostly on optics/photonics platform with few analog silicon implementations³³⁻³⁵. In contrast to prior silicon RC, the proposed RC is based on the architecture in³⁶ and does not require large capacitors to realize biological time-constants which is energy-inefficient, and does not require background calibration for analog delay elements or nonlinearity element. Output of the RC with N reservoir neurons can be mathematically expressed as

$$\vec{R}_k[n] = H(G_i \vec{W} \times \vec{X}[n] + G_f \vec{W}_r \times \vec{R}_k[n-1]) \quad (1)$$

where \vec{X} is analog ECG input with D samples, \vec{W} is $N \times D$ input weight matrix, \vec{W} ($D \gg N$), \vec{W}_r is $N \times N$ inter-connection weight matrix for the reservoir layer, $H(\cdot)$ is nonlinear activation for RC, G_i is input scaling factor and G_f is feedback gain. As in³⁶, identity matrix is used for \vec{W}_r , which simplifies the hardware implementation since \vec{W}_r can be realized using a single-cycle delayed feedback. The restriction on \vec{W}_r is consistent with³⁷ which has shown through systematic investigation that a simple reservoir architecture with sparsely inter-connected reservoir provides comparable accuracy as more complicated reservoir architectures. G_i and G_f and N are set to 0.6, 0.1 and 63 respectively to optimize prediction accuracy and ensure stability of the reservoir, $D = 6000$ corresponding to 20 s ECG segments.

Figure 4 shows the circuit schematics of the reservoir layer and input layer and the timing diagram. Elements of \vec{W} are set to '0/1' which converts matrix multiplication in the input layer to addition. Switched-capacitor (SC) integrator is used to perform charge-domain accumulation and store partial results in the feedback capacitor, C_{intg} (Fig. 4a). The accumulated results from the input layer are sent to the reservoir layer shown in Fig. 4b. An operational-transconductance amplifier (OTA) is used to sum input to the reservoir layer with delayed feedback

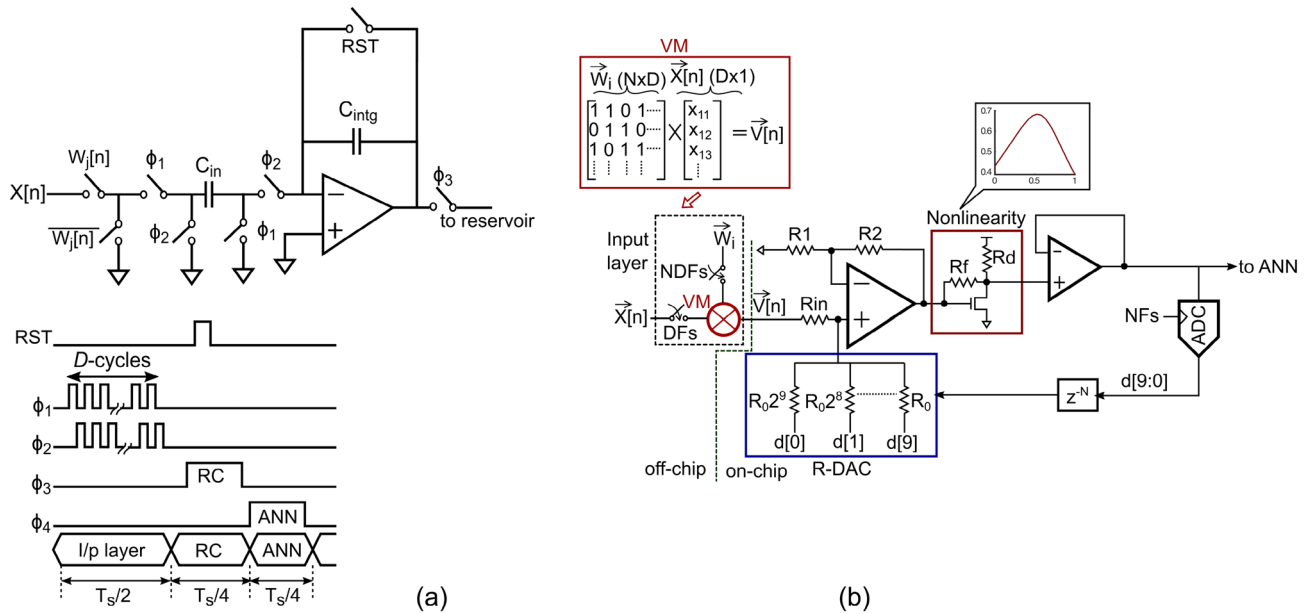


Figure 4. Circuit schematics for the reservoir-computer for analyzing ECG signals. (a) Schematic of the input layer of the reservoir showing the switched-capacitor multiplier (b) schematic of a single reservoir neuron.

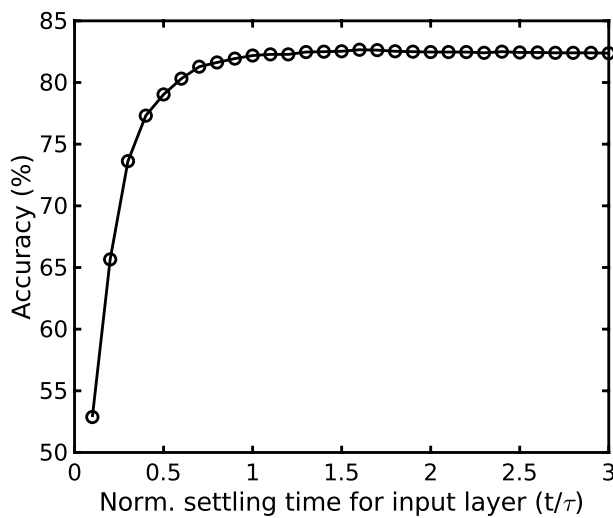


Figure 5. Simulated accuracy versus input layer bandwidth.

from the reservoir neuron. Output of the OTA represents the term within parenthesis in (1) and is passed through the nonlinearity $H(\cdot)$ which is implemented using a feed-forward common-source amplifier as shown in Fig. 4b. The non-linear activation function $H(\cdot)$ is based on Mackay-Glass nonlinearity. Output of the nonlinearity circuit is buffered and drives a 10-bit successive approximation register (SAR) ADC, and its delayed output is feedback to the input OTA through a resistive digital-to-analog converter (R-DAC). The reservoir layer is time-multiplexed to save on-chip area such that one physical neuron is used to realize N virtual neurons by operating the reservoir layer at NF_s where $F_s = 1/T_s$ is the frequency of operation of the RC+ANN and ECG input is sampled at DF_s . The ADC is used in the reservoir loop for accurate generation of N -cycle delay in the time-multiplexed feedback path since generation of precise analog delay is difficult in practice. The RC input layer is off-chip for this design to allow testing with different W .

In contrast to conventional analog design, the circuits components in the RC can be nonlinear since all nonlinearity is absorbed into the reservoir dynamics. Relaxed linearity requirements allow amplifiers in the input layer and RC as well as the ADC to be low bandwidth, which results in increased nonlinearity due to slewing and incomplete settling, but reduces both noise and power. Figure 5 shows the simulated accuracy from ECG analysis as a function of settling time in the switched-capacitor input layer, and time-constant of the amplifier (τ) is set to $T_s/8/D$. Size of the sampling capacitor in the input layer is set by noise and accuracy requirements. Input-referred noise in the input layer can be shown to be

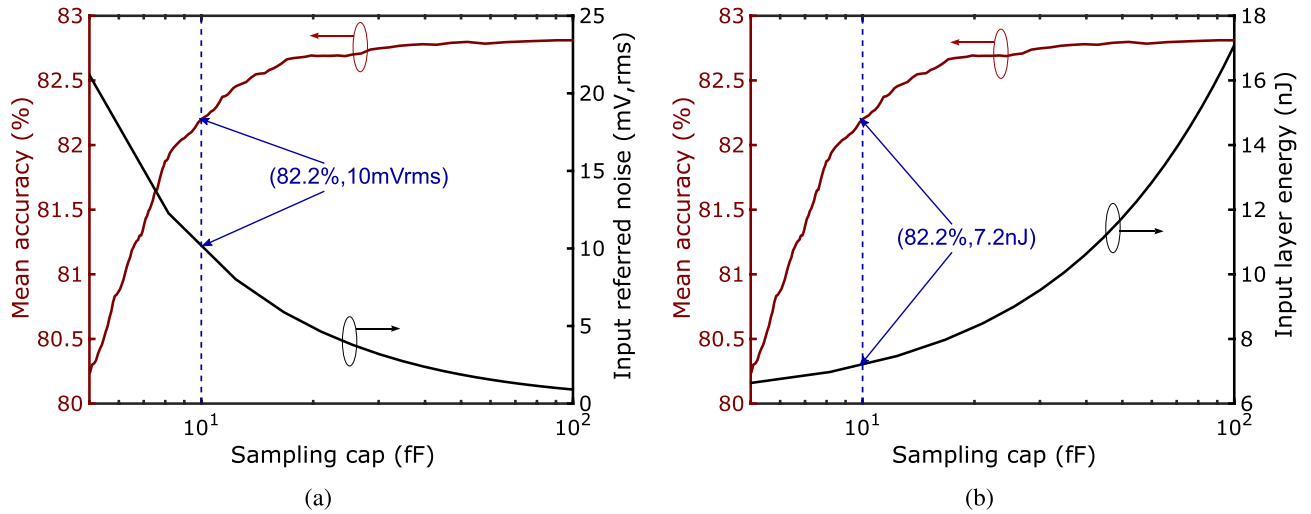


Figure 6. Simulated (a) accuracy and noise; and (b) energy of input layer versus sampling capacitor.

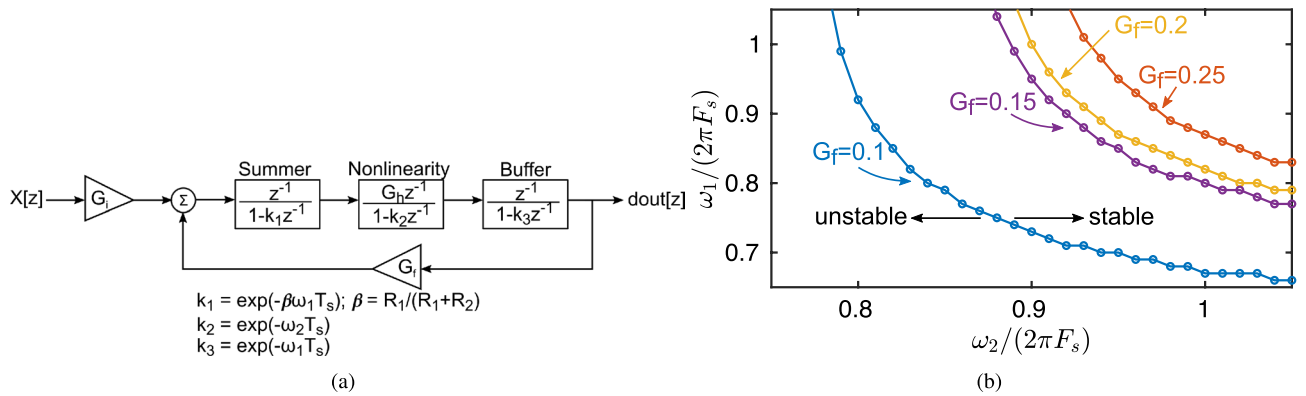


Figure 7. (a) Linearized model of the RC (b) stability contours.

$$\overline{V_{n,in}^2} = \frac{kT}{C_{in}} \frac{1}{1-\beta} + \frac{1}{\beta} \frac{kT}{C_{eq}} \left(\frac{4}{3}\right) \left(\frac{C_{intg}}{C_{in}}\right)^2 \tag{2}$$

where β is the feedback factor and $C_{eq} = C_L + (1 - \beta)C_{intg}$, C_L is the load capacitor. Keeping the feedback capacitor C_{intg} fixed to 400fF and load capacitor of 100fF, the sampling capacitor C_{in} is swept and the RC+ANN accuracy and energy of the input layer are plotted in Fig. 6. Based on the simulation results, C_{in} is set to 10fF.

The lower bound on bandwidth of amplifiers in the reservoir layer is set by stability requirements. Since the reservoir is strongly nonlinear, the RC loop has to be linearized around its operating point to theoretically analyze stability. The worst-case scenario from stability perspective occurs when the RC loop has the highest gain, corresponding to the highest gain of the nonlinearity function $H(\cdot)$ that occurs for the smallest input seen by the nonlinearity circuit. The highest possible gain for $H(\cdot)$ is found through simulations for different values of feedback gain, G_f . Figure 7a shows the discrete-time, linearized model of the RC with G_h denoting gain of $H(\cdot)$. The summing amplifier and the unity-gain buffer in Fig. 4b use the same OTA with unity-gain bandwidth of ω_1 and feedback factor of the summing amplifier is β , and 3-dB bandwidth of the nonlinearity circuit is ω_2 . Stability of the RC is analyzed by finding the roots of (3)

$$1 + \frac{z^{-3}}{(1 - k_1z^{-1})(1 - k_1z^{-1})(1 - k_1z^{-1})} = 0 \tag{3}$$

Figure 7b plots stability contours versus normalized values of ω_1 and ω_2 as a function of β . The stable region shrinks as G_f increases, and ω_1, ω_2 reduce. ω_1 and ω_2 are set to $2\pi \times 0.9F_s$ ($2\pi \times 0.9NF_s$ after time-multiplexing) for $G_f = 0.1$ to ensure a wide stability margin.

ANN model training and circuit design. The ANN has 20 neurons in the first hidden layer, and 6 neurons in the second hidden layer. The hidden layers use custom tanh activation function, while the output layer uses a custom softmax activation function. The voltage output of the softmax function is compared with a threshold voltage to generate the ANN decision, i.e., non-sepsis/sepsis. The activation circuits are designed

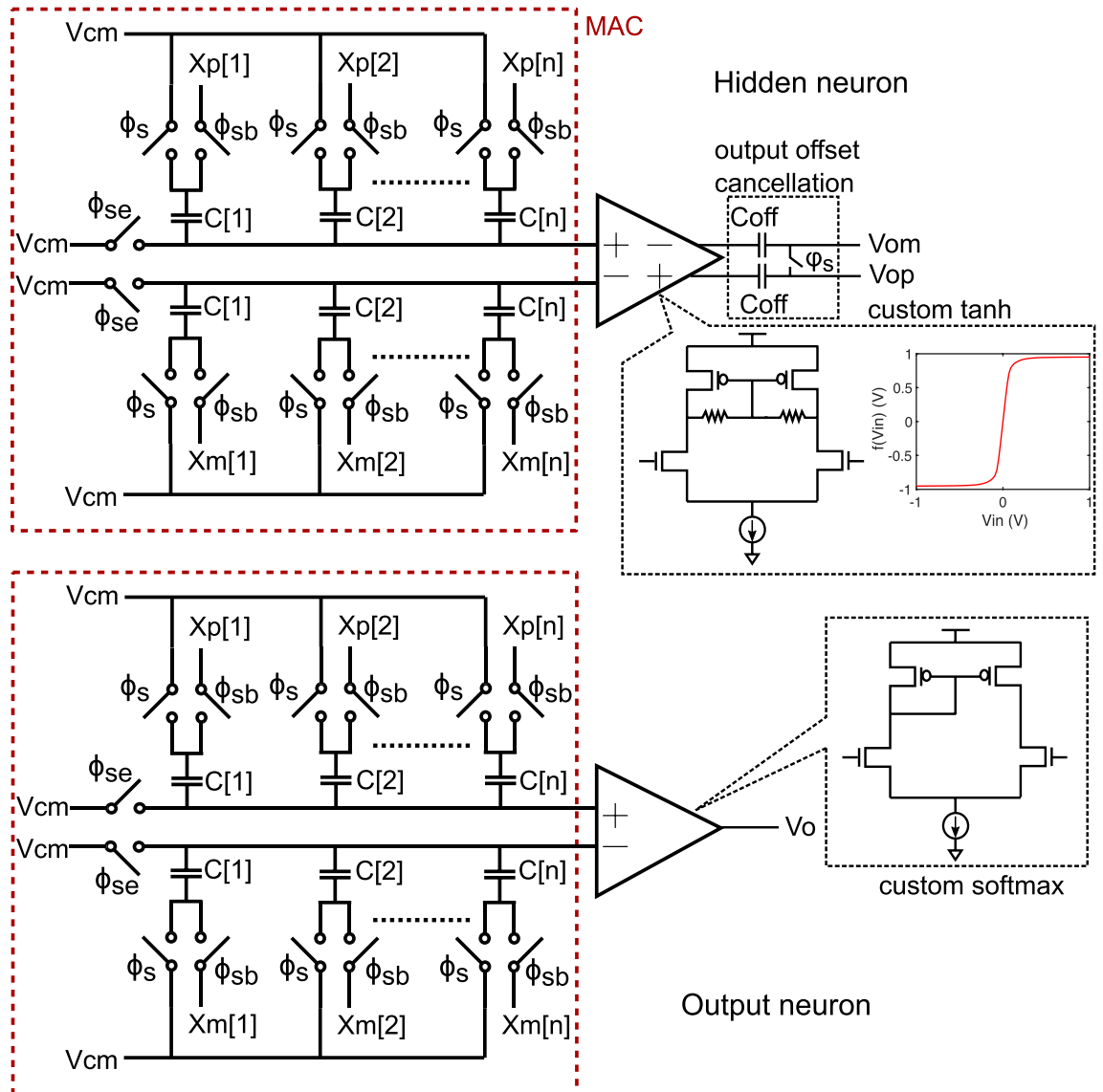


Figure 8. Circuit schematic of custom hidden and output neurons.

using single-stage, common-source differential amplifiers as shown in Fig. 8. The fully differential amplifiers in the hidden layers use output offset cancellation technique to reduce amplifier offset. Offset in the output layer is removed through foreground calibration as described later. The custom analog activation functions resemble their ideal, mathematical counterparts, but are not exactly the same. To ensure good matching between software ANN model and IC measurements, we use a hardware-software co-design methodology in which amplifier transfer curves, and their derivatives, are used to train the ANN model iteratively³⁸. Stochastic gradient descent is used to optimize the ANN model by minimizing the loss function at each epoch. Once the ANN is fully trained, the model weights are encoded as capacitor values in the SC-CIM. The ANN weights are quantized to 4-bit in the hidden layers, and 6-bit in the output layer. The weight quantization is done during the training iterations to minimize effect of quantization error. A 4fF unit capacitor is used to realize an LSB weight in the SC-CIM without degrading ANN accuracy due to capacitor mismatch.

Data availability

The dataset used in the current study is publicly available in the MIMIC-III database (<https://physionet.org/content/mimiciii/1.4/>). MIMIC-III integrates de-identified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, and makes it widely accessible to researchers internationally under a data use agreement³⁹.

Received: 9 September 2022; Accepted: 25 October 2022

Published online: 29 October 2022

References

- Mondal, S., Hsu, C.-L., Jafari, R. & Hall, D. A. A dynamically reconfigurable ECG analog front-end with a $2.5\times$ data-dependent power reduction. *IEEE Trans. Biomed. Circuits Syst.* **15**, 1066–1078 (2021).
- Mondal, S. & Hall, D. A. A $67\text{-}\mu\text{W}$ ultra-low power PVT-Robust medradio transmitter. In *2020 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, 327–330 (IEEE, 2020).
- Yan, L. et al. 24.4 A 680nA fully integrated implantable ECG-acquisition IC with analog feature extraction. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 418–419 (IEEE, 2014).
- Dixon, A. M., Allstot, E. G., Gangopadhyay, D. & Allstot, D. J. Compressed sensing system considerations for ECG and EMG wireless biosensors. *IEEE Trans. Biomed. Circuits Syst.* **6**, 156–166 (2012).
- Deepu, C. J., Heng, C.-H. & Lian, Y. A hybrid data compression scheme for power reduction in wireless sensors for IoT. *IEEE Trans. Biomed. Circuits Syst.* **11**, 245–254 (2016).
- Guo, W., Kim, Y., Tewfik, A. H. & Sun, N. A fully passive compressive sensing SAR ADC for low-power wireless sensors. *IEEE J. Solid-State Circuits* **52**, 2154–2167 (2017).
- Yazicioglu, R. F., Kim, S., Torfs, T., Kim, H. & Van Hoof, C. A $30\mu\text{W}$ analog signal processor ASIC for portable biopotential signal monitoring. *IEEE J. Solid-State Circuits* **46**, 209–223 (2010).
- Li, Y., Mansano, A. L., Yuan, Y., Zhao, D. & Serdijn, W. A. An ECG recording front-end with continuous-time level-crossing sampling. *IEEE Trans. Biomed. Circuits Syst.* **8**, 626–635 (2014).
- Trakimas, M. & Sonkusale, S. R. An adaptive resolution asynchronous ADC architecture for data compression in energy constrained sensing applications. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **58**, 921–934 (2010).
- Valavi, H., Ramadge, P. J., Nestler, E. & Verma, N. A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement. In *IEEE Symposium on VLSI Circuits*, 141–142 (2018).
- Yin, S., Jiang, Z., Seo, J.-S. & Seok, M. XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks. *IEEE J. Solid-State Circuits* **55**, 1733–1743 (2020).
- Si, X. et al. A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning. In *IEEE International Solid-State Circuits Conference-(ISSCC)*, 396–398 (2019).
- Gongondla, S. K., Kang, M. & Shanbhag, N. R. A variation-tolerant in-memory machine learning classifier via on-chip training. *IEEE J. Solid-State Circuits* **53**, 3163–3173 (2018).
- Jiang, Z., Yin, S., Seo, J.-S. & Seok, M. C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism. *IEEE J. Solid-State Circuits* **55**, 1888–1897 (2020).
- Xue, C.-X. et al. A 1Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors. In *IEEE International Solid-State Circuits Conference-(ISSCC)*, 388–390 (2019).
- Chen, Z., Chen, X. & Gu, J. A 65 nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44TOPS/W system energy efficiency. In *IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, 240–242 (2021).
- Seo, J.-O., Seok, M. & Cho, S. ARCHON: A 332.7 TOPS/W 5b variation-tolerant analog CNN processor featuring analog neuronal computation unit and analog memory. In *IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 258–260 (2022).
- Hsu, T.-H. et al. A 0.5-V real-time computational CMOS image sensor with programmable kernel for feature extraction. *IEEE J. Solid-State Circuits* **56**, 1588–1596 (2020).
- Hsu, T.-H. et al. A 0.8 V intelligent vision sensor with tiny convolutional neural network and programmable weights using mixed-mode processing-in-sensor technique for image classification. In *IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 1–3 (2022).
- Ma, T., Cao, W., Qiao, F., Chakrabarti, A. & Zhang, X. HOGEye: neural approximation of hog feature extraction in rram-based 3d-stacked image sensors. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 1–6 (2022).
- Mercier, P. P., Bandyopadhyay, S., Lysaght, A. C., Stankovic, K. M. & Chandrakasan, A. P. A sub-nW 2.4 GHz transmitter for low data-rate sensing applications. *IEEE J. Solid-State Circuits* **49**, 1463–1474 (2014).
- Murmann, B. Adc performance survey 1997–2021.
- Liu, R. et al. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Sci. Rep.* **9**, 1–9 (2019).
- Medina, M. & Sala, P. On the early detection of Sepsis in MIMIC-III. In *IEEE 9th International Conference on Healthcare Informatics (ICHI)*, 171–180 (2021).
- Rosnati, M. & Fortuin, V. MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis. *PLoS ONE* **16**, e0251248 (2021).
- Nemati, S. et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit. Care Med.* **46**, 547 (2018).
- Kaji, D. A. et al. An attention based deep learning model of clinical events in the intensive care unit. *PLoS ONE* **14**, e0211057 (2019).
- Chen, Y., Wang, Z., Patil, A. & Basu, A. A 2.86-TOPS/W current mirror cross-bar-based machine-learning and physical unclonable function engine for Internet-of-Things applications. *IEEE Trans. Circuits Syst. I* (66), 2240–2252 (2019).
- Dong, Q. et al. A 351TOPS/W and 372.4 GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications. In *ISSCC*, 242–244 (2020).
- Harpe, P., Gao, H., van Dommele, R., Cantatore, E. & van Roermund, A. 21.2 A 3nW signal-acquisition IC integrating an amplifier with 2.1 NEF and a 1.5 fJ/conv-step ADC. In *IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, 1–3 (2015).
- Shen, L., Lu, N. & Sun, N. A 1-V $0.25\text{-}\mu\text{W}$ inverter stacking amplifier with 1.07 noise efficiency factor. *IEEE J. Solid-State Circuits* **53**, 896–905 (2018).
- Chen, Y.-P., Blaauw, D. & Sylvester, D. A 266 nW multi-chopper amplifier with 1.38 noise efficiency factor for neural signal recording. In *IEEE Symposium on VLSI Circuits Digest of Technical Papers*, 1–2 (2014).
- Bauer, F. C., Muir, D. R. & Indiveri, G. Real-time ultra-low power ECG anomaly detection using an event-driven neuromorphic processor. *IEEE Trans. Biomed. Circuits Syst.* **13**, 1575–1582 (2019).
- Bai, K. & Yi, Y. DFR: An energy-efficient analog delay feedback reservoir computing system for brain-inspired computing. *ACM J. Emerg. Technol. Comput. Syst. (JETC)* **14**, 1–22 (2018).
- Chen, Y., Yao, E. & Basu, A. A 128-channel extreme learning machine-based neural decoder for brain machine interfaces. *IEEE Trans. Biomed. Circuits Syst.* **10**, 679–692 (2015).
- Chandrasekaran, S. T., Bhanushali, S. P., Banerjee, I. & Sanyal, A. A bio-inspired reservoir-computer for real-time stress detection from ECG signal. *IEEE Solid-State Circuits Lett.* **3**, 290–293 (2020).
- Rodan, A. & Tino, P. Minimum complexity echo state network. *IEEE Trans. Neural Netw.* **22**, 131–144 (2010).
- Chandrasekaran, S. T., Jayaraj, A., Karnam, V. E. G., Banerjee, I. & Sanyal, A. Fully integrated analog machine learning classifier using custom activation function for low resolution image classification. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **68**, 1023–1033 (2021).

39. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).

Acknowledgements

This work is supported in part by National Science Foundation grant CCF-1948331 and Air Force Research Laboratory under agreement number FA8650-18-2-5402. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

Author contributions

S.S. and S.B. designed the test-chips and performed lab measurements, I.B. and A.S. wrote the main manuscript text, A.S. prepared the figures and all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022