arrWNN: Arrhythmia-detecting Weightless Neural Network FlexIC

Velu Pillai*
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, Texas, USA
velu.pillai@my.utsa.edu

Mugdha Jadhao
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, USA
mugdhaj17@utexas.edu

Alan T. L. Bacellar
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, USA
alanbacellar@utexas.edu

Diego L. C. Dutra
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
ddutra@cos.ufrj.br

Felipe M. G. França
Instituto de Telecomunicações
Universidade do Porto
Porto, Portugal
felipe.franca@lx.it.pt

Igor D. S. Miranda*
Division of Electrical and Computer Engineering
Federal University of Recôncavo da Bahia
Cruz das Almas, Brazil
igordantas@ufrb.edu.br

Paulo C. R. Souza Neto
Division of Electrical and Computer Engineering
Federal University of Recôncavo da Bahia
Cruz das Almas, Brazil
paulo.rocha@aluno.ufrb.edu.br

Mael Lhostis
Pragmatic Semiconductor
Cambridge, UK
mlhostis@pragmaticsemi.com

Eugene B. John
Department of Electrical and Computer Engineering
University of Texas at San Antonio
San Antonio, Texas, USA
eugene.John@utsa.edu

Emre Ozer
Pragmatic Semiconductor
Cambridge, UK
eozer@pragmaticsemi.com

Tejas Musale
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, USA
f20190409@goa.bits-pilani.ac.in

Zachary Susskind
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, USA
zsusskind@utexas.edu

Priscila M. V. Lima
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
priscilamvl@cos.ufrj.br

Mauricio Breternitz Jr.
ISTAR Laboratory
ISCTE - Instituto Universitário de Lisboa
Lisbon, Portugal
mbjrz@iscte-iul.pt

Lizy K. John
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, USA
ljohn@ece.utexas.edu

***Abstract*—This paper proposes a technique for incorporating machine learning into a wearable medical patch by combining two key technologies: weightless neural networks (WNNs), known for their efficiency and low hardware overhead, and Flexible Integrated Circuits (FlexICs) - ultra low-cost circuits on flexible substrates. We develop a special WNN model called "arrWNN" for detecting arrhythmia events from ECG signals that has an average prediction accuracy of 89% over the MIT BIH Arrhythmia datasets. We, then, design and implement the arrWNN model in hardware, and fabricate it using Pragmatic's FlexIC technology. The arrWNN FlexIC contains 5,706 NAND2-equivalent gates with a core area of 24 mm² consuming less than 10 mW at 3V. Our wafer-level test and measurement results show the full functionality of the fabricated arrWNN FlexICs validated against the simulation.**

Keywords—machine learning, arrhythmia, FlexIC, weightless neural networks and ECG patch

I. Introduction

Over the course of the last two decades, flexible electronics has evolved into a mature platform, delivering low-cost, slim, flexible, and conformable devices. The emergence of ultra-cost FlexIC technology [15][16] from *Pragmatic* has opened doors to a plethora of potential applications within the medical field, impacting various aspects of healthcare delivery and individual well-being. This paper investigates the application of FlexIC technology in arrhythmia detection in cardiology demonstrating the potential of implementing area-efficient neural networks (NNs) as flexible chips for real-time detection of irregular heart rhythm conditions.

Deep neural networks (DNNs) are compute and memory intensive algorithms difficult to be implemented in hardware for resource-constrained applications such as smart-packaging and healthcare patches. Weightless neural networks (WNNs) [1][2][3][4][5] offer a key advantage in terms of their hardware efficiency because they require fewer computational resources and no memory for weight storage. Thus, this makes WNNs an excellent candidate for the FlexIC technology that currently has some technology constraints such as larger device geometries, lack of large on-chip memories and the limited scale of integration. However, it has certain qualities that cannot be matched by the silicon chips such as low-cost materials, low-capex and low-carbon footprint fabrication, thinness and physical flexibility. In particular, the physical flexibility is a desirable feature for wearable devices such as ECG patches.

In this paper, we propose a WNN specifically designed to detect arrhythmia from ECG data (arrWNN). arrWNN is then implemented as an application specific integrated circuit and

* These authors have equal contribution.

fabricated as a FlexIC, which makes it the first arrhythmia detection NN hardware fabricated as a FlexIC.

II. Background and Related Work

Cardiovascular diseases, with a significant proportion attributed to arrhythmias, continue to pose a substantial global health burden. According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death worldwide, accounting for 32% of all global deaths in 2019. Arrhythmias, characterized by abnormal heart rhythms, contribute significantly to this alarming statistic. Timely identification of arrhythmias allows for the implementation of appropriate interventions, reducing the risk of morbidity, hospitalization, and mortality. Continuous monitoring over an extended period allows for the identification of intermittent or asymptomatic arrhythmias that may go undetected during short-term assessments.

There have been prior works in detecting arrhythmia using NNs, however, these NNs are computationally intensive and demand large on-chip memory. Some were implemented on FPGAs [6][7][8] and some as ASICs [9][10]. However, these NN models are too complex to fabricate in the FlexIC technology. Recent advances in edge inferencing using WNN [2][3][4][5] yield an area-efficient methodology to create lightweight neural networks that can be fabricated as FlexIC. There has been a prior work by *Ozer et al* [17] that developed a very small binary NN as a FlexIC targeting an odour recognition application rather than arrhythmia detection. *Ozer et al* [19] also develop a FlexIC detecting atrial fibrillation (i.e., a type of arrhythmia) events but the atrial fibrillation detecting FlexIC was not based on a machine learning model.

WNNs rely on value lookups implemented using RAMs or look up tables (LUTs) instead of Multiply-Accumulate operations [1][11], and have shown success for edge inferencing [2][3][4][5]. LUTs can capture a variety of non-linear functions, and NNs built with them can learn patterns with few parameters. *WiSARD* [1], an early WNN, has been shown in [12] to have a large VC dimension best suited to classifier tasks, where inputs are partitioned into different categories. It uses a sub-model called discriminator for each class created for each output category, and these discriminators are composed of small RAM nodes [1][3]. During inference, the outputs of the RAM nodes in each discriminator are summed, and the index of the discriminator with the strongest response is the predicted output. Recent research [3][5][13] have demonstrated that WNNs and their variations are effective for energy-efficient edge applications.

III. arrWNN Design

A. Model Development

We use the MIT-BIH Arrhythmia Database [18] for developing the arrWNN model and evaluating the performance of its hardware implementation fabricated as a FlexIC. The database comprises 48 half-hourly excerpts extracted from 2-channel ECG recordings obtained from 47 subjects from 23 recordings representing routine clinical scenarios alongside 25 recordings featuring arrhythmias. The data are sampled at a rate of 360 Hz per channel with an 11-bit resolution in a 10-mV range. The database labels each heartbeat as either normal or one of 18 arrhythmia types, but we reduce the number of classes into two: Normal versus Arrhythmia (all types of arrhythmia considered as a single class) to simplify the arrWNN model and its hardware implementation.

The arrhythmia detection pipeline is shown in **Fig. 1a**. The ECG signal is pre-processed by an analog frontend and then converted into a digital signal using an analog-to-digital converter. Next, a bandpass filter similar to the one used in the *Pan-Tompkins* QRS detection algorithm [14] to filter the noise. Filtered digital data is then encoded into a unary (thermometer) format and mapped into 8-bit addresses to access the WNN model where arrhythmia is predicted.

In developing the arrWNN model, we build upon the methodologies presented in *LogicWiSARD* [2] and *COIN* [4]. These architectures convert LUTs to equivalent function minterms, so no memory storage for weights is required. To reduce the number of pins used in the arrWNN FlexIC, we feed inputs to the FlexIC in a serial manner. Hence, inside the FlexIC, we convert the serial input bit stream into an 8-bit parallel format. Then, the logic function minterms corresponding to the model using the *COIN* training methodology are implemented. The model uses 741 minterms. During inference, corresponding minterm groups are selected using a multiplexer for each input tuple, and up-down counters are employed to adjust the score of each class. The argMax determines the predicted class with the highest score.

We conduct 50 iterations of Monte Carlo cross-validation to test the robustness of our WNN model. We select the best model whose AUC (Area Under the Curve) is closest to the mean AUC of all iterations. Also, we adopt a "leave-some-patients-out" strategy and stratify cross-validation to mitigate biases arising from patient data overlap and unfair partitioning. Despite the inherent imbalance in the dataset, we maintain a representative distribution of classes. The Monte Carlo cross-validation results show a mean accuracy of 0.8604

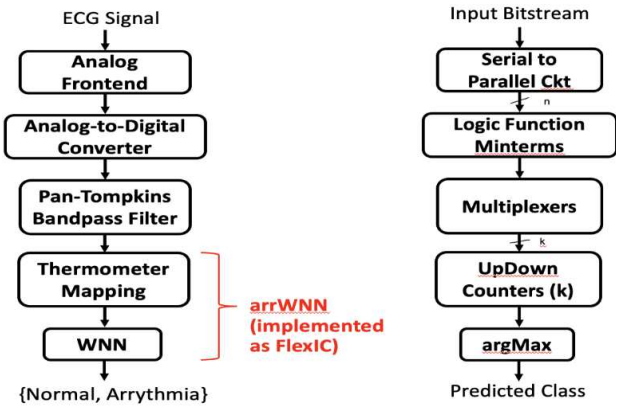
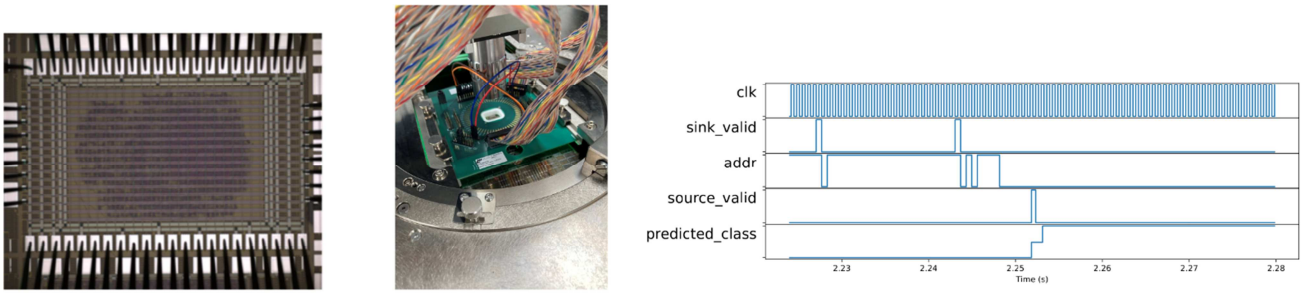


Fig. 1 (a) Arrhythmia detection pipeline showing the arrWNN implemented as an FlexIC (b) Microarchitecture of arrWNN

with a variance of 0.0038, and a mean AUC of 0.8429. Thus, the arrWNN model has an overall accuracy of 0.8827 with a sensitivity of 0.6862 and a specificity of 0.9983 achieving AUC of 0.8422.

B. Hardware Design and Implementation

arrWNN leverages *Pragmatic*'s FlexIC technology of 0.6µm n-type metal-oxide thin-film transistor technology that uses indium-gallium-zinc-oxide (IGZO) and resistors to fabricate flexible chips on a 200 mm polyimide wafer.



(a) (b) (c)

Fig. 2 (a) Die photograph of the arrWNN FlexIC of a 9x6mm pad layout with 60 pins, (b) Flexible wafer containing arrWNN FlexICs tested on a wafer probe station, (c) Waveform of the tested arrWNN FlexIC, displaying the synchronization of input and output signals during a test sequence showing the final stages of the input data stream (*sink_valid* and *addr*) and the inference outputs *(source_valid* and *predicted_class*).

The pre-trained arrWNN model is converted to Verilog. **Fig. 1b** illustrates the major blocks in the microarchitecture of arrWNN such as serial to parallel converter, function min-terms, up-down counters and the argMax to determine the counter with the highest count (i.e., the predicted class). Logic synthesis converts the design in Verilog into a gate-level netlist, and optimizations are applied to reduce die area while ensuring logical correctness. *Pragmatic*'s 0.6µm FlexIC PDK and the standard cell library are used for developing the arrWNN FlexIC. Complex logic cells such as OR-AND-Invert and AND-OR-Invert help reduce area and timing critical paths. The design is synthesised and timed for a clock frequency of 100 kHz. Timing analysis tools are used to ensure that the gate-level netlist meets these constraints. Formal verification and simulation are employed to detect and correct design errors. The logic synthesis is followed by Place and Route, when the physical placement of logic gates and the routing of interconnections is determined. Effective floorplanning, assigning logic gates to predefined locations on the chip, is followed by optimization algorithms to determine the optimal positions for these gates, ensuring that wires are as short as possible to meet timing constraints and minimize signal delay. A well-designed clock tree is used to minimize clock skew and optimize clock signal delivery.

A well-designed power grid reduces voltage drop and supports low-power operation. During Route, ensuring a reliable and efficient power grid is crucial. In the initial stages of Place and Route, meticulous power planning is undertaken to define the layout of the power grid. This process encompasses the strategic positioning of decap cells, which are miniature capacitors strategically placed to alleviate voltage fluctuations and guarantee a stable power supply. Cell routing density determines the signal routing congestion and final area. A balanced and calculated approach is applied to avoid timing violations. Design Rule Checking (DRC) and Layout-versus-Schematic (LVS) are performed to verify that the physical layout matches the original design. Then, the physical layout is converted to GDSII format for tape-out and fabrication. The physical implementation results show that the arrWNN FlexIC has a maximum clock frequency of 100 kHz, a NAND2-equivalent gatecount of 5,706 with a core area (excluding the chip-level power grid, pads and IOs) of 24 mm^2 and consume 9.4 mW power at 3V.

C. FlexIC Fabrication and Test

Fig. 2a shows the die photo of the arrWNN FlexIC fabricated on a 200 mm polyimide wafer. Two flexible wafers are fabricated and all arrWNN FlexICs in the wafers undergo rigorous testing to ensure functionality. A probe card is used to test the arrWNN FlexICs on the flexible wafer as shown in **Fig. 2b**. Due to high capacitive loading in the test equipment, the maximum clock frequency generated by the test harness is 6.25 kHz, so the arrWNN FlexICs are operated at a clock frequency of 6.25 kHz.

The tests involve 306 inference test vectors created from the ECG datasets in the MIT-BIH Arrhythmia Database. The test results exactly match the simulation results indicating the full functionality. **Fig. 2c** visualizes these results demonstrating accurate synchronization between input and output signals throughout the test sequence. This includes the final stages of the input data stream (*sink_valid* and *addr*) and their corresponding outputs (*source_valid* and *predicted_class*). The slow rise time observed on the *predicted_class* signal is attributed to the combined effect of capacitive loading from the test equipment cable and the limited drive strength of the output buffers on the chip.

IV. Conclusion

In this paper, we have presented "arrWNN" - an innovative machine learning network designed for arrhythmia detection, harnessing the computational efficiency of weightless neural networks and the ultra-cost and physical flexibility of the FlexIC technology. We have developed the arrWNN model, designed and implemented its hardware, fabricated it as a FlexIC. The mean accuracy of the arrWNN model has been around 89%. When physically implemented as a FlexIC, arrWNN has a core area of 24 mm^2, which is equivalent to a NAND2 gatecount of 5,706 consuming less than 10 mW at 3V. Our test and measurement results have shown the full functionality of arrWNN operating at a clock frequency of 6.25 kHz. arrWNN will enable a disruptive class of arrhythmia detection hardware for emerging low-cost wearable ECG patches.

Acknowledgment

This research was partially supported by Semiconductor Research Corporation (SRC) Task 3148.001, National Science Foundation (NSF) Grants #2326894 and #2425655, NVIDIA Applied Research Accelerator Program Grant, and by FCT/MCTES through national funds and, when applicable, co-funded by EU funds under the projects UIDB 50008/2020, UIDB/04466/2020, UIDP/04466/2020 and by Next Generation EU through PRR Project Route 25 (C645463824-00000063). Any opinions, findings, conclusions, or recommendations are those of the authors and not of the funding agencies.

REFERENCES

[1] I. Aleksander, W.V. Thomas, and P.A. Bowden. “WISARD·a radical step forward in image recognition,” Sensor Review, vol. 4, no. 3, pp. 120–124, Mar. 1984.

[2] I. D. S. Miranda et al. “LogicWiSARD: Memoryless Synthesis of Weightless Neural Networks,” IEEE 33rd International Conference on Application-specific Systems, Architectures and Processors (ASAP), pp. 19–26, July 2022.

[3] Z. Susskind et al. “Weightless Neural Networks for Efficient Edge Inference,” 31st International Conference on Parallel Architectures and Compilation Techniques (PACT), pp. 279-290, Oct. 2022.

[4] I. D. S. Miranda et al. “COIN: Combinational Intelligent Networks,” IEEE 34th International Conference on Application-specific Systems, Architectures and Processors (ASAP), pp. 27–28, July 2023.

[5] Z. Susskind et al. “ULEEN: A Novel Architecture for Ultra Low-Energy Edge Neural Networks,” ACM Trans. Archit. Code Optim. vol. 20, no.4, pp. 1-24, Dec. 2023.

[6] D. L. T. Wong, Y. Li, D. John, W. K. Ho and C. -H. Heng “An Energy Efficient ECG Ventricular Ectopic Beat Classifier Using Binarized CNN for Edge AI Devices,” IEEE Transactions on Biomedical Circuits and Systems, vol. 16, no. 2, pp. 222– 232, Apr. 2022.

[7] J. Lu et al. “Efficient hardware architecture of convolutional neural network for ECG classification in wearable healthcare device,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 7, pp. 2976–2985, July 2021.

[8] M. Wess, P. D. S. Manoj, and A. Jantsch. “Neural network based ECG anomaly detection on FPGA and trade-off analysis,” IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–4, May 2017.

[9] M. Janveja, R. Parmar, M. Tantuway and G. Trivedi, “A DNN-based low power ECG co-processor architecture to classify cardiac arrhythmia for wearable devices,” IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 4, pp. 2281–2285, Apr. 2022.

[10] S. -Y. Hsu, Y. Ho, P. -Y. Chang, C. Su and C. -Y. Lee, “A 48.6-to-105.2 µW Machine Learning Assisted Cardiac Sensor SoC for Mobile Healthcare Applications,” IEEE Journal of Solid-State Circuits, vol. 49, no. 4, pp. 801–811, Apr. 2014.

[11] I. Aleksander, M. De Gregorio, F.M.G. França, P.M.V. Lima and H. Morton, “A brief introduction to Weightless Neural Systems,” 17th European Symposium on Artificial Neural Networks (ESANN), pp. 299–305, Apr. 2009.

[12] H. C. C. Carneiro, C. E. Pedreira, F. M. G. França and P. M. V. Lima, "The Exact VC Dimension of the WiSARD n-Tuple Classifier," Neural Computation, vol. 31, no. 1, pp. 176-207, Jan. 2019.

[13] L. Santiago et al. “Weightless neural networks as memory segmented bloom filters,” Neurocomputing, vol. 416, pp. 292–304, Nov. 2020.

[14] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 3, pp. 230-236, Mar. 1985.

[15] E. Ozer et al., “A hardwired machine learning processing engine fabricated with submicron metal-oxide thin-film transistors on a flexible substrate,” Nat Electron, vol. 3, 419–425, July 2020.

[16] J. Biggs et al., “A natively flexible 32-bit Arm microprocessor,” Nature, vol. 595, pp. 532-536, July 2021.

[17] E. Ozer et al., “Binary Neural Network as a Flexible Integrated Circuit for Odour Classification,” IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS), pp. 1-4, Oct. 2020.

[18] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 3, pp. 45-50, May-June 2001.

[19] E. Ozer et al., “A Custom-designed Atrial Fibrillation Detection hardware on a Flexible Substrate,” IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS), pp. 1-4, July 2024.