Abstract—This work presents the first on-chip, mixed-signal echo state network (ESN) for early prediction of heart disease. The ESN comprises an input layer, a non-linear projection (NP) layer, and an output layer. Only the output layer of the ESN requires training. The input layer weights are time-invariant and drawn from a static binary random distribution. Thus, the proposed ESN has significantly lower trainable parameters compared to other non-linear neural networks used for similar prediction tasks. A 65nm prototype is validated with the Cleveland Heart Disease (CHD) dataset. The ESN achieves a mean accuracy of 84.6% over 5 test chips while consuming 7.5nJ energy/inference.

Index Terms—Machine learning, echo state network, cardiac diseases prediction, data fusion and medical wearable

I. INTRODUCTION

Each year, one-third of global deaths are caused by heart diseases. Recent advances in machine learning (ML) can automate risk prediction for heart diseases and prevent death by analyzing patient physiological signals in real-time. However, conventional ML algorithms are computationally intensive and consume significant energy, thus making their integration on resource-constrained wearables challenging. Prior works 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 [1]–[5].

Instead of optimizing conventional ML architectures, this work presents an on-chip, mixed-signal echo state network (ESN) for reducing energy consumption. An ESN nonlinearly projects the input data to high-dimensional space using a nonlinear layer, and the output is typically obtained by a linear combination of the projected states. While ESNs have been used extensively in the ML literature, hardware implementation of ESNs have been mostly on photonics platforms [6], [7], and silicon implementation is limited to digital IC [5]. This work presents the first mixed-signal, on-chip ESN. Performance of the proposed on-chip ESN is demonstrated for classification of heart diseases on 297 patient data from the Cleveland Heart Disease (CHD) dataset, which is widely used for heart disease research. The CHD dataset contains demographic information and physiological measurements, as well as labels indicating if the patient has heart diseases indicated by narrowing of the epicardial artery. An early fusion ML model [8] is used, which combines patient demographic information with physiological data to create a consolidated input feature vector.

The rest of this paper is organized as follows: Section II presents the ESN architecture, measurement results are presented in Section III, and the conclusion is brought up in Section IV.

II. ESN ARCHITECTURE

Fig. 1: ESN architecture with early fusion showing linear separability of data classes after passing through ESN.

Fig. 1 shows the proposed ESN architecture with early fusion of demographic (gender and age), and physiological measurements. The input layer weights are restricted to \{0/1\} which replaces multipliers by adders and reduces hardware cost. The simple architecture and reduced training requirements make ESN very attractive candidate for low energy wearables. The ESN accepts an input vector \( \hat{X} \) with \( D \) features, which is multiplied with an \( N \times D \) input weight \( \tilde{W} \), and passed through the non-linear projection (NP) layer with \( N \) neurons. State of the \( k \)-th neuron is expressed mathematically as shown in Fig. 1 where \( \tilde{W}_f \) is the \( N \times N \) inter-connection weight matrix for the NP layer, \( H(\cdot) \) is the non-linearity function, \( G_i \) is input scaling factor and \( G_f \) is feedback gain. The inter-connect matrix \( \tilde{W}_f \) is typically sparsely filled and provides memory to the NP layer which
allows the ESN to exhibit properties of high dimensionality with a small number of neurons [9]. Fig. 2 shows the simulated prediction accuracy for different configurations of the NP inter-connect matrix. The ESN model is simulated 100 times for each $W_i$, and the input layer weights are selected randomly each time. The best average accuracy of 86.4%, with standard deviation of 0.8%, is obtained if $W_i$ is an identity matrix. As shown in Fig. 2, increasing the number of inter-connects increases memory in the NP layer and leads to over-fitting of the model during training, which reduces prediction accuracy. The inter-connect identity matrix is realized through 1-cycle delayed feedback to each neuron.

As shown in Fig. 2, increasing the number of inter-connects increases memory in the NP layer and leads to over-fitting of the model during training, which reduces prediction accuracy. The inter-connect identity matrix is realized through 1-cycle delayed feedback to each neuron.

Fig. 2: ESN classification performance for different inter-connect architectures in the non-linear projection layer

Fig. 3 shows circuit implementation of the ESN. In order to save area, the NP layer is time-multiplexed such that only one physical neuron is used which also eliminates effect of random mismatch between neurons. Computations in the NP 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. Output of the non-linear activation is digitized using a 10-bit 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. Non-linearity due to static mismatches in the SAR DAC or feedback DAC are absorbed into overall non-linearity of the NP neuron and does not need correction.

Fig. 4a) shows the mean and standard deviation of simulated accuracy as a function of number of NP neurons and $(G_i,G_f)$. The highest mean accuracy is obtained for 30 neurons and $(G_i,G_f) = (0.6,0.1)$, which are used as design parameters for the proposed ESN. Small value of $G_f$ indicates that the NP layer requires short-term memory for predictions, which can be intuitively explained by the fact that static features are used as inputs to the ESN rather than continuous-time signals. Fig. 4b) shows the mean and standard deviation of simulated accuracy as $(G_i,G_f)$ are varied independently from 1% to 19%. The mean accuracy remains > 85% for perturbation < 10%. Fig. 5 shows simulated accuracy as a function of ADC resolution. A 10-bit ADC resolution is selected for this design for high classification accuracy. Fig. 6 shows the simulated classification accuracy as a function of ESN noise referred to the ADC input. For each value of noise standard deviation, the ESN is simulated 100 times, and Fig. 6 shows mean and standard deviation of classification accuracies. The R-DAC, OTA-summer, unity gain buffer, and non-linearity contribute 0.34mV$^2$rms noise at referred to ADC input, while the ADC has an input referred noise of 0.54mV$^2$rms, which results in simulated accuracy of 85.6% with 0.6% standard deviation.
III. MEASUREMENT RESULTS

Fig. 7(a) shows die micro-photograph of on-chip ESN components. 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 NP layer runs at $NF_s = 1.2$MHz. Fig. 7(b) shows the breakdown in energy consumed/inference by the different circuit components. The R-DAC consumes the highest energy of 2.8nJ out of the total energy consumption of 7.5nJ, while the synthesized output layer has an estimated energy consumption of 1.5nJ.

Fig. 8 shows the measured confusion matrix. The classification accuracy is calculated through 5-fold cross-validation on the CHD dataset and collating the predictions on test data from each cross-validation fold. The measured accuracy is 80% if only physiological data is used as input to the ESN. Once demographics information is combined with physiological data, the accuracy improves to 84% for early fusion model.

Fig. 9(a) shows measured accuracies of 5 test-chips. The average accuracy is 84.6% if the output LR layer is trained for each chip. The average accuracy drops slightly to 83.9% if the LR weights from chip 1 is re-used for the other chips. Fig. 9(b) shows the measured accuracy for chip 1 with variation in supply voltage. Classification accuracy reduces with supply voltage. Fig. 10 shows the measured histogram of classification for 1000 evaluations with chip 1 to capture the effect of thermal noise. The mean classification accuracy is 83.74% with a low standard deviation of 0.62% which has good agreement with simulation result (Fig. 6) and shows that the ESN is relatively robust against thermal noise.

Table I compares the proposed ESN with state-of-the-art software AI models demonstrated on CHD dataset. The proposed ESN 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. To compare performance of the proposed mixed-signal ESN with digital ICs, we synthesized the NP layer digitally in the same 65nm process. The synthesized NP layer consumes 272.2nJ which is 45× higher than the energy consumption of our mixed-signal NP layer.

IV. CONCLUSION

This work has presented a time-multiplexed, mixed-signal ESN prototype for prediction of coronary heart diseases by performing fusion of patient demographic and physiological data. The energy consumption of the proposed ESN can be
technology scaling, and further improve energy efficiency, since the analog components in the ESN do not need high precision, linearity or gain.

ACKNOWLEDGMENT
This material is based on research sponsored by 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.

REFERENCES