A 43.6 TOPS/W AI Classifier with Sensor Fusion for Sepsis Onset Prediction

Sudarsan Sadasivuni*, Sumukh Prashant Bhanushali‡, Imon Banerjee†, and Arindam Sanyal‡
*Department of Electrical Engineering, University at Buffalo, NY, USA; †Mayo Clinic, Phoenix, AZ, USA;
‡School of Electrical, Computer and Energy Engineering, Arizona State University, AZ, USA

***Abstract*—This work presents an artificial intelligence (AI) framework for real-time, personalized sepsis prediction four hours before onset through fusion of electrocardiogram (ECG) and patient electronic medical record. An on-chip classifier combines analog reservoir-computer and artificial neural network to perform in-sensor classification at 43.6 TOPS/W (normalized efficiency of 528 TOPS/W) which reduces energy by $155\times$ compared to conventional sensors and $4\times$ compared to state-of-the-art bio-medical AI circuits. The proposed AI framework predicts sepsis onset with state-of-the-art 92.9% accuracy on patient data from MIMIC–III. The proposed framework is non-invasive and does not require lab tests which makes it suitable for at-home monitoring.**

***Index Terms*—sepsis, artificial intelligence, in-memory computing, data fusion, artificial neural network, reservoir-computer**

I. Introduction

Sepsis is a life-threatening medical condition that arises when the body initiates an extreme response to an infection in the bloodstream. The key to treating sepsis is early detection. Real-time, at-home monitoring of at-risk patients using smart wearable is a potential solution for predicting sepsis onset and timely intervention. This work presents an artificial intelligence (AI) framework that combines patient electronic medical record (EMR) and electrocardiogram (ECG) data to automate risk prediction of sepsis onset without requiring a clinical expert in the loop. The proposed framework is shown in Fig. 1, and comprises three components – a) in-sensor processing AI circuit for analyzing ECG signal and predicting risk of sepsis onset; b) a classifier that predicts risk of sepsis onset from EMR – patient demographics (age, gender, race and ethnicity) and co-morbidity data; and c) a meta-learner that combines prediction results from ECG and EMR to predict risk of sepsis onset with high accuracy. Wireless transmission of continuous sensor data is energy inefficient since information rate of ECG signal is much lower than its sampling rate. In-sensor AI to classify ECG segments and transmitting prediction score instead of raw data can significantly reduce sensor energy which is dominated by transmission energy. However, integrating computationally intensive AI classifier into a resource constrained sensor is challenging. The majority of attempts [1]–[5] to reduce energy consumption of AI circuits use a) in-memory/near-memory computing b) reduced precision computations. With reduced transmission energy and optimized AI computations, front-end analog-to-digital conversion (ADC) and digital feature extraction becomes a major energy bottleneck for bio-medical sensors. To address this energy bottleneck, we propose an analog signal processing neural network that directly processes analog ECG samples without front-end ADC.

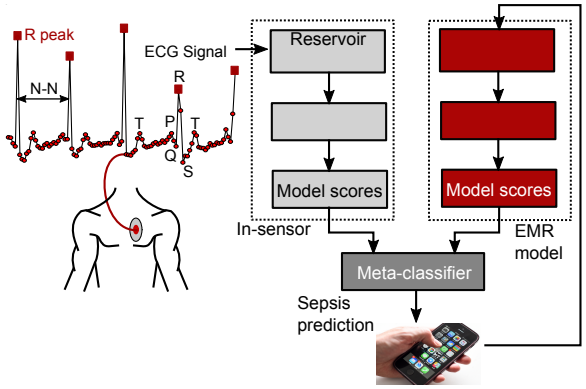

Fig. 1: Overview of the proposed AI framework for fusion of sensor and EMR data for sepsis onset prediction

The key contributions of this work are - a) demonstration of on-chip AI 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 $13\times$ compared to digital baseline (front-end ADC followed by digital ANN) and reduces transmission energy by $2700\times$ compared to direct transmission of digitized ECG segments; b) a fusion model that combines patient ECG and demographics to predict sepsis onset with high accuracy without requiring laboratory test results as in current state-of-the-art sepsis onset prediction works. The AI models are trained on de-identified data of 800 patients obtained from Emory University Hospital and tested on publicly available MIMIC–III dataset with 4559 patients.

II. On-chip AI classifier

A. 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. Hardware implementation of RC has been mostly on optics/photonics platform with few analog silicon implementations [6]–[8]. In contrast to prior silicon RC, the proposed RC is based on the architecture in [9] and requires neither large capacitors to realize biological time-constants nor background

calibration for analog delay elements or nonlinearity element.

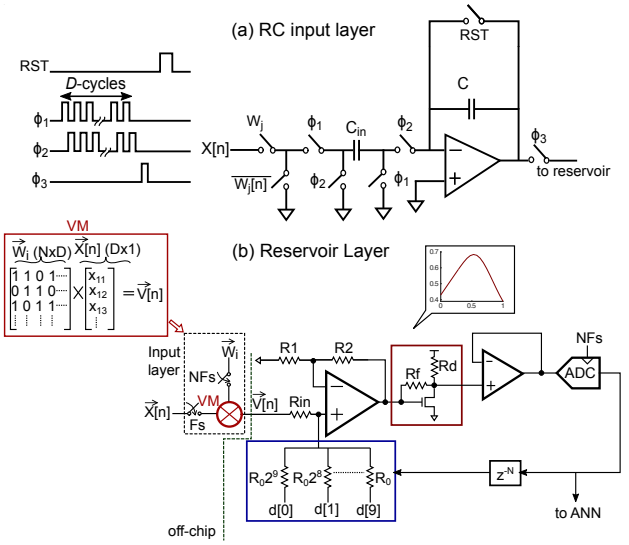

Fig. 2: Circuit schematics for the reservoir-computer for analyzing ECG signals

Output of the RC with N reservoir neurons can be mathematically expressed as

$$\vec{R_k}[n] = h\left(G_i\vec{W} \times \vec{X}[n] + G_f\vec{W_r} \times \vec{R_k}[n-1]\right) \qquad (1)$$

where $\vec{X}$ is analog ECG input with D samples, $\vec{W}$ is $N \times D$ input weight matrix, $\vec{W}$ ($D >> N$), $\vec{W_r}$ is $N \times N$ interconnection 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 in [9], 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. 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.

Fig. 2 shows the circuit schematics of the reservoir layer and input layer. 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. 2(a)). The accumulated results from the input layer are sent to the reservoir layer shown in Fig. 2(b). An operational-transconductance amplifier (OTA) is used to sum input to the reservoir layer with delayed feedback 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. 2(b). 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 fedback 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 is the frequency at which the ECG input is sampled by the input layer. 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 $\vec{W}$.

In contrast to conventional analog design, the circuit components (amplifiers and comparators) in the RC can be nonlinear and allows for slewing and incomplete settling which reduces both noise and power. The lower bound on bandwidth of amplifiers is set by stability requirements in the RC. 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 $H(\cdot)$. The highest possible gain for $H(\cdot)$ is found through simulations for different values of feedback gain, G_f. Fig. 3(a) 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. 2(c) uses 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 (2)

$$1 + \frac{z^{-3}}{(1-k_1z^{-1})(1-k_1z^{-1})(1-k_1z^{-1})} = 0 \qquad (2)$$

Fig. 3(b) 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.

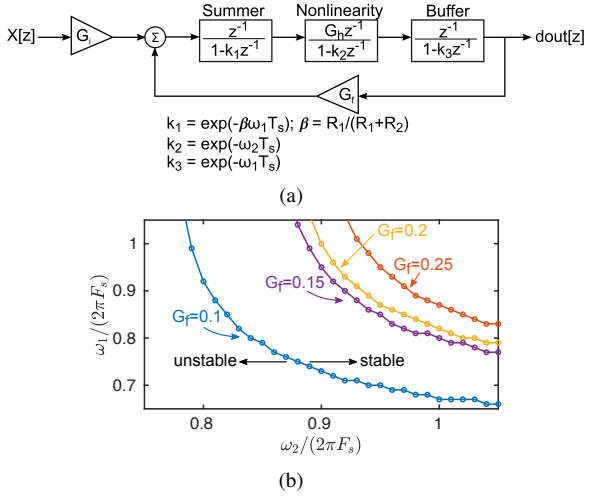

Fig. 3: a) Linearized model of the RC b) stability contours

B. 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 (V_{th}) to generate the ANN decision. The activation circuits are designed using single-stage, common-source differential amplifiers as shown in Fig. 4. 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 iteratively. 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-b in the hidden layers, and 6-b 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. The unit capacitor value is selected to ensure mismatch does not degrade ANN accuracy.

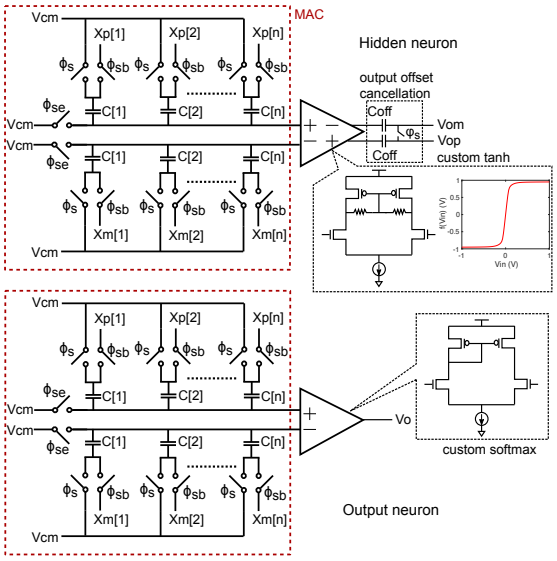

Fig. 4: Circuit schematic of custom hidden and output neurons

C. EMR model

An EMR AI model is used to predict sepsis onset from patient demographics and co-morbidities. A series of pre-processing steps are performed on the EMR data prior to analysis and model development. As the first data pre-processing step, standard data cleaning steps, including removing empty cells and special characters, are applied. Label encoding technique is used to convert categorical features to numerical quantities. The co-morbidity data is in the form of ICD-10 codes and is converted to vector format using Term Frequency- Inverse Document Frequency (TF-IDF) algorithm that computes a score for each word in proportion to its significance in the corpus. The TF-IDF tokenizer is trained on the training dataset for the vectorization. Finally the numeric representation of the categorical features and TF-IDF representation of the co-morbidities are combined using linear concatenation, and normalized by removing the mean and scaling to unit variance. Given the static nature of EMR, a single-point prediction model is used using only EMR data. The optimal value of the hyper-parameters is tuned through 10-fold cross validation on the training data. Table I shows accuracy of different EMR models on the test set, with random forest achieving the highest accuracy.

TABLE I: Sepsis prediction results with EMR models

	Linear SVM	Logistic regression	Random forest	ANN
Accuracy (%)	49	53	76	51

III. Measurement Results

Fig. 5 shows the measurement setup. The separately fabricated RC and ANN chips are integrated on printed circuit board level for lab measurement. The on-chip reservoir layer consumes 2nJ/inference and the ANN consumes 7nJ/inference while the off-chip reservoir input matrix multiplier consumes 8.4nJ/inference from 1.2V supply at 1kHz operating frequency. The energy for communication between the test chips will be amortized once the two chips are integrated on the same die.

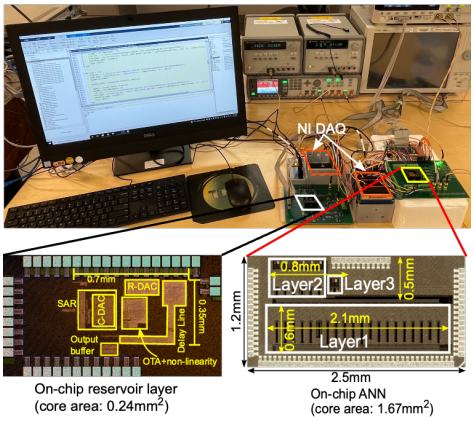

Fig. 5: Lab measurement setup with die photos

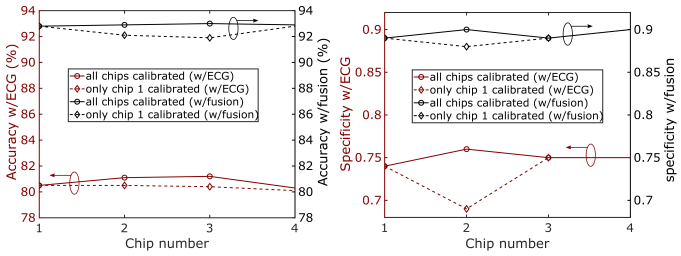

Fig. 6: a) Accuracy and b) specificity on MIMIC–III dataset

Patient EMR and ECG data is applied to the RC+ANN test-chips and the fusion AI models for predicting sepsis 4 hours before onset. Amplifier offset in the ANN output layer is calibrated by applying the training samples from Emory dataset to the test-chip and setting the decision threshold voltage to maximize prediction accuracy on Emory dataset. Fig. 6 shows the measured accuracy and specificity on the MIMIC–III dataset before and after fusion with all chips calibrated and with chips 2-4 using the calibrated threshold from chip 1. The test-chips achieve mean accuracy of 80.8% with ECG data and 92.9% after fusion.

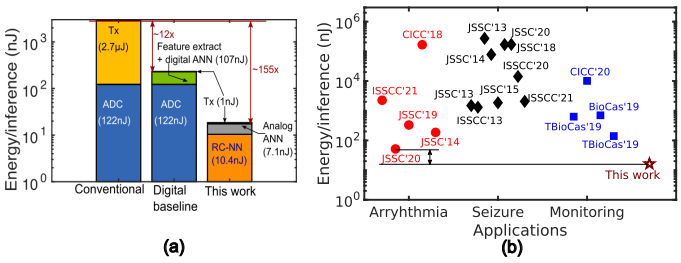

Fig. 7: Comparison with (a) baseline and (b) state-of-the-art AI ASICs for different bio-medical applications

TABLE II: Comparison with state-of-the-art AI models

	[10]	[11]	[12]	[13]	[14]	This work
Data	MIMIC–III					
Model	RNN	CNN	DL-ATT[1]	Cox	LSTM	Fusion
time-to-onset	7 hr	6 hr	4 hr	4 hr	1 hr	4 hr
Accuracy	–	84.7%	–	64%	–	92.9%[2]
Sensitivity	0.88	0.87	0.49	0.89	0.85	0.95[2]
Specificity	0.84	0.86	–	0.90	0.64	0.89[2]
Vitals	10	0	7	10	6	1
Lab tests	6	13	17	30	27	0
EMR[3]	0	0	3	19	3	5

[1]attention-based deep-learning model; [2]average of 4 test-chips; [3]demographics and co-morbidities

TABLE III: Comparison with AI accelerator macros

	[1] VLSI'18	[2] JSSC'20	[3] ISSCC'19	[4] JSSC'18	[5] JSSC'20	This work
Computation type	SRAM					Analog RC+ANN
	10T1C	12T	8T	6T	8T1C	
Process (nm)	65	65	55	65	65	65
Weight precision	1	1	2	8	1	1(RC) 4-4-6(ANN)[1]
Input precision	1	1	1	8	1	12(RC) 10-8-8(ANN)[1]
Efficiency TOPS/W	658[2]	403[2]	18.4[2]	6.25[2]	671.5[2]	43.6[3]
Norm. eff. (TOPS/W)[4]	658	403	36.8	400	671.5	528[3]

[1]precision for 2 hidden layers and output layer; [2]one MAC is considered as 2 OPS (multiplication and addition) and does not include energy for data movement and output activations; [3]excludes output layer of RC; nonlinearity, ADC and DAC of RC are considered as 1 operation each; [4]normalized efficiency is given by efficiency (TOPS/W) × input precision × weight precision

A. Comparison with state-of-the-art

Table II compares performance of the proposed fusion model with state-of-the-art software AI models. The proposed technique has the highest accuracy using single modality sensor data source and no laboratory test results which is a key differentiation from state-of-the-art. Table III compares efficiency (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 matrix multiplier macros even after including energy for data movement and output activations. Fig. 7(a) compares the proposed RC+ANN with direct transmission of all digitized sensor data, and digital baseline which performs in-sensor classification with digital ANN before transmission of prediction scores. Transmission energy is assumed to be state-of-the-art 38pJ/bit [15], and the ADC for digitizing ECG segment is assumed to consume 5fJ/conversion-step at 1kHz and 12-bit resolution [16]. RC+ANN reduces energy/inference by 13× compared to digital baseline at 3% loss in accuracy, and by 155× compared to conventional technique. Fig. 7(b) plots energy/inference of recent state-of-the-art AI ICs for different bio-medical applications. The proposed RC+ANN technique consumes the lowest energy/inference which is 4× lower than state-of-the-art.

IV. Conclusion

This work has presented a fusion AI framework for prediction of sepsis onset using in-sensor classification to reduce transmission energy. The proposed AI circuits are expected to benefit from technology scaling, and further improve energy efficiency, since the analog components do not need high linearity or gain.

Acknowledgment

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.

References

[1] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement," in *IEEE Symposium on VLSI Circuits*, 2018, pp. 141–142.

[2] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE JSSC*, vol. 55, no. 6, pp. 1733–1743, 2020.

[3] X. Si *et al.*, "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE ISSCC*, 2019, pp. 396–398.

[4] S. K. Gonugondla, M. Kang, and N. R. Shanbhag, "A variation-tolerant in-memory machine learning classifier via on-chip training," *IEEE JSSC*, vol. 53, no. 11, pp. 3163–3173, 2018.

[5] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism," *IEEE JSSC*, vol. 55, no. 7, pp. 1888–1897, 2020.

[6] F. C. Bauer, D. R. Muir, and G. Indiveri, "Real-time ultra-low power ECG anomaly detection using an event-driven neuromorphic processor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1575–1582, 2019.

[7] K. Bai and Y. Yi, "DFR: An energy-efficient analog delay feedback reservoir computing system for brain-inspired computing," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 14, no. 4, pp. 1–22, 2018.

[8] Y. Chen, E. Yao, and A. Basu, "A 128-channel extreme learning machine-based neural decoder for brain machine interfaces," *IEEE Trans. Biomedical Circuits and Sys.*, vol. 10, no. 3, pp. 679–692, 2015.

[9] S. T. Chandrasekaran, S. P. Bhanushali, I. Banerjee, and A. Sanyal, "A Bio-Inspired Reservoir-Computer for Real-Time Stress Detection From ECG Signal," *IEEE SSC-L*, vol. 3, pp. 290–293, 2020.

[10] R. Liu *et al.*, "Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[11] M. Medina and P. Sala, "On the early detection of Sepsis in MIMIC-III," in *IEEE 9th International Conference on Healthcare Informatics*, 2021, pp. 171–180.

[12] M. Rosnati and V. Fortuin, "MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis," *Plos one*, vol. 16, no. 5, p. e0251248, 2021.

[13] S. Nemati *et al.*, "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Critical care medicine*, vol. 46, no. 4, p. 547, 2018.

[14] D. A. Kaji *et al.*, "An attention based deep learning model of clinical events in the intensive care unit," *PloS one*, vol. 14, no. 2.

[15] P. P. Mercier *et al.*, "A sub-nW 2.4 GHz transmitter for low data-rate sensing applications," *IEEE JSSC*, vol. 49, no. 7, pp. 1463–1474, 2014.

[16] B. Murmann, "Adc performance survey 1997-2021." [Online]. Available: {http://web.stanford.edu/~murmann/adcsurvey.html}