Toward Real-Time, At-Home Patient Health Monitoring Using Reservoir Computing CMOS IC

Sanjeev Tannirkulam Chandrasekaran, *Member, IEEE*, Sumukh Prashant Bhanushali,
Imon Banerjee, and Arindam Sanyal, *Member, IEEE*

***Abstract*— This work presents a mixed-signal, reservoir-computing neural network (RC-NN) for at-home, real-time health monitoring using intelligent wearable device. The proposed technique is demonstrated on stress detection from electrocardiogram (ECG) signal, and heart diseases detection using a fusion artificial intelligence (AI) model that combines demographic and physiological information. The RC-NN uses a static, random reservoir layer with short-term memory to nonlinearly project input data to high-dimensional plane, and allow easy separation using linear AI model at the output layer. The RC-NN is designed in 65nm CMOS process, and detects stress and heart-diseases with mean accuracies of 92.8% and 86.8% respectively, while consuming 10.97nJ/inference and 2.57nJ/inference respectively.**

***Index Terms*— Machine learning, reservoir computing, health monitoring, cardiac diseases prediction, stress detection, data fusion and medical wearable.**

I. Introduction

CARDIOVASCULAR diseases (CVD) is the leading cause of death and disability, as well as healthcare spending, in the USA. Recent studies estimate that one person dies every 36 seconds in the USA from CVD which accounts for close to 25% of all deaths in the USA annually [1]. There are several underlying causes of CVDs including high blood pressure, obesity, long-term stress etc. While a large number of tools are available for early diagnosis of CVD, such as CT heart scans, chest X-rays, stress tests, and heart MRI, these tests need to performed in a clinical setting and are expensive. Hence, the impact of CVDs is disproportionately felt more on minority communities who have inadequate insurance coverage, and lack access to healthcare facilities on a regular basis [2]. Recent advances in artificial intelligence (AI) has the potential to enable equity in healthcare by automating risk prediction for CVDs from real-time analysis of patient physiological signals acquired using low-cost wearable device. Integrating AI circuits on wearables is a viable solution for real-time continuous monitoring that improves patient data security and increases device battery life-span by not transmitting raw patient data over the network.

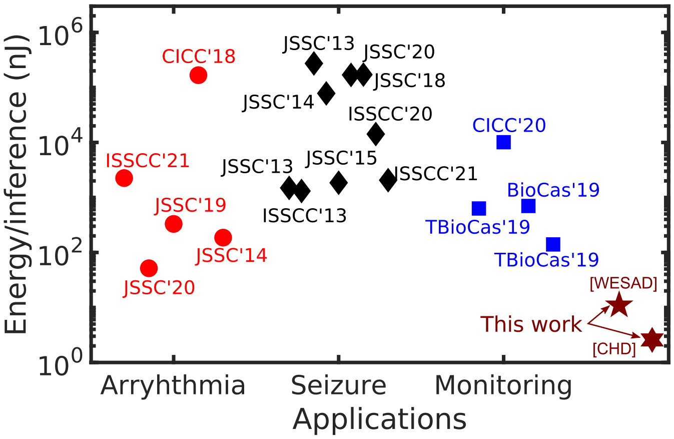


Fig. 1. Energy efficiency of state-of-the-art AI ASICs for different bio-medical applications.

Energy efficiency of conventional AI computing systems is limited by communication costs of bringing together many input activations and neuron weights, and distributing output activations which makes them unsuitable for resource constrained wearable devices. Prior works have attempted to reduce energy consumption through low bit precision and in-memory/near-memory computation, but state-of-the-art medical ML ICs still consume hundreds of nJ to few μJ for inference [3]–[7]. Fig. 1 shows energy/inference of recent state-of-the-art ML ICs for arrhythmia and seizure detection, and patient monitoring applications.

Instead of optimizing AI computing systems for wearables, this work presents an on-chip, mixed-signal reservoir computing neural network (RC-NN) for high energy efficient CVD detection. An RC-NN nonlinearly projects the input data to high-dimensional space using a static random, nonlinear reservoir layer and the output is typically obtained by a linear combination of the projected states. The reservoir

Manuscript received May 7, 2021; revised September 11, 2021 and November 13, 2021; accepted November 14, 2021. Date of publication November 16, 2021; date of current version December 13, 2021. This work was supported by the Air Force Research Laboratory under Grant FA8650-18-2-5402. This article was recommended by Guest Editor X. Zhang. *(Corresponding author: Sanjeev Tannirkulam Chandrasekaran.)*

Sanjeev Tannirkulam Chandrasekaran is with the Department of Electrical Engineering, University at Buffalo, Buffalo, NY 14260 USA (e-mail: stannirk@buffalo.edu).

Sumukh Prashant Bhanushali and Arindam Sanyal are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA.

Imon Banerjee is with Mayo Clinic, Phoenix, AZ 85054 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JETCAS.2021.3128587.

Digital Object Identifier 10.1109/JETCAS.2021.3128587

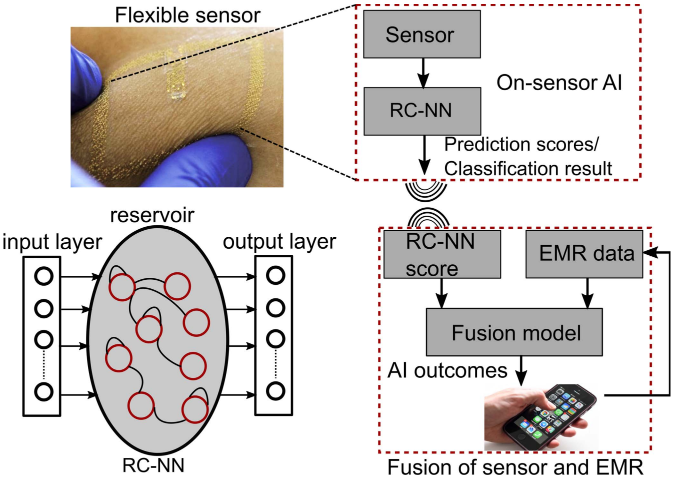


Fig. 2. Overview of the proposed health monitoring technique.

layer is untrained, and only the output layer is trained in a supervised manner. While reservoir computing was invented almost two decades earlier [8] and has been extensively used in the machine-learning literature, hardware implementation of reservoir computing have been mostly on optics/photonics platform [9]–[12], with few silicon implementations [7], [13]–[19]. The work in [7] developed a spiking RC-NN with analog neurons based on differential integrators biased in subthreshold. The analog neurons require large capacitors with values around 1pF which results in relatively large energy consumption in the order of μJ/inference. The work in [13] utilizes an analog delay chain in the feedback path to create virtual reservoir neurons. The analog delay elements require background calibration to maintain equal delay contributions from each neuron. The work in [14] presents an extreme learning machine (ELM) which is equivalent to RC-NN with a memoryless reservoir. The reservoir neurons are realized using current-controlled ring oscillators (CCOs) which acts as saturating nonlinearity. However, high sensitivity of CCOs to environmental conditions will require background calibration. The work in [15] presents a random-projection neural network with differential-amplifier for implementing sigmoid nonlinearity that leverages mismatch and offset introduced during fabrication to create a diverse pool of hidden neurons to increase encoding capacity of the network. Reference [16] presents an ELM with current-mirror based crossbar arrays for random projection in analog domain, and implements nonlinear activation and output layers in digital domain, while consuming 114nJ/classification. References [17]–[19] presents digital, ensemble ELMs for anomaly detection and consumes 24-477nJ/classification. The analog reservoir layer in this work uses a feedforward common-source amplifier for creating strong nonlinearity, and delayed feedback loop to impart memory to the reservoir layer. In contrast to prior silicon RC-NNs, the proposed reservoir neurons do not require large capacitors to realize biological time-constants, and have low sensitivity to environmental conditions which obviates need for background calibration.

Fig. 2 shows the driving vision of this work. The proposed RC-NN will be part of physiological sensor that will monitor patient vitals in real time. The RC-NN output will be directly used as classification result and displayed on a mobile device, or combined with electronic medical record (EMR) data through cloud-based fusion AI model for precise and personalized healthcare outcome. To that end, performance of the proposed RC-NN is demonstrated on two dataset - WESAD [20] and Cleveland Heart Disease (CHD) [21] dataset. ECG chest sensor data from the WESAD dataset is directly applied to the proposed RC-NN without feature extraction for predicting stress. We perform fusion of demographic and physiological vitals from the CHD dataset to identify if a patient has heart diseases indicated by narrowing of the epicardial artery. The proposed RC-NN prototype achieves similar or better performance than ideal software models reported in literature for both dataset. Section II presents the proposed RC-NN architecture, circuit design, and measurement results on WESAD dataset, while Section III presents fusion model and measurement results on CHD dataset. The conclusion is brought up in Section V.

II. Stress Detection From ECG Signal

The WESAD dataset [20] contains multi-modal, physiological sensor data of 15 patients collected from wrist and chest at 700Hz sampling rate. The sensor data for each patient is recorded for approximately 2 hours out of which the patient was in a baseline condition for 20 minutes and under stress for roughly 10 minutes. The sensor data is annotated as baseline, stress, amusement or meditation conditions. For this work, we used ECG signal from the chest and performed binary stress versus baseline classification. We used 5 seconds window segments of ECG signal, and collected 20,000 ECG segments having roughly equal proportions of stress and baseline events.

A. RC-NN Architecture

Fig. 3 shows the proposed RC-NN architecture with 3 layers – input, reservoir and output. Weights in the input and reservoir layer are typically drawn from random distributions, and only the output layer is trained. The simple architecture and reduced training requirements make RC-NN very attractive candidate for low energy wearables. The RC accepts an input vector $\vec{X}$ with D features, multiplies it with an $N \times D$ input weight matrix $\vec{W}$, and passes the result to the reservoir which performs non-linear projection with N neurons. State of the k-th neuron is expressed mathematically as shown in Fig. 3 where $\vec{W}_r$ is the $N \times N$ inter-connection weight matrix for the reservoir layer, $H(\cdot)$ is the non-linearity function, G_i is input scaling factor and G_f is feedback gain. The interconnect matrix $\vec{W}_r$ is typically sparsely filled and provides memory to the reservoir layer which allows the RC-NN to exhibit properties of high dimensionality with a small number of neurons [22]. The reservoir layer outputs are read-out by a memoryless output layer which typically performs linear, weighted combination of the NP layer outputs to provide classification/prediction result. To further reduce energy consumption, elements in the input and reservoir weight matrices are restricted to {0/1} which replaces multipliers by adders

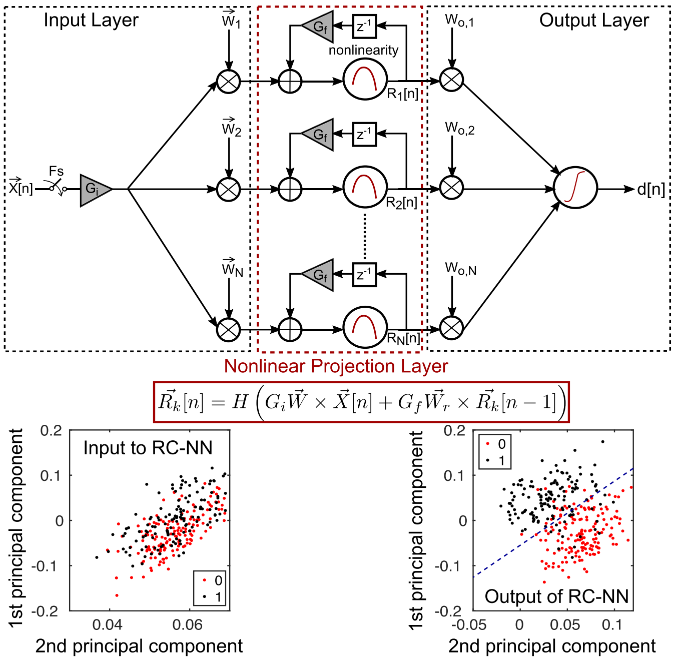


Fig. 3. RC-NN architecture showing linear separability of data classes after passing through reservoir.

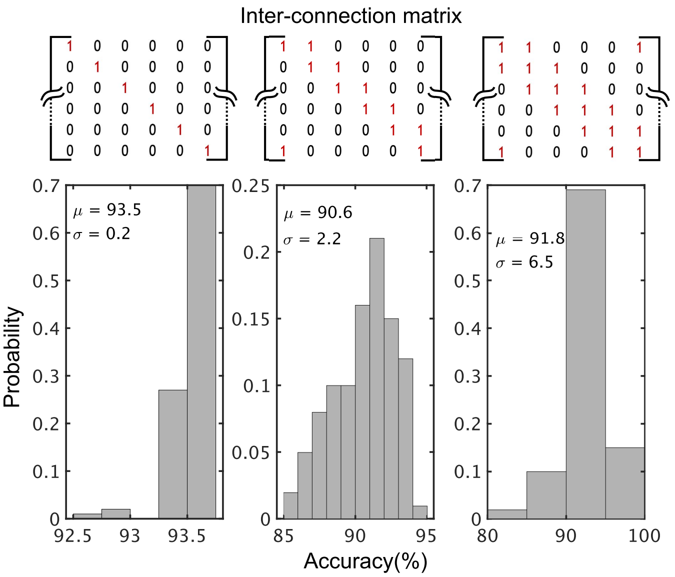


Fig. 4. Simulated classification performance for different inter-connection architectures in the reservoir layer.

and reduces hardware cost. D is set to 350 for the WESAD dataset which corresponds to 350 samples of ECG data in each inference window.

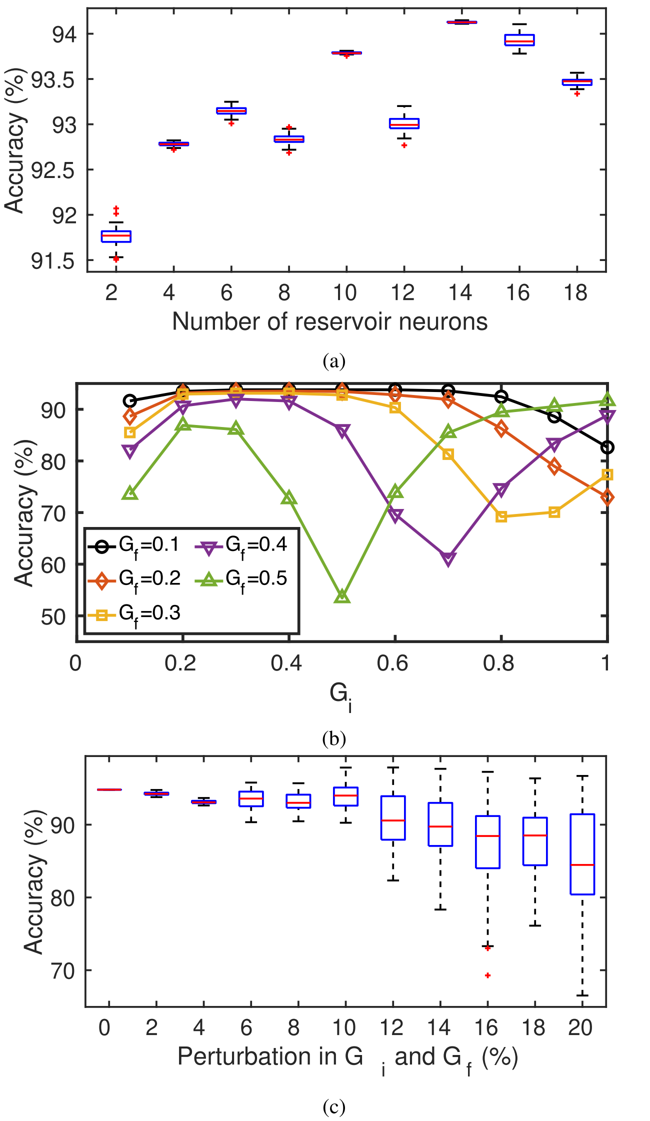
Optimizing an RC-NN typically requires balancing several parameters – reservoir size, sparsity of input connections and reservoir, and ranges for input scaling factor and feedback gain. As in [23], we set the following restrictions on the RC-NN design to simplify the parameter optimization space – 1) same input scaling factor, G_i, for all inputs, 2) same feedback gain G_f for all reservoir neurons, and 3) a single weight value for all reservoir connections. The sparsity of input connections is not optimized, and elements in the input weight matrix $\vec{W}$ are randomly set to $\{0/1\}$. This leaves 4 parameters for optimization – reservoir sparsity, G_i, G_f and N. To optimize sparsity of reservoir layer, we select 3 simple inter-connection matrices, $\vec{W}_r$, as shown in Fig. 4 based on templates provided in [23] and simulate the RC-NN on the WESAD dataset. The dataset is randomly partitioned into 80% training and 20% test data. The RC-NN model is simulated 100 times for each $\vec{W}_r$, and the input layer weights are selected randomly each time. Fig. 4 shows the mean and standard deviation of classification accuracy on the test set for the 3 inter-connection weight matrices. The best accuracy is obtained for the identity inter-connect matrix which is realized through 1-cycle delayed feedback to each neuron.

Our architecture also allows easy change of reservoir sparsity through tapping the output at different locations in the feedback delay chain. As an example, if the reservoir requires connection between adjacent neurons, i.e, if $W_r[i][j+1] = 1$, feedback from the output of $N-1$-th elements of the delay chain to the input will satisfy the sparsity requirement. Similarly, if the reservoir requires both self-connections and connection between adjacent neurons, i.e., $W_r[i][j] = 1$ and $W_r[i][j+1] = 1$, then the feedback connection has to be made from output of N-th and $N-1$-th elements of the delay chain, and the architecture will need 2 DACs. The reservoir sparsity can be set arbitrarily by drawing the appropriate feedback connections, but this comes at the cost of increased area and energy due to increase in number of DACs.

Fig. 5a) shows the mean and standard deviation of simulated classification accuracy as a function of reservoir neurons with the same simulation setup as for reservoir inter-connection matrix optimization. For this design, 10 reservoir neurons are used. Fig. 5b) shows the mean and standard deviation of simulated classification accuracy as a function of (G_i, G_f) for $N = 10$. The highest mean accuracy is obtained for $0.2 < G_i < 0.7$ and $G_f = 0.1$. For this design, $(G_i, G_f) = (0.6, 0.1)$ is used. Small value of G_f indicates that the reservoir layer requires short-term memory for predictions. Fig. 5c) shows the mean and standard deviation of simulated accuracy as (G_i, G_f) are varied independently from 0% to 20%. The mean accuracy remains $\sim$ 93% for perturbation $\leq$10%. G_i and G_f are defined as ratios of resistors in the design, and are not expected to vary significantly.

B. Stability of RC-NN

The RC-NN needs to be stable for the predictions to be repeatable. The strong nonlinearity of $H(\cdot)$ makes it difficult to analyze stability of the closed-loop system using standard techniques based on location of poles. The RC-NN needs to be linearized around an operating point before stability analysis. The worst-case scenario from stability perspective occurs when the RC-NN 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. We simulated the RC-NN on the entire WESAD dataset, repeated the simulation 100 times for different random values of input weight matrix, $\vec{W}$, and extracted



(a)

(b)

(c)

Fig. 5. Simulated accuracy (a) as a function of number of reservoir neurons; (b) (G_i, G_f); (c) with perturbations in G_f and G_i.

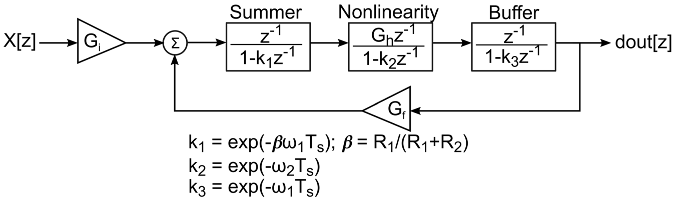
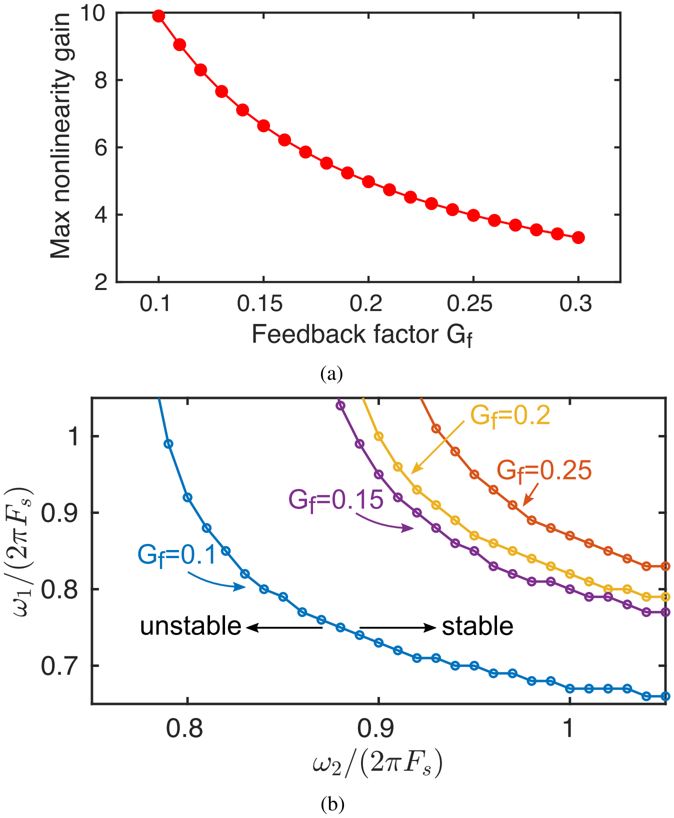

X[z]
G_i
Σ
Summer
$\frac{z^{-1}}{1-k_1z^{-1}}$
Nonlinearity
$\frac{G_hz^{-1}}{1-k_2z^{-1}}$
Buffer
$\frac{z^{-1}}{1-k_3z^{-1}}$
dout[z]
G_f
$k_1 = \exp(-\beta\omega_1T_s)$; $\beta = R_1/(R_1+R_2)$
$k_2 = \exp(-\omega_2T_s)$
$k_3 = \exp(-\omega_1T_s)$

Fig. 6. Linearized mathematical model of the RC-NN.

the input and output swings of the nonlinearity circuit, and gain of the nonlinearity circuit. The mean (±standard deviation) of voltage swing at the input and output of the nonlinearity circuit are 56mV(±0.49mV) – 560mV(±0.5mV), and 460mV(±0.28mV) – 681mV(±0.5mV) respectively. The gain of the nonlinearity circuit varies from 8.3(±0.067) – 1.22(±0.001). The linearized discrete-time RC-NN model for stability analysis is shown in Fig. 6. The summer and the buffer uses the same OTA and has unity-gain bandwidth of ω_1 and feedback factor of the closed-loop summer is β.



(a)

(b)

Fig. 7. (a) Maximum nonlinearity gain as a function of G_f for stable RC-NN (b) stability contours as a function of G_f and bandwidths of OTA and nonlinearity circuit.

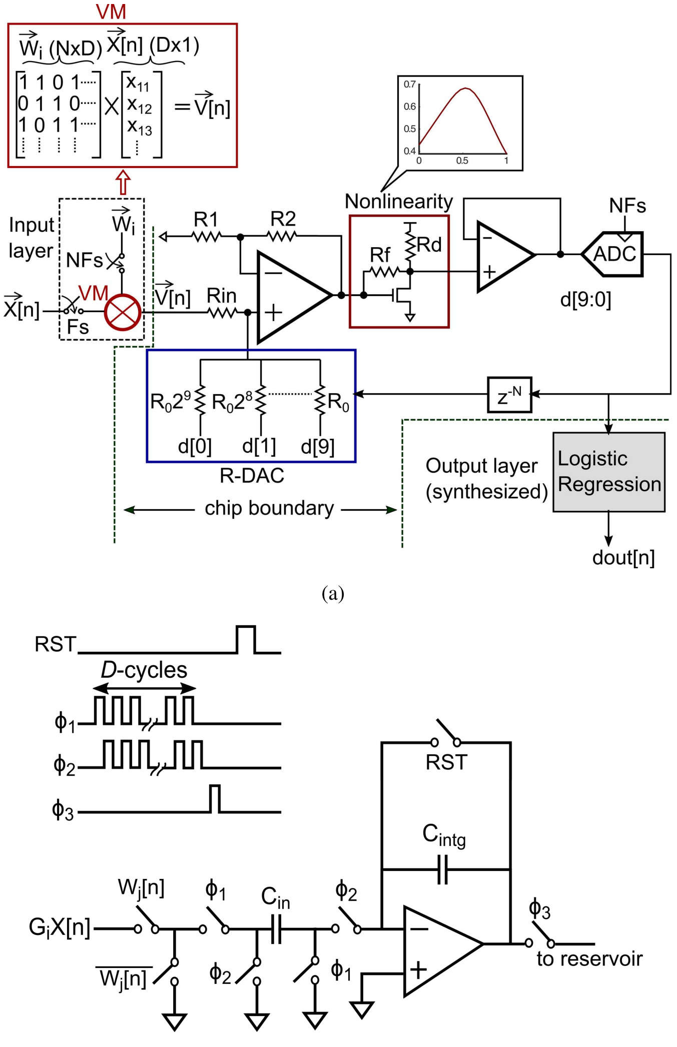
The nonlinearity circuit is replaced by a linear amplifier with dc gain of G_h and 3-dB bandwidth of ω_2. Stability of the RC-NN is analyzed by finding the roots of (1)

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

To analyze stability of the RC-NN, we first find the maximum value of G_h that results in a stable RC-NN for different values of feedback scaling factor G_f, and the result is plotted in Fig. 7(a). The maximum allowed value of G_h reduces as G_f increases. Fig. 7(b) plots the stability contour of the RC-NN as a function of the OTA unity-gain bandwidth and 3-dB bandwidth of the nonlinearity circuit for different values of G_f. The RC-NN is stable in the region to the right of the contour plots. For each value of G_f, G_h is set to the maximum allowed value from Fig. 7(a). RC-NN is stable for a wider range of (ω_1, ω_2) for lower values of G_f which is intuitive. G_f is set to 0.1 from the classification accuracy requirement in Fig. 5(b), and ω_1 and ω_2 are both set to $2\pi \times 0.9F_s$ (where F_s is the sampling frequency) to ensure the RC-NN has sufficient stability margin. In the circuit implementation of the RC-NN, the reservoir layer is time-multiplexed, and ω_1 and ω_2 are scaled up by the number of reservoir neurons, N, since the reservoir runs at NF_s.

C. RC-NN Circuit Design

Fig. 8(a) shows circuit implementation of the RC-NN. In order to save area, the reservoir layer is time-multiplexed



(a)

(b)

Fig. 8. (a) Time-multiplexed RC-NN circuit (b) single slice of input layer.

such that only one physical neuron is used. Computations in the reservoir layer are performed in mixed-signal domain which eliminates memory access associated with storing intermediate results in a digital implementation. The input is summed with feedback signal using an indirect miller-compensated OTA, while a common-source amplifier with resistive feedforward path implements the non-linear activation function $H(\cdot)$ in the NP neuron. The non-linear activation function is based on Mackay-Glass nonlinearity as used in [24]. Output of the non-linear activation is digitized using a successive approximation register (SAR) ADC with unit DAC capacitance of 2.4fF. Placing the ADC inside the feedback loop in the neurons allows accurate construction of N-cycle delayed feedback. Logistic regression (LR) is used for the output layer which is implemented off-chip through digital synthesis. Vector matrix-multiplication in the input layer is performed off-chip to allow demonstration of multiple dataset with different number of features using the same RC-NN. Non-linearity due to static mismatches in the SAR DAC or feedback DAC are absorbed into overall non-linearity of the reservoir neuron and does not need correction.

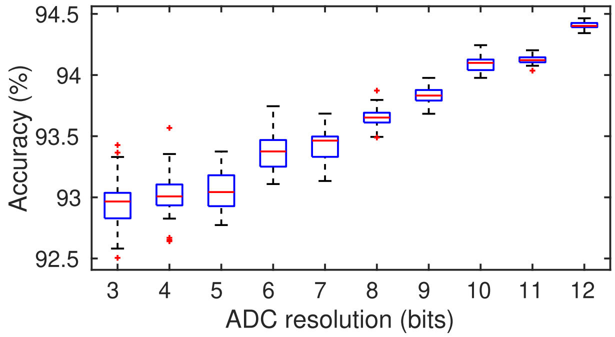


Fig. 9. Simulated accuracy as a function of ADC resolution.

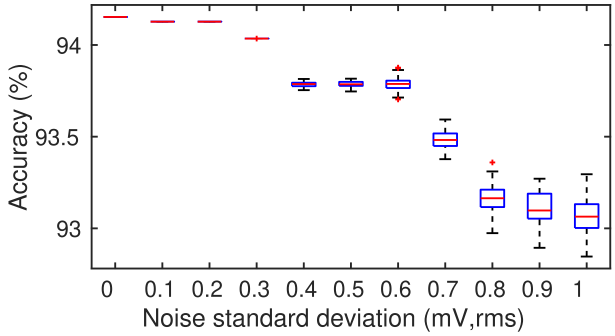
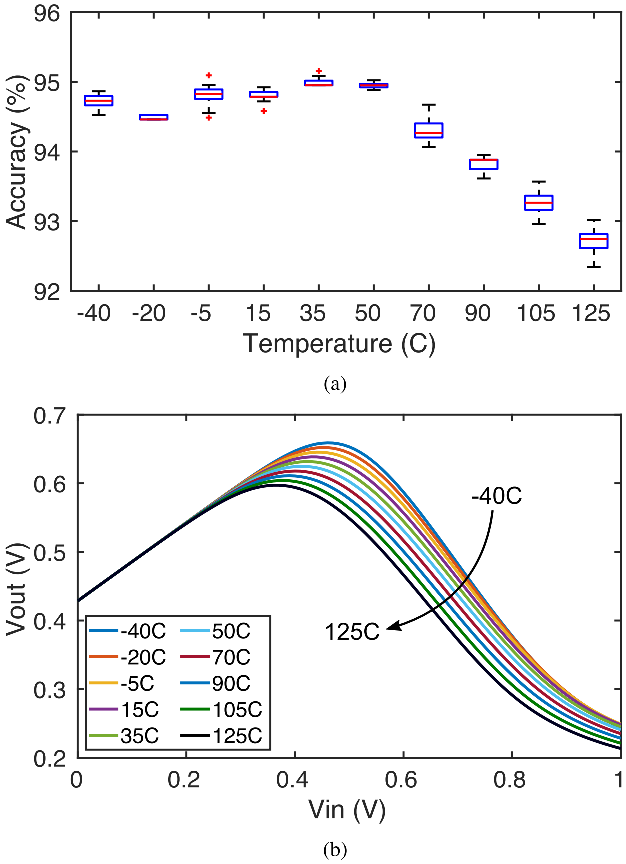


Fig. 10. Simulated accuracy versus noise.

Fig. 8(b) shows the j-th slice of the input layer with N slices with associated timing diagram. The same input signal is applied to all the slices. A switched-capacitor integrator is used for accumulating partial sums of $\vec{W}_j \times X$ over D-cycles which is stored on the integrating capacitor, C_{intg}. The amplifier in the integrator can be low gain and low bandwidth since nonlinearity due to incomplete settling and gain error is absorbed into the reservoir nonlinearity. The input layer is off-chip in this work to allow using the same reservoir layer with different dataset having different number of features.

Fig. 9 shows simulated classification accuracy as a function of ADC resolution. A 10-bit ADC resolution is selected for this design for high classification accuracy. Fig. 10 shows the simulated classification accuracy as a function of reservoir noise referred to the ADC input. For each value of noise standard deviation, the ESN is simulated 100 times, and Fig. 10 shows mean and standard deviation of classification accuracies. The R-DAC, OTA-summer, unity gain buffer, and non-linearity contribute 0.34mV,rms noise at referred to ADC input, while the ADC has an input referred noise of 0.54mV,rms, which results in a total noise of 0.64mV,rms. Based on the noise simulation result shown in Fig. 10, the RC-NN is expected to have a mean classification accuracy of 93.5% with standard deviation of 0.43%. Fig. 11(a) shows the simulated accuracy as a function of temperature in the range of −40°C – 125°C. The mean accuracy changes by 2.5% over the temperature range. The accuracy starts dropping beyond 70°C. The reason for this drop is due to shift in the nonlinearity transfer function $H(\cdot)$ as shown in Fig. 11(b). At low input voltages, the output of $H(\cdot)$ is set by resistor ratios and does not vary with temperature. At input voltages



(a)

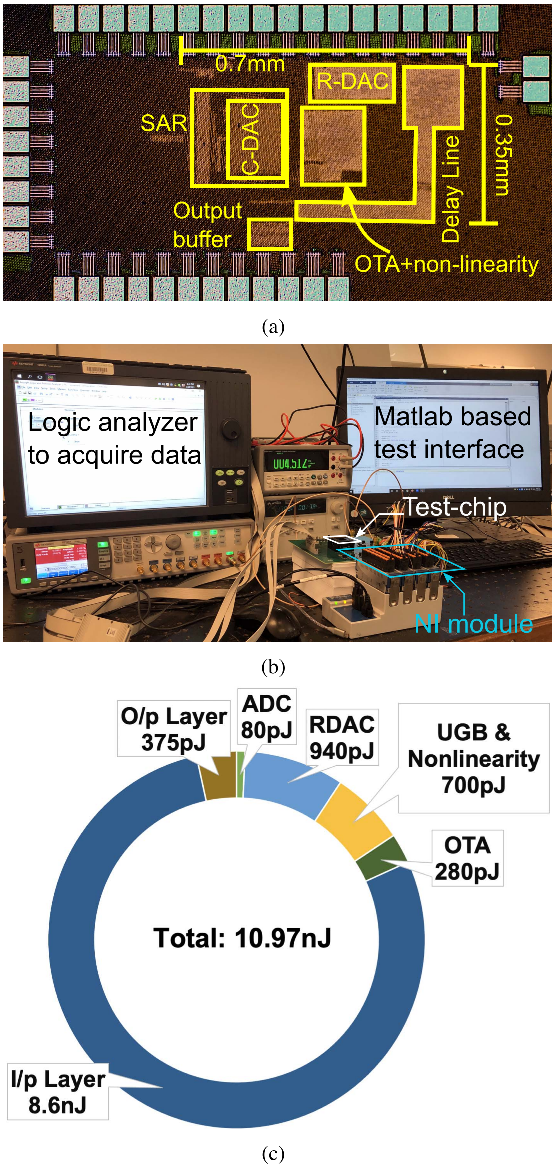
(b)

Fig. 11. Simulated (a) accuracy versus temperature (b) reservoir nonlinearity versus temperature.

beyond 0.4V, $H(\cdot)$ is set by transconductance of the amplifier, and starts changing with temperature. At high temperatures, $H(\cdot)$ output drops which reduces differences between the classes and reduces classification accuracy. While the RC-NN shows strong sensitivity to temperature, the target application is at-home monitoring of patients by embedding the RC-NN into wearable sensor, and temperature is not expected to vary significantly and the classification accuracy is not expected to degrade over narrow range of temperature as shown in Fig. 11(a). The nonlinear activation function $H(\cdot)$ is also susceptible to shifts due to offset in the amplifiers and in the ADC. However, static offset can be corrected using one-point calibration. During calibration, the digital output layer is re-trained with a small subset of known dataset. The weights of the output layer are re-tuned to maximize accuracy on the training set during calibration which corrects for static shifts in $H(\cdot)$. This one-time calibration is not computationally expensive since this is performed once and the output layer is a simple, one-layer, logistic regression model.

D. Measurement Results

Fig. 12(a) shows die microphotograph of the on-chip RC-NN. The RC-NN has a core area of 0.24mm^2. The output layer is synthesized off-chip. The test chip operates from 1.2V power supply at a speed of $F_s = 40$kHz, while the time-multiplexed reservoir layer runs at $NF_s = 400$kHz.



(a)

(b)

(c)

Fig. 12. (a) RC-NN chip micro-photograph (b) measurement setup (c) energy breakdown.

Fig. 12(b) shows the laboratory measurement setup used to characterize the test chips. The input data is loaded into the test chip through National Instruments Data Acquisition module, and the test-chip outputs are captured using logic analyzer. Data input to test-chip and output acquisition is synchronized through Matlab interface running on a desktop computer. Fig. 12(c) shows the breakdown in energy consumed/inference by the different circuit components. The input layer is estimated to consume 8.6nJ/inference with the switched-capacitor integrator having unity gain bandwidth of 80kHz. The R-DAC is the next highest consumer with 940pJ, while the ADC consumes 80pJ. The synthesized output layer has an estimated energy consumption of 375pJ.

Fig. 13 shows the measured confusion matrix. The classification accuracy is calculated on the test set. The test-chip has accuracy of 93.9%, sensitivity of 0.95 and specificity of 0.93. Fig. 14(a) shows measured accuracies of

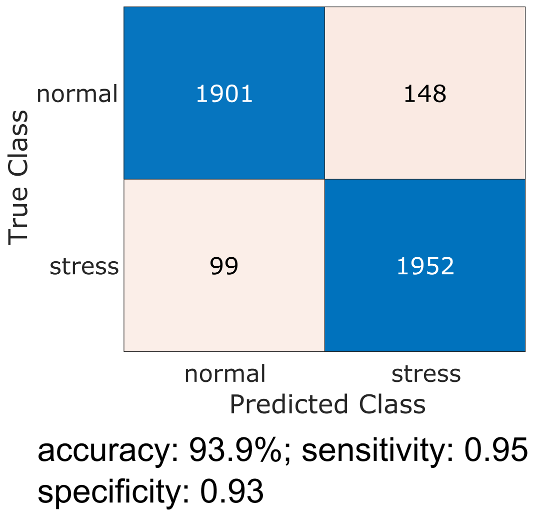
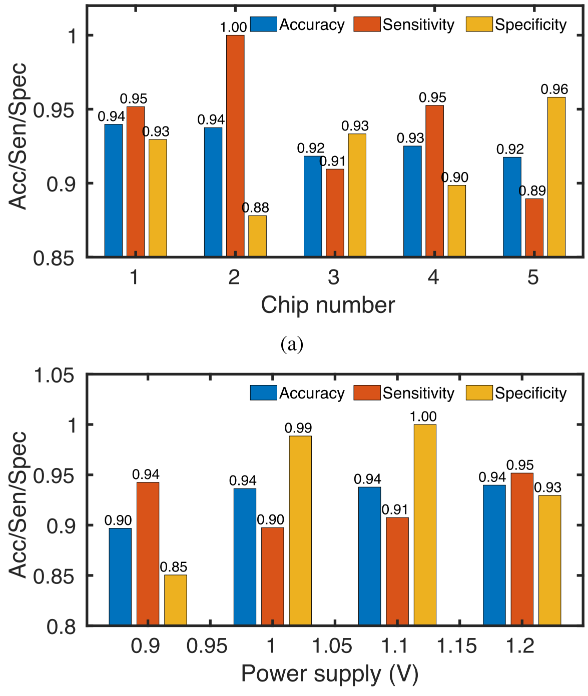


Fig. 13. Measured confusion matrix.



(a)

(b)

Fig. 14. Measured accuracy for (a) multiple chips (b) versus power supply voltage.

5 test-chips. The output layer is trained for each test-chip. The average accuracy, sensitivity and specificity are 92.8%, 0.94 and 0.92 respectively. Fig. 14(b) shows the measured accuracy for chip 1 with variation in supply voltage. The accuracy remains unchanged from 1.2V down to 1V and drops slightly at 0.9V supply voltage.

Fig. 15 shows the measured histogram of classification for 500 evaluations with chip 1 to capture the effect of thermal noise. The mean classification accuracy is 93.8% with a low standard deviation of 0.9% which has good agreement with simulation result (Fig. 10) and shows that the RC-NN is relatively robust against thermal noise.

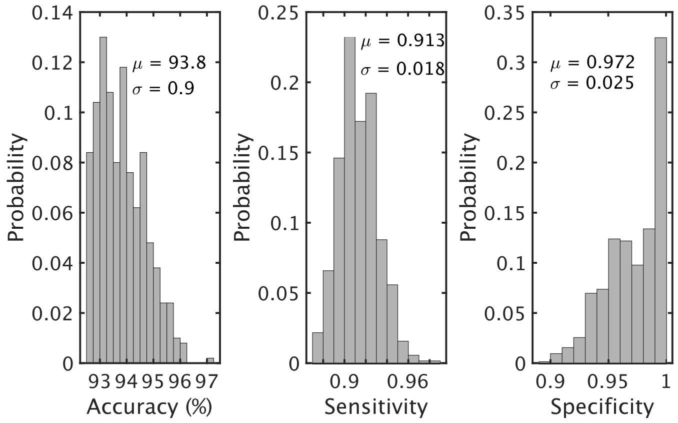


Fig. 15. Measured histogram for repeated evaluations.

TABLE I
COMPARISON WITH STRESS DETECTION WORKS

	Model	Classes	Accuracy(%)
[20]	LDA[1]	2	85.4
[25]	fusion	2	92.0
[26]	Naive bayes	3	85.7
[27]	k-NN[2]	2	87
This work	RC-NN	2	92.8[3]

[1] LDA: linear discriminant; [2] k-NN: k nearest neighbor; [3] average of 5 chips

Table I compares this work with state-of-the-art stress detection works. To the best of our knowledge, no custom ASIC has been reported in the literature that performs stress detection using WESAD dataset. To compare performance of the proposed mixed-signal RC-NN with digital ICs, we synthesized the reservoir layer digitally in the same 65nm process. The synthesized reservoir layer consumes 68nJ which is 39× higher than the energy consumption of our mixed-signal reservoir layer.

III. Heart Disease Detection From CHD Dataset

A. Fusion Model

The CHD dataset has 13 attributes of 297 patients. The 13 attributes are summarized in Table II. The physiological attributes are derived from patient vital signs, and can be recorded at-home with wearable sensors, while the laboratory result attributes requires the patient to be in a clinical setting for recording. For this work, we have used only the demographics and physiological attributes so that our design can be used for at-home monitoring of patients. D is set to 8 for this dataset corresponding to the 8 physiological attributes.

We designed a late fusion prediction model to comprehensively integrate patient demographics and physiological information as shown in Fig. 16. The fusion model combines prediction scores from two separate classifiers operating only on demographics and physiological data, and uses a meta-classifier to provide the final prediction. The proposed RC-NN is used as the classifier for physiological data, while a logistic regression model is used for predicting heart disease from demographics data. In an application scenario, the prediction

TABLE II
ATTRIBUTES IN CHD DATASET

Demographics
age, gender
Physiological data
blood pressure, blood sugar, ECG, maximum heart rate,
ST depression induced by exercise, exercise induced angina,
slope of ST segment during exercise, chest pain type
Laboratory test results
serum cholesterol, number of major vessels colored by
flourosopy, heart status from thalium test

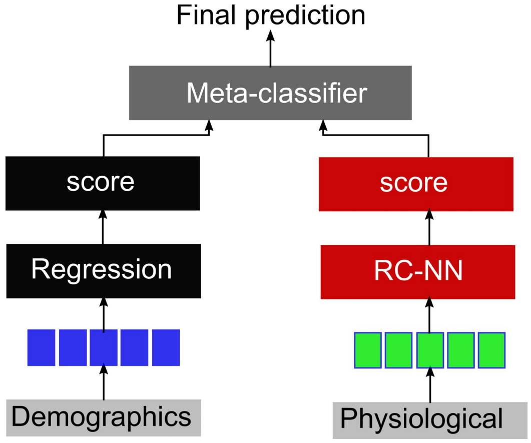


Fig. 16. Late fusion model used for heart disease prediction.

scores from the RC-NN will be sent to a mobile device which will also allow the user to upload their demographics information, and the mobile device will use cloud computing resources for implementing the meta-classifier.

B. Measurement Results

We used the same RC-NN classifier as used for detecting stress, and measured 5 test-chips. Table III summarizes the measurement performance. The classification with only the demographic attributes is 52.9%, while the RC-NN classifier, with physiological attributes as inputs, has classification accuracies of 76.9%–78.3% across 5 test-chips. We used 3 different meta-classifiers – logistic regression, linear support vector machine (SVM) and neural network for performing fusion of prediction scores from the RC-NN and logistic regression models on physiological and demographic data respectively. The neural network meta-classifier is a two-layer neural network with 20 neurons in the hidden layer, and tanh activation function. The best accuracy, sensitivity and specificity for each classifier and for each chip are highlighted in Table III.

Fig. 17 shows the breakdown in energy consumption/ inference for CHD dataset. The input layer is estimated to consume 200pJ out of total energy consumption of 2.57nJ/inference. Fig. 18 shows the measured accuracy of RC-NN and late fusion model for chip 1 as a function of supply voltage. The RC-NN accuracy reduces by 0.7% going from 1.2V supply voltage to 0.9V. SVM meta-classifier has the highest accuracy at 1.2V supply voltage while the logistic regression meta-classifier has the highest accuracy at 0.9V supply voltage. Fig. 19 shows the measured histogram of classification accuracies with RC-NN and fusion model for

TABLE III
ATTRIBUTES IN CHD DATASET

Demographics			
Model	**Accuracy (%)**	**Sensitivity**	**Specificity**
Logistic regression	52.9%	0.52	0.53
Physiological data			
Model	**Accuracy (%)**	**Sensitivity**	**Specificity**
Chip 1			
RC-NN	77.3	0.73	0.81
Chip 2			
RC-NN	78.3	0.74	0.82
Chip 3			
RC-NN	76.9	0.71	0.82
Chip 4			
RC-NN	77.3	0.72	0.82
Chip 5			
RC-NN	76.9	0.73	0.80
Fusion using meta-classifier			
Model	**Accuracy (%)**	**Sensitivity**	**Specificity**
Chip 1			
Logistic regression	87.2	**0.83**	0.89
Linear SVM	**87.2**	0.82	**0.89**
Neural network	86.1	0.80	0.89
Chip 2			
Logistic regression	87.5	0.83	0.89
Linear SVM	**87.5**	**0.83**	**0.89**
Neural network	86.5	0.82	0.89
Chip 3			
Logistic regression	87.0	**0.82**	0.89
Linear SVM	**87.1**	0.81	**0.90**
Neural network	85.8	0.81	0.88
Chip 4			
Logistic regression	87.1	0.82	0.89
Linear SVM	**87.1**	**0.82**	**0.89**
Neural network	86.4	0.82	0.88
Chip 5			
Logistic regression	87.0	0.82	0.89
Linear SVM	**87.1**	**0.82**	**0.89**
Neural network	86.2	0.81	0.89

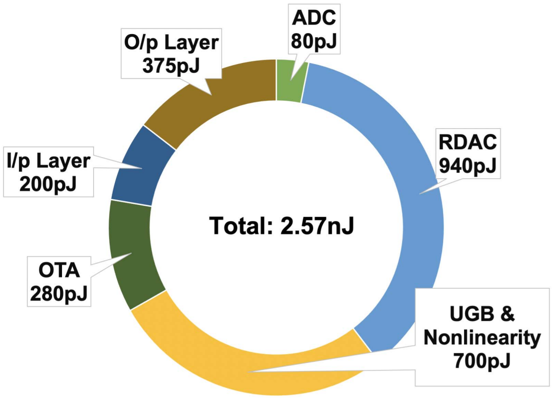


Fig. 17. Energy breakdown of RCNN for CHD dataset.

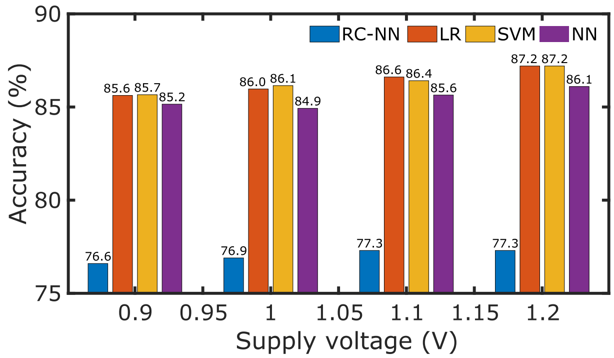


Fig. 18. Measured accuracy of RC-NN and late fusion model versus supply voltage.

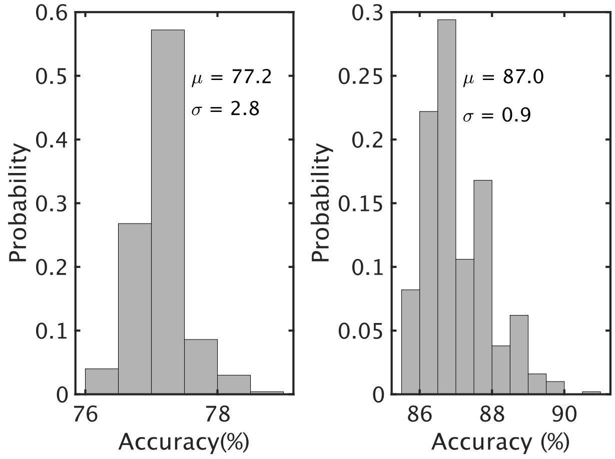


Fig. 19. Histogram of accuracy of RC-NN and fusion model.

TABLE IV
COMPARISON WITH HEART DISEASES DETECTION WORKS

	[28]	[29]	[30]	[31]	This work
Accuracy	81%	75-84%	84%	78%	**86.8%**[1]
Model	k-NN[2]	Ensemble	SVM+MLP[3]	SVM	Fusion

[1]average of 5 chips, [2]k nearest neighbor; [3]stacking of support vector machine (SVM) and multi-layer perceptron (MLP)

500 repeated evaluations with chip 1. The standard deviation of accuracy with RC-NN model is 2.8% which reduces to 0.9% after fusion with demographic attributes. The measured histogram shows the RC-NN has relatively low sensitivity to thermal noise.

Table IV compares the proposed fusion model with state-of-the-art software AI models demonstrated on CHD dataset. The proposed model achieves state-of-the-art accuracy. To the best of our knowledge, no custom ASIC has been reported in the literature that performs heart disease detection using CHD dataset.

We have demonstrated fusion model on the CHD dataset. Since the CHD dataset has only 2 demographic attributes, an early fusion model could have been used which would have combined both demographic and physiological attributes, and applied to an on-chip classifier without loss in classification performance. However the goal of this work is to demonstrate feasibility of a more comprehensive technique that will combine other demographic attributes such as race, and co-morbidities (such as hypertension, obesity etc.) of patients with patient vitals to predict heart diseases with high accuracy and also provide personalized predictions based on demographic, co-morbidity and physiological information.

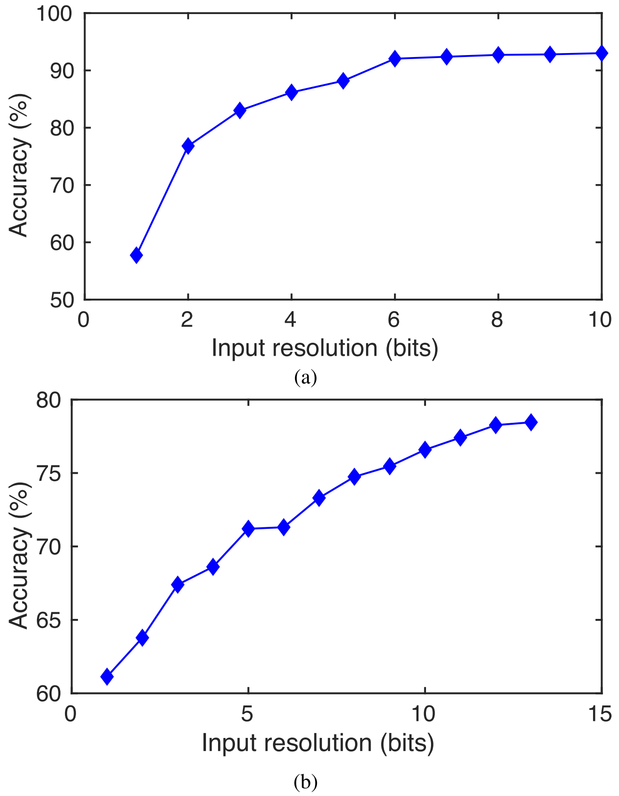


Fig. 20. Simulated accuracy vs input resolution for (a) WESAD and (b) CHD dataset.

IV. DISCUSSION

While the proposed RC-NN prototype achieves high energy-efficiency/inference compared to state-of-the-art bio-medical AI ASICs (see Fig. 1), it is difficult to compare energy efficiency of AI ASICs demonstrated on different applications and dataset. The AI ASICs can be compared at a lower level by looking at their energy efficiency for matrix multiplications which is a common computation shared across AI algorithms. Table V compares the energy-efficiency of the proposed RC-NN with state-of-the-art matrix-multiplier macros based on in-memory computing using SRAM arrays in terms of TOPS/W. ADC, DAC and nonlinearity circuit are all considered as 1 operation each for calculating energy-efficiency of the proposed RC-NN. The output layer is not included in this calculation. The energy-efficiency of our RC-NN is much lower than state-of-the-art SRAM macros. There are several reasons for this - 1) our energy-efficiency calculation includes energy consumption for data movement and activation function, while the SRAM macros report energy-efficiency only for matrix multiplication 2) the SRAM macros are designed for high throughput, and their energy-efficiency is likely to be limited by leakage at lower frequencies for bio-medical applications. Compared to the WESAD dataset, the energy

TABLE V
COMPARISON WITH STATE-OF-THE-ART AI ACCELERATOR MACROS

	[32] **JSSC'18**	**[33]** **TCAS1'19**	**[34]** **VLSI'18**	**[35]** **JSSC'20**	**[36]** **ISSCC'19**	**[37]** **JSSC'18**	**[38]** **JSSC'20**	**This** **work**	
Computation type	SRAM							Analog	
	10T	6T	10T1C	12T	8T	6T	8T1C		
Process (nm)	65	65	65	65	55	65	65	65	
Weight precision	1	1	1	1	2	8	1	1	
Input precision	6	1	1	1	1	8	1	7 (WESAD)	12 (CHD)
Efficiency (TOPS/W)	40.3[1]	30.5[1]	658[1]	403[1]	18.4[1]	6.25[1]	671.5[1]	0.66[2](WESAD)	0.08[2](CHD)
Norm. Efficiency[3](TOPS/W)	241.8	30.5	658	403	36.8	400	671.5	4.6 (WESAD)	0.96 (CHD)
Throughput (GOPS)	8	1112.8	589.9	665	269.6	8.26	1638	0.28 (WESAD)	0.007 (CHD)

[1] one MAC is considered as 2 operations (multiplication and addition)
[2] excludes output layer; nonlinearity, ADC and DAC are considered as 1 operation each
[3] normalized energy efficiency is given by energy efficiency × input precision × weight precision

efficiency is lower for CHD dataset since the RC-NN is not optimized for the CHD dataset. Input resolution for the proposed RC-NN for the WESAD and CHD dataset are estimated by quantizing the input signal with different resolutions and observing the simulated classification accuracy. The simulation results are plotted in Fig. 20. The WESAD dataset requires 7-bit input resolution while the CHD dataset requires 12-bit input resolution for no degradation in classification accuracy compared to analog input.

V. CONCLUSION

This work has presented a time-multiplexed, mixed-signal RC-NN prototype for stress detection from ECG signal, and prediction of coronary heart diseases by performing fusion of patient demographic and physiological data. The energy consumption of the proposed RC-NN can be reduced further by using a single-stage, low gain dynamic amplifier for OTA summer as well as for the unity-gain buffer. The amplifiers will be allowed to slew to further reduce power consumption since amplifier nonlinearity will be absorbed in the nonlinearity of the reservoir layer and does not affect prediction performance as the training happens entirely in the digital output layer. The proposed RC-NN is also expected to benefit from technology scaling, and further improve energy efficiency, since the analog components do not need high precision, linearity or gain.

ACKNOWLEDGMENT

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 the Air Force Research Laboratory or the U.S. Government.

REFERENCES

[1] S. S. Virani *et al.*, "Heart disease and stroke statistics—2020 update: A report from the American Heart Association," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.

[2] A. Baciu, Y. Negussie, A. Geller, and J. N. Weinstein, "The state of health disparities in the United States," in *Communities in Action: Pathways to Health Equity*. Washington, DC, USA: National Academies Press, 2017.

[3] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE J. Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, Jul. 2013.

[4] S. M. Abubakar, M. R. Khan, W. Saadeh, and M. A. B. Altaf, "A wearable auto-patient adaptive ECG processor for shockable cardiac arrhythmia," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2018, pp. 267–268.

[5] S. Yin *et al.*, "A 1.06-μW smart ECG processor in 65-nm CMOS for real-time biometric authentication and personal cardiac monitoring," *IEEE J. Solid-State Circuits*, vol. 54, no. 8, pp. 2316–2326, May 2019.

[6] J. Liu *et al.*, "BioAIP: A reconfigurable biomedical AI processor with adaptive learning for versatile intelligent health monitoring," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 62–64.

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

[8] H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks-with an erratum note," German Nat. Res. Center Inf. Technol., Bonn, Germany, GMD Tech. Rep., 2001, p. 13, vol. 148, no. 34.

[9] A. Katumba, J. Heyvaert, B. Schneider, S. Uvin, J. Dambre, and P. Bienstman, "Low-loss photonic reservoir computing with multimode photonic integrated circuits," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Dec. 2018.

[10] C. Sugano, K. Kanno, and A. Uchida, "Reservoir computing using multiple lasers with feedback on a photonic integrated circuit," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, pp. 1–9, Jan. 2020.

[11] K. Takano *et al.*, "Compact reservoir computing with a photonic integrated circuit," *Opt. Exp.*, vol. 26, no. 22, pp. 29424–29439, 2018.

[12] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nature Commun.*, vol. 4, no. 1, pp. 1–7, Jun. 2013.

[13] K. Bai and Y. Yi, "DFR: An energy-efficient analog delay feedback reservoir computing system for brain-inspired computing," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 14, no. 4, pp. 1–22, Dec. 2018.

[14] Y. Chen, E. Yao, and A. Basu, "A 128-channel extreme learning machine-based neural decoder for brain machine interfaces," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 3, pp. 679–692, Jun. 2016.

[15] C. S. Thakur, R. Wang, T. J. Hamilton, J. Tapson, and A. V. Schaik, "A low power trainable neuromorphic integrated circuit that is tolerant to device mismatch," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 2, pp. 211–221, Feb. 2016.

[16] Y. Chen, Z. Wang, A. Patil, and A. Basu, "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, Reg. Papers*, vol. 66, no. 6, pp. 2240–2252, Jun. 2019.

[17] B. Kar, P. K. Gopalakrishnan, S. K. Bose, M. Roy, and A. Basu, "ADIC: Anomaly detection integrated circuit in 65-nm CMOS utilizing approximate computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 12, pp. 2518–2529, Dec. 2020.

[18] Y.-C. Chuang, Y.-T. Chen, H.-T. Li, and A.-Y.-A. Wu, "An arbitrarily reconfigurable extreme learning machine inference engine for robust ECG anomaly detection," *IEEE Open J. Circuits Syst.*, vol. 2, pp. 196–209, 2021.

[19] H.-T. Li, C.-Y. Chou, Y.-T. Chen, S.-H. Wang, and A.-Y. Wu, "Robust and lightweight ensemble extreme learning machine engine based on eigenspace domain for compressed learning," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 12, pp. 4699–4712, Dec. 2019.

[20] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[21] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.

[22] M. Le Berre, E. Ressayre, A. Tallet, H. Gibbs, D. Kaplan, and M. Rose, "Conjecture on the dimensions of chaotic attractors of delayed-feedback dynamical systems," *Phys. Rev. A, Gen. Phys.*, vol. 35, no. 9, p. 4020, 1987.

[23] A. Rodan and P. Tino, "Minimum complexity echo state network," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 131–144, Jan. 2011.

[24] L. Appeltant *et al.*, "Information processing using a single dynamical node as complex system," *Nature Commun.*, vol. 2, no. 1, pp. 1–6, Sep. 2011.

[25] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct*, Sep. 2016, pp. 1185–1193.

[26] N. Keshan, P. V. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ECG signals in automobile drivers," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 2661–2669.

[27] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 671–676.

[28] B. Moghaddam and G. Shakhnarovich, "Boosted dyadic kernel discriminants," in *Proc. NIPS*, 2002, pp. 745–752.

[29] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.

[30] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 204–207.

[31] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Inform.*, vol. 36, pp. 82–93, Mar. 2019.

[32] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[33] X. Si *et al.*, "A dual-split 6T SRAM-based computing-in-memory unit-macro with fully parallel product-sum operation for binarized DNN edge processors," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 11, pp. 4172–4185, Nov. 2019.

[34] 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 *Proc. IEEE Symp. VLSI Circuits*, Jun. 2018, pp. 141–142.

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

[36] X. Si *et al.*, "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

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

[38] 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 J. Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, Jul. 2020.

Sanjeev Tannirkulam Chandrasekaran (Member, IEEE) received the B.Tech. degree in electronics and instrumentation from SASTRA University, Thanjavur, India, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering with the University at Buffalo, Buffalo, NY, USA. He has held internship positions with Silicon Laboratories, Austin, TX, USA, Mythic–AI, Austin, and GE Global Research, Niskayuna, NY, USA, where he was involved in mixed-signal IC design. His research interests are geared toward developing scalable energy-efficient circuits for the IoT applications with a focus on data converters and edge-AI. He was a recipient of the Best Paper Award in the 2020 IBM AI Compute Symposium, the 2019 MWSCAS Student Participation Grant, and the 2019 CICC Student Travel Grant Award. He serves as a Reviewer for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS (TCAS—I), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS (TCAS—II), and IEEE SOLID-STATE CIRCUITS LETTERS (SSC-L).

Sumukh Prashant Bhanushali received the B.Tech. degree from the K. J. Somaiya College of Engineering and the M.S. degree from the University at Buffalo, where he is currently pursuing the Ph.D. degree in electrical engineering with the Analog/Mixed-Signal VLSI Group. He was an Intern at the Atom CPU Core Group, Intel, Austin. His current research areas include design and validation of analog/mixed-Signal circuits for machine learning, data converters, and image sensors.

Imon Banerjee received the M.Tech. degree from the National Institute of Technology, Durgapur, India, in 2011, and the Ph.D. degree from the University of Genoa, Italy, in 2016.

She has done the post-doctoral training at Stanford University. She was an Assistant Professor with the Department of Biomedical Informatics and the Department of Radiology, Emory University, and the Department of Biomedical Engineering, Emory University and Georgia Tech. She is currently the Lead AI Scientist at Mayo Clinic, AZ, USA, and an Associate Professor at the School of Computing and Augmented Intelligence, Arizona State University. Her research interests are in the areas of application of machine learning for biomedical data mining and predictive modeling. She was a recipient of the 2012 Marie Curie Fellowship in European FP7 Marie Curie Initial Training Networks.

Arindam Sanyal (Member, IEEE) received the B.E. degree from Jadavpur University, India, in 2007, the M.Tech. degree from the Indian Institute of Technology Kharagpur, Kharagpur, in 2009, and the Ph.D. degree from The University of Texas at Austin in 2016.

He was an Assistant Professor with the Department of Electrical Engineering, University at Buffalo. He is currently an Assistant Professor with the School of Electrical, Computer and Energy Engineering, Arizona State University. His research interests include analog/mixed signal design, bio-medical sensor design, analog security, and neuromorphic computing. He serves as a member for the Analog Signal Processing Technical Committee (ASP-TC) and the VLSI Systems and Applications Technical Committee (VSA-TC) within the IEEE Circuits and Systems Society. He was a recipient of the 2020 NSF CISE Research Initiation Initiative (CRII) Award, the Intel/Texas Instruments/Catalyst Foundation CICC Student Scholarship Award in 2014, and the Mamraj Agarwal Award in 2001. He also serves as an Associate Editor for *Electronics Letters* (IET).