An Optimized Hardware Inference of SABiNN: Shift-Accumulate Binarized Neural Network for Sleep Apnea Detection

Omiya Hassan, Tanmoy Paul, Nazmul Amin, Twisha Titirsha, Rushil Thakker, Dilruba Parvin, *Member, IEEE*, Abu Saleh Mohammad Mosa, and Syed Kamrul Islam, *Senior Member, IEEE*

***Abstract*— This article presents the design of an optimized hardware-based neural network (NN) called a shift-accumulate binarized NN (SABiNN). SABiNN is used in detecting respiratory-related diseases such as sleep apnea (SA) among adults. Initially, a three-hidden-layer-based NN model was trained, validated, and tested with open-source apnea polysomnography (PSG) datasets collected from the PhysioNET databank. Single-lead electrocardiography (ECG) and pulse oximeter data were collected, preprocessed, and digitized for network training. The NN was later transformed into SABiNN, demonstrating model accuracy of 81.5% (CI: ±3.5) with an energy efficiency of 5 mJ on reprogrammable hardware. The precision rate of the model was further increased by redesigning the XNOR gate of the multiply–accumulate (MAC) operation with the NAND gate-based XNOR. This redesign process significantly improved the overall model's classification and precision. Further expansion of SABiNN was carried out to achieve a higher accuracy rate (over 88%) which was designed on the CMOS platform using a 130-nm open-source process design kit (PDK) developed by Google and Skywater. The proposed model on the CMOS platform used a chip area of 0.16 mm^2 and showcased an energy efficiency of 1 pJ. A total of ∼11k CMOS cells with two 16-bit input and one 1-bit output pins were used to realize the SABiNN on CMOS. The success of this design technique can be leveraged in developing a fully system-on-a-chip (SoC) integrated wearable system for SA detection.**

***Index Terms*— 130-nm process design kit (PDK), apnea, binarized neural network (BiNN), biomedical, field-programmable gate array (FPGA), Google–SkyWater, machine learning (ML), multiply–accumulate (MAC), neural network (NN), PhysioNET, sleep apnea (SA) detection, system-on-a-chip (SOC), TinyML, XNOR.**

I. Introduction

ONE-SEVENTH of the world's population is suffering from respiratory-related diseases [1], [2], of which 80% are associated with sleep apnea (SA). Despite the significant rise in the cases of SA-related episodes in the general population, it is marked as one of the most neglected diseases due to its expensive and inconvenient diagnosis and treatment methods. Although the recent outbreak of COVID-19 created mass awareness of respiratory diseases, available diagnosis and treatment methods are far from adequate. It has been statistically proven that the excruciating factor leading to significant death tolls was the lack of rapid detection tools, particularly in developing nations [3], [4], [5], [6]. In addition, the severity of lung infection was observed to be worse in patients having a history of SA [7], [8], [9]. Therefore, it is necessary to treat and diagnose respiratory-related diseases such as SA at their early stages [10]. Over the past few decades, significant advancement has been made in medical Internet of Things (IoT) devices [11] and point-of-care (PoC) systems [12]. However, to this date, the only reliable technique for screening SA is the historical cardiorespiratory polysomnography (PSG). This gold standard method still uses wires and often invasive patches and requires overnight sleep studies. Developing an accurate, portable, cost-effective, and less intrusive system is well overdue. The change and the shift from traditional PSG screening to an advanced wearable and smart device will enable a range of possibilities. Moreover, precautionary measures can be taken for prediagnosed patients to prevent them from getting infected with other infectious diseases such as COVID.

Competent healthcare is estimated to be a 10 trillion-dollar market as of 2022 [13]. The growing demands for such a healthcare system can be resolved by incorporating artificial intelligence (AI) in medical device technologies. As the reliability of deep learning (DL) models is improving, it has filled in various sides of healthcare, from monitoring [14] to prediction [15], diagnosis [16] treatment, and prognosis [17]. Significant contributions have been made in training and developing sophisticated AI/machine learning (ML) models in detecting and predicting SA events with accuracy achieving over 99% [18], [19], [20], [21], [22], [23], [24], [25], [26]. However, as the field of DL is growing in terms of performance, network size, and training run time, the development of dedicated hardware in biomedical applications needs to be improved. Previously, it has been demonstrated that the capacity of AI can meet or exceed the performance of human experts in medical diagnosis, especially for SA detection [27]. While this is possible for software-based applications or in a laboratory setting enabling cloud-computing or IoT devices [28], [29], realizing

Manuscript received 7 November 2022; revised 28 March 2023; accepted 30 April 2023. Date of publication 25 May 2023; date of current version 12 June 2023. The Associate Editor coordinating the review process was Dr. Rajarshi Gupta. *(Corresponding author: Omiya Hassan.)*

Omiya Hassan, Tanmoy Paul, Twisha Titirsha, Abu Saleh Mohammad Mosa, and Syed Kamrul Islam are with the Department of Electrical Engineering and Computer Science (EECS), University of Missouri, Columbia, MO 65211 USA (e-mail: ohbk4@mail.missouri.edu).

Nazmul Amin and Dilruba Parvin are with Texas Instruments Inc. Dallas, TX 75243 USA.

Rushil Thakker is with Caterpillar, Peoria, IL 61630 USA.

Digital Object Identifier 10.1109/TIM.2023.3279880

a portable and accurate, real-time SA detection tool on edge is challenging [27]. An exact DL network for ambulatory SA detection is computationally expensive and typically requires data centers and cloud computing, which can compromise the privacy of patient data. Recent publications focusing on sleep-related abnormalities and diagnosis have demonstrated that for models using a minimal number of sensors, efficient ML models [30], [31], [32], [33], [34], [35], [36], [37], [38], dedicated hardware, or a secured cloud-based approach can provide higher security [39], [40], [41], [42], [43], [44]. Research also shows that these algorithmic models can be designed and implemented optimally and energy-efficiently on hardware [11]. However, any realistic energy-efficient AI/ML edge hardware is yet to be developed.

In this extended paper of [45], we propose an energy-efficient neural network (NN)-based hardware architecture, which has been implemented and explored on both reconfigurable and CMOS platforms. The proposed NN-based hardware can monitor real-time respiratory-related diseases such as SA in a portable and secure manner. This design scheme aims at improving the processing limitations between the NN models and the edge devices. The notable contributions of this work are as follows.

1) Using hardware and software tools when training and optimizing an NN model.
2) Applying class balance on datasets and binarization techniques on the NN model for minimizing memory utilization.
3) Low-power hardware techniques such as the shift-accumulate (SAC) method have been adopted and improved from [46] to meet the power efficiency requirements.
4) NAND-based XNOR gates are replaced with traditional XNOR gates in SAC connection to achieve higher precision and accuracy.
5) Fully optimized NN model trained for automatic SA detection has been integrated into ASIC with a power consumption of $\sim$10 μW introducing a gateway to future ML-on-chip applications.

The rest of this article is organized as follows. Section II overviews the entire system and develops the proposed decision-making block. Section III describes the design methodology, from the preprocessing and collection of the open-source medical datasets to the training, compression, and optimization of the NN model. Section IV showcases the experimental results from the proposed SABiNN model on reprogrammable hardware, Section V gives an overview of the expanded SABiNN model on the CMOS platform using the 130-nm open-source process design kit (PDK) process, and Section VI discusses the proposed SABiNN and its novelty in contrast to current research work on SA detection using AI/ML models. Finally, concluding remarks are made in Section VII.

II. System Overview

Clinical representations of SA include irregularity in oxygen saturation levels, breathing, and heart rate. The monitoring equipment used during PSG includes electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), pulse oximeter, etc. As a result, the procedure becomes uncomfortable due to multiple patches being placed all over the head and the body of the patients. Besides, PSG is time-consuming and requires sleep experts to extract features from the sensor readings. To overcome these limitations, we propose a real-time SA detection system using two types of input: single-channel biopotential ECG data from the chest strap and oxygen saturation (SpO_2) signal from finger-tip pulse oximeters. ECG measures the electrical activity of the heart and includes multiple chest leads. Studies have shown that ECG signal is a valuable tool for early detection of SA [34], [35]. Each heartbeat consists of three primary waves: P, QRS complex, and T, representing regularity in ECG signals. The deviation and the extraction of the R-R peak intervals of these different waves from their usual pattern confirm the presence of SA. In addition, a pulse oximeter is a low-cost tool that measures the oxygen saturation level in blood using a red LED signal. It can contribute to the diagnosis of SA based on the rise and drop of the SpO_2 levels.

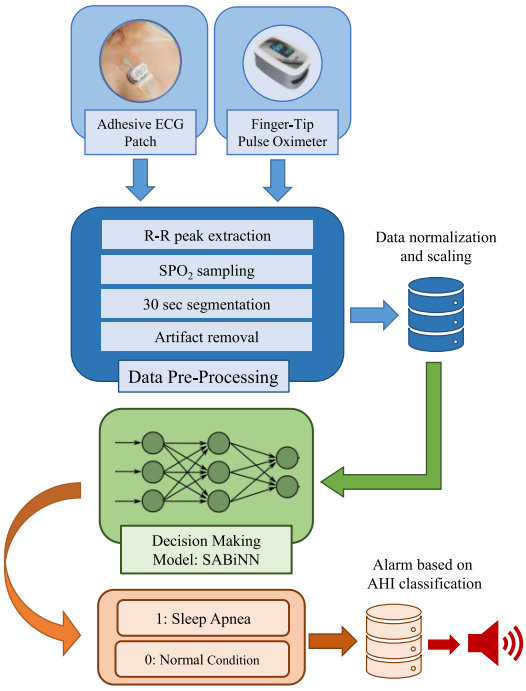

Fig. 1. Block diagram of the proposed system design which is used in screening SA events. The system uses single-lead ECG patch and finger-tip pulse oximeter as front-end sensors and predicts SA events through the decision-making model SABiNN.

Fig. 1 depicts a block diagram of the proposed SA detection system. The decision-making model (SABiNN) classifies the real-time data for potential SA events. The data include the output signals generated by the single-lead ECG patch and the finger-tip pulse oximeter. SABiNN generates a binary output (1: sleep apnea, 0: absence of apnea/normal condition) based on the apnea—hypopnea index. The binary "1" signal triggers an alarm that helps patients wake up and resume breathing and alerting caregivers to respond quickly. The proposed design of the SABiNN block can successfully predict apneic occurrences with an accuracy of 81.5% (CI: $\pm$3.5).

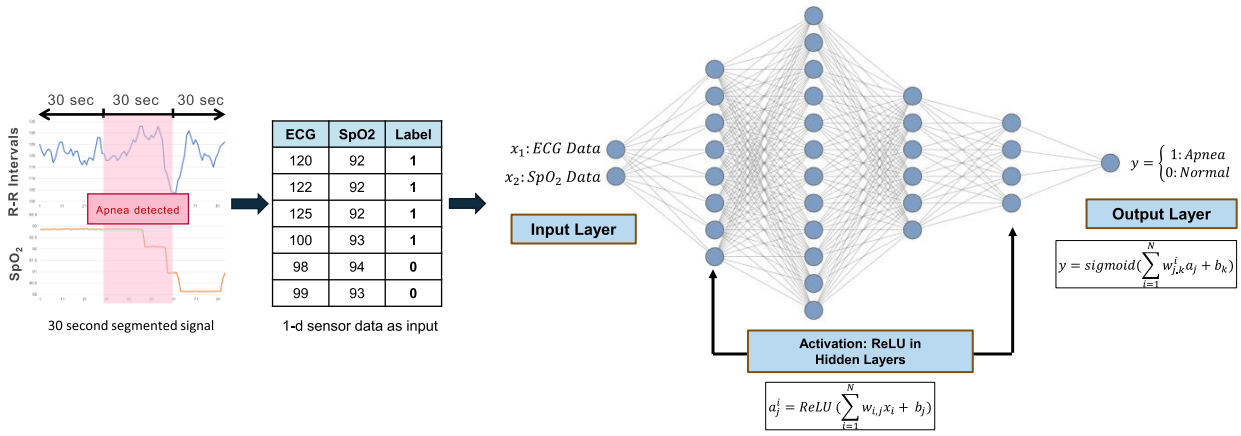

Fig. 2. Construction of a four-layer 2-(8-12-6-4)-1 FNN model with two 30-s segmented 1-D sensor inputs (x1: R-R interval and x2: SpO_2) generated from ECG patch and pulse oximeter. The hidden layer consisting of ReLU activation function and sigmoid as output layer activation function.

III. Design Methodology

A. Data Generation and Preprocessing

There are publicly available datasets that aid in benchmarking newly developed DL methods. In biomedical applications, there are widely available open-source medical datasets collected from various existing biosensors. In our research, two popular datasets that are publicly available from two different institutions were used for benchmarking the proposed SABiNN model. Both the datasets contained signals from ECG patches and pulse oximeters.

Philips University Hospital (Apnea-ECG) [48] records eight overnight ECG recordings and finger-tip SpO_2 data. An annotation was placed for every minute of the data stream, indicating whether any apneic event had occurred.

St. Vincent University Hospital [49] database contains 25 complete overnight PSG records with ECG and fingertip SpO_2 data over six months. About 25 subjects were referred to as having SA. These include subjects over 18 years of age, and the whole set contains 21 males and four females (age: range 28–68 years; AHI: 24.1 $\pm$ 20.3). Sleep technologists scored and annotated the sleep stages according to the standard Rechtschaffen and Kales rules [38]. The apnea-ECG dataset [48] was annotated after each minute by sleep experts. The ECG and the SpO_2 data from this dataset had a sampling rate of 100 Hz. Our work divided the data stream into 30-s segments instead of 1 min for a higher precision rate. On the other hand, the St. Vincent University Hospital database [49] recorded the ECG and the SpO_2 data at a sampling rate of 128 and 8 Hz, respectively. Both the datasets were further processed, and the existing artifacts were removed for a higher accuracy rate. Any SpO_2 value less than 50% and any sudden change in oxygen saturation level greater than 4% within a 1-s interval were marked as artifacts since such cases are physiologically impossible [51]. Once the artifacts were rejected, the signal was resampled at 1 Hz using a simple moving average filter. In the case of ECG signals, the dataset provided machine-generated QRS annotation. From the QRS, an R-R interval series was created by taking the time interval between two successive R-peaks. This was done using an R-peak feature extraction tool provided by PhysioNET. A sliding window technique was implemented to remove the ectopic sample points from the R-R interval series [36]. The window length was 5 s, and any R-R interval more significant than 20% of the average value within the window was marked as an ectopic beat and was removed.

B. Feedforward Neural Network (FNN)

The decision-making block has been designed based on a DL model called feedforward NN (FNN). FNN was chosen due to its straightforward calculation and minimal design features. A basic NN has multiple numbers of neurons that multiply the input data with a learnable parameter called weight and bias and sends the resultant data to the subsequent network. In our work, the proposed pretrained FNN inference module takes input data from the sensors and predicts apneic occurrences in binary format. It consists of a three-hidden layer FNN model with units 2-(8-6-4)-1 where the ReLU activation function is used in the input and the hidden layers and sigmoid function in the output layer. The model is further expanded for CMOS implementation in four hidden layers 2-(8-12-6-4)-1 while keeping the rest of the functions and loss function the same.

Fig. 2 showcases the graphical representation of the expanded FNN model where two types of datasets are fed in as input (x1 and x2), and the binary output is generated at the output layer consisting of two nodes. Dense layers were used from the Keras library in constructing the FNN model. During training, the mean-squared-error function was used for loss calculation, ADAM optimization was leveraged, and the k-fold cross-validation technique was implemented due to the limited number of sample points from the datasets. For designing the FNN architecture, the TensorFlow ML library was used on the Google Collaboratory platform for training, validating, and testing the model.

C. Class Imbalance and Normalization Technique

Following preprocessing of the datasets, a significant amount of class imbalance was present in the annotations. As a result, the proposed FNN model showed a sensitivity of $\sim$100% but an accuracy of 80%. This demonstrates that the model was biased toward detecting only the dataset's

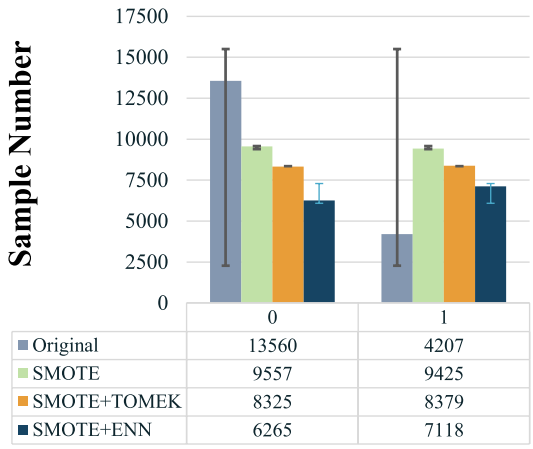

Fig. 3. Number of samples versus data labels obtained by executing various class-imbalance techniques, such as SMOTE, Tomek, ENN, and a combination of them. x-axis defines the sample number before and after class-imbalance technique was applied and y-axis defines the binary classes.

true positives (TPs). Researchers generally use various techniques to avoid class imbalance and such a training bias. These techniques include either generating minor class labels or removing the major class labels of the datasets. In this research, a combination of three popular techniques such as Synthetic Minority Oversampling TEchnique (SMOTE), Tomek, and edited nearest neighbor (ENN) was explored. The SMOTE technique generates synthetic data from the minority class, while the Tomek technique removes major class data by locating all the cross-class nearest neighbors. The ENN technique finds the observation's k-nearest ($k = 3$) neighbors and detects the difference between the major and minor classes. If different, then the observation and its k-nearest neighbor are removed. Generally, a combination of SMOTE-Tomek and SMOTE-ENN generates a balanced dataset [52], [53].

Fig. 3 illustrates the original imbalanced training dataset and the balanced dataset generated by the three techniques. The x-axis defines the number of samples before and after class balance, and the normalization technique was applied. The y-axis represents the binary classification of the dataset (1: apnea occurred, 0: normal). The chart below the graph is a numeric representation of the graph. A receiver operating characteristic (ROC) curve of the proposed FNN was executed to understand its performance over each method. Fig. 4 graphically showcases the ROC curve where the SMOTE + ENN dataset performed the best with a TP rate of ∼83%. With a balanced dataset, the proposed FNN model showed optimal accuracy (∼80%) for medical screening and diagnosis.

Further evaluation was performed based on the evaluation metrics such as precision, sensitivity, and F1-score [27]. These are essential components in validating the quality of the NN model. To validate the model accuracy further, a comparative study between the existing and ongoing research works focused on SA detection has been performed with the same dataset called ApneaECG, which is collected from PhysioNET bank [48], as shown in Table I. From Table I, it can be concluded that the proposed model has a higher accuracy rate with a balanced percentage rate of precision, sensitivity, and F1 score.

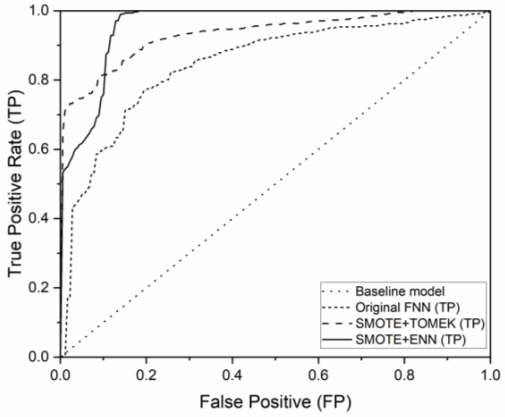

Fig. 4. ROC curve between Baseline (50%), FNN (65%), SMOTE + Tomek (73%), and SMOTE + ENN-based FNN (83%) model.

TABLE I
COMPARATIVE STUDY OF FNN MODEL WITH APNEIC DATASET [48]

Parameter	FNN	TW-MLP [30]	LS-SVM[54]	HMM-SVM[55]
Accuracy	**87%**	87%	83%	80%
Precision	**88%**	88%	84%	85%
Sensitivity	**85%**	85%	84%	85%
F1-Score	**86%**	85%	79%	72%

D. Shift-Accumulate-Based Binarized Neural Network (SABiNN)

Deploying NN models on edge is critical in designing real-time automated devices and systems. It significantly reduces the communication cost with the cloud regarding network latency and power consumption. On the contrary, edge devices have limited memory, power, and computing resources, constraining extensive computationally intensive models from embedding them into the hardware. As a result, these networks must be compact and optimized for embedded deployment. The proposed compact design technique (SABiNN) enables users to quickly deploy trained binarized NN (BiNN) models onto any edge device.

According to [56], a significant reduction is observed in the memory size when binarizing the weights between +1 and −1. This is because the arithmetic operations are replaced with bitwise functions, significantly reducing the power consumption. In the work [56], NN with binarized weights and activations was trained and benchmarked on the MNIST, CIFAR-10, and SVHN datasets, producing near state-of-the-art results. A recent work [57] introduced an XNOR gate in a parallel computation scheme where the weights and the input values were binarized. This drastically reduced the memory size and showcased high energy efficiency. However, these models were benchmarked with image datasets. Images can be easily transformed into their histogram version where the pixel values are in binary "−1/0" and "+1" forms. But physiological signals generated from biosensors cannot be translated into the binary format as a significant amount of data will be lost due to its 1-D nature. Thus, this binarization method [56], [57] results in a low accuracy rate even if it is power-efficient. Considering this, the post-training SABiNN model was developed where the binarization technique is used only on the hyperparameters

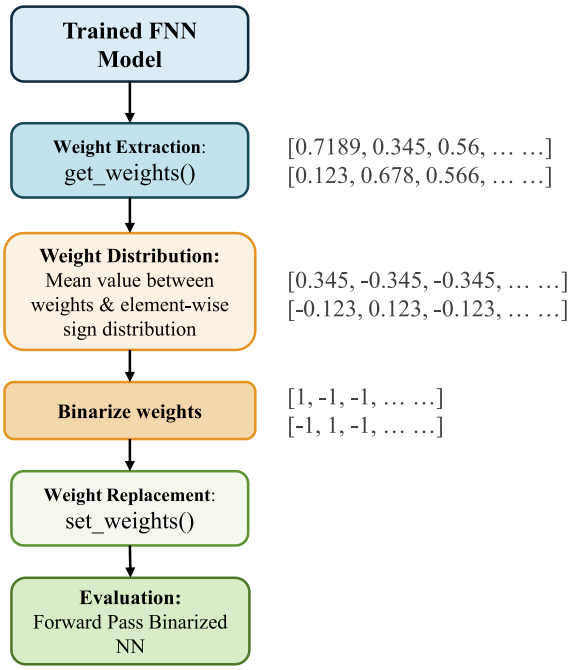

Fig. 5. Step-by-step optimization and conversion technique of the SABiNN model from FNN in software using Google Collab and Keras Library.

TABLE II
EVALUATION METRICS OF SABiNN AND FNN USING COMBINED DATASET FROM [48], [49]

Parameter	SABiNN	FNN
Accuracy	**77%**	80%
Precision	**71%**	78%
Sensitivity	**73%**	78%
F1-Score	**81%**	84%

instead of the input values [45]. A significant reduction in memory and power was observed when the data were forward passed through the proposed network. Instead of using the stochastic binarization technique during training [56], the sign function was used post-training for balanced distribution. In Fig. 5, an activity diagram showcases the algorithm of the proposed SABiNN model. The accuracy of the model is still maintained without any significant loss, as shown in Table II. The comparative study between SABiNN and FNN used the accumulated dataset from Phillips University [48] and St. Vincent [49]. The final binarized model used ReLU as its hidden layer activation function and sign function in its output layer.

Four widely used quality metrics are calculated for NN model validation to validate the performance of the SABiNN model on the SA dataset: accuracy, precision, sensitivity, and F1-score [27]. Even though the accuracy of SABiNN decreased compared with that of the FNN, it still maintained an accuracy level of over 70%, which is medically accepted and realistic. To further verify the proposed model, SABiNN was tested with the same testing dataset [48] that was used in other machine learning models that predicted SA events, followed by a metric evaluation test. The comparative study is shown in Table III.

Table III shows that the proposed model outperformed the existing models for predicting and detecting SA events. The

TABLE III
COMPARATIVE STUDY OF SABiNN MODEL WITH APNEAECG DATASET [48]

	Accuracy	Sensitivity	F1-Score
LS-SVM [54]	83%	84%	79%
HMM-SVM[55]	80%	85%	72%
Decision Fusion[58]	86%	82%	88%
TW-MLP [30]	87%	85%	85%
SABiNN	**88%**	**94%**	**91%**

TABLE IV
POWER CONSUMPTION STUDY

Function	Max Power (W)	Junction Temp (°C)	Thermal Margin (W)
Sigmoid	1.257	30.5	15.0
Sign	0.084	25.4	16.1
Tanh	2.061	34.4	14.1
ReLU	1.213	30.5	15

next step of our work was to infer the SABiNN model into hardware by making design changes to the activation function and the neuron connection of the model without sacrificing model accuracy and precision.

1) Shifter-Based Activation Function on the Edge: In NNs, the product of the weights and the input data gets classified through activation functions. Therefore, choosing such functions for different layers is essential for achieving a high accuracy rate [27]. Due to its classification characteristics, the activation function usage increases as each neuron unit uses this block. Thus, it is essential to choose an activation function where the power consumption is kept low while using limited resources from hardware. In our model, two activation functions, ReLU and sign, have been used. Other popular functions were implemented, but ReLU and sign demonstrated low power consumption rates of 1.21 and 0.084 W, respectively, as shown in Table IV. The reported consumption ranges are the maximum power the logic blocks can reach, and the junction temperature defines their maximum temperature increase when operated. ReLU is a piecewise linear function that sends the input data as output if it results in a positive value and results in zero if negative [45]. The 8-bit ReLU function was designed using an *if-else* logic unit on the hardware, represented by stacking multiple 2:1 multiplexers. The sign function was used in the output layer, which successfully categorizes the input values from the previous layer between 0 and 1. This function was also designed with a stacked 2:1 multiplexer [45].

2) 8-Bit Quantization and Post-Train Binarization: The hardware architecture was designed in n-bit ($n = 8$ for reprogrammable hardware and $n = 16$ on CMOS process) integer value. Thus, all the floating points from the input were quantized into n-bit using the quantization method [46]. According to the method, the floating-point values were first shifted into 16 bits using a 10-bit shifter. To convert it from 16-bit into 8-bit, it was back-shifted 7-bits using a 7-bit reverse shifter. This way, no significant data loss was observed, and the integrity of each input value was maintained.

3) Binarized Neuron Design: NN can be categorized into diverse networks with multiple computational schemes. Their neuron and synapse units are the most fundamental

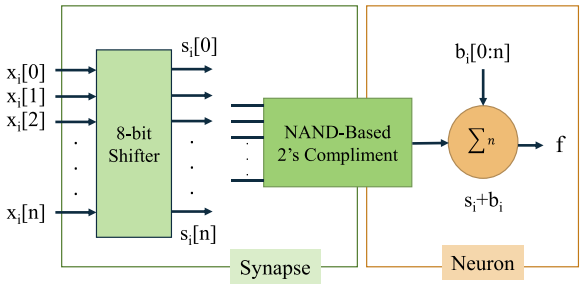

Fig. 6. Block diagram of a digital NAND gate-based binarized synapse-neuron unit for the SABiNN model.

components despite the diversity. The synapse unit takes in the output value from the last neuron of the layer and is multiplied by the trained weights. Then the associated activation function classifies the resultant value and sends it to the subsequent neuron layer [27]. When translating onto the edge devices, the matrix-multiplier components are used in the synapse-neuron unit of the NNs. The major drawback of using such a component is that the multiplier is a high power-consuming logic element [27], [45] in a high-power consumption rate. In our work, a NAND gate-based n-bit 2's complement logic has been proposed, which completely removes the multipliers whenever the values of the weights are -1 and directly forward passes it if the weights are $+1$. These weight values are based on the binarized hyperparameters achieved from the extracted model. Fig. 6 illustrates the proposed binarized neuron unit [45].

IV. Experimental and Measurement Results

A. Hardware and Edge Inference

Each network component was translated into the digital logic block and embedded into digital hardware after successfully training and evaluating the optimized model on software. The model hyperparameters were extracted from the Google Collaboratory Platform using TensorFlow. General-purpose Nexys Artix-7 field-programmable gate array (FPGA) was selected in the first phase of hardware inference to avoid error rate and increase flexibility in redesigning the SABiNN model if significant accuracy degradation occurs. The FPGA implementation was performed using Xilinx HLX software. In the network design, MAC operation was replaced by shifter, adders, and NAND gate-based 2's complement, as shown in Fig. 6. Hardware performance analysis, such as resource utilization, power consumption rate, and output signal generation, was measured to validate the proposed technique shown in Table V. By replacing MAC units, the model reduces its power consumption by 13× [27]. Current BiNNs use an XNOR gate in their MAC unit [57]. However, in CMOS ASIC design, the XNOR gate results in lower precision and higher noise-induced signal, contributing to significant accuracy degradation when using a dense number of XNOR gates in neuron–synapse connection. When implemented in CMOS, a significant -279 to 456-mV spike was observed on the XNOR gate. This spike repeatedly occurred when input A flipped the bit from "1" to "0," and input B flipped the bit from "0" to "1." To ensure clean voltage reading and reduce voltage spikes, we designed a NAND-based XNOR gate on the MAC unit, which achieved a higher precision rate and clean signal generation. Transient

TABLE V
Power Consumption Report and Hardware Utilization of SABiNN Block on Artix-7

Parameter	Power (W)	Parameter	Number
Signals	1.720	Slice LUT	108
Logic	2.472	Bonded I/O	18
I/O	0.068	Buffer	1
Dynamic	4.260	Registers	7
Static	0.68	Slice	33
Total Power	**4.365**		
Thermal Margin	11.8		

TABLE VI
Measurement of SABiNN Chip on CMOS

Parameter	Unit
Area	0.16mm^2
Detection Accuracy	88%
Network Layer	2-(8-12-6-4)-1
Frequency	9 MHz
CMOS Cell Count	113119
Clock Period	10 ns
Supply Voltage	1.8V
Power	~10 μW

analysis of the XNOR and NAND-based XNOR gate is presented in Fig. 7(a) and (b), indicating the voltage spikes in the XNOR gate. NN's main bottleneck for analyzing data is memory access. Each MAC operation requires three memory reads (weight, activation function, and bias) and one memory write for the calculated value (partial sum) [56]. A higher number of MAC blocks are used in the worst case scenarios where NN structures are deep, consisting of more than two hidden layers. Such models must use on-chip RAMs resulting in a high-power consumption rate [59]. Using the proposed model where each weight is binarized, the post-trained model requires near-zero on-chip memory access. Each activation function accesses low memory allocation due to using shifters instead of multipliers. The physical implementation of the SABiNN model on Nexys Artix-7 FPGA is shown in Fig. 8, where the switches are used in logging data points from the ECG, pulse-oximeter dataset, and seven-segment display showcases the binary output classified by the SABiNN block.

V. Design Space Exploration on CMOS Process

The SABiNN block was designed on Google–Skywater's open-source digital 130-nm PDK which provides a realistic approximation of area utilization, speed, and power consumption rate on silicon. We expanded our hardware model from 8-bit to 16-bit, and the network size was increased to a four-hidden layer model 2-(8-12-6-4)-1 for a higher accuracy rate. The overall accuracy of our model reached ~88% with a power consumption of ~10 μW. Table VI shows the measurement of the proposed model on the CMOS platform. The total area utilization of the proposed SABiNN model was only ~32%, enabling the opportunity to integrate

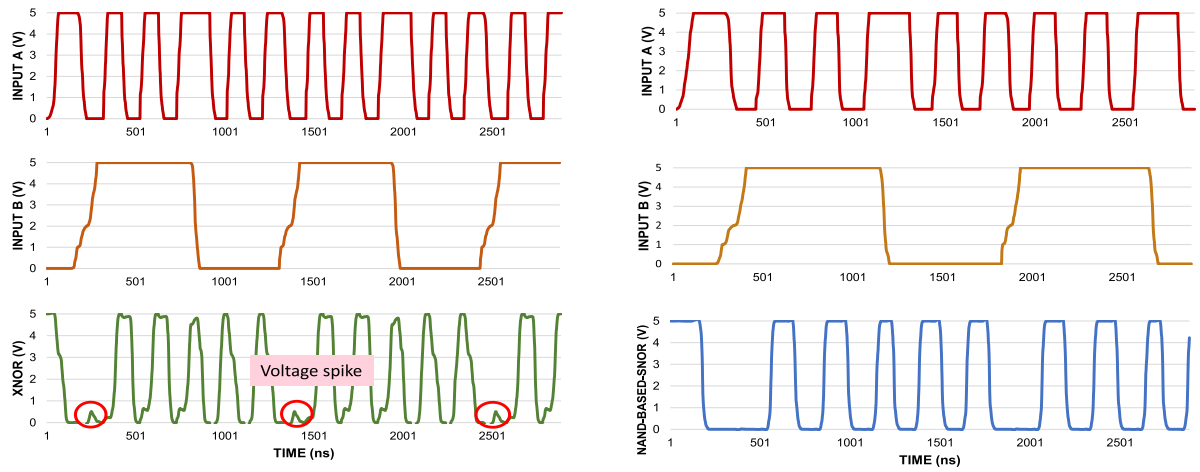

Fig. 7. Comparative study of the output values during transient analysis between (a) XNOR gate and (b) NAND-based XNOR gate. In traditional XNOR gate ∼(−279 to 456 mV), voltage spike is observed during bit flipping in transient analysis indicated at (a).

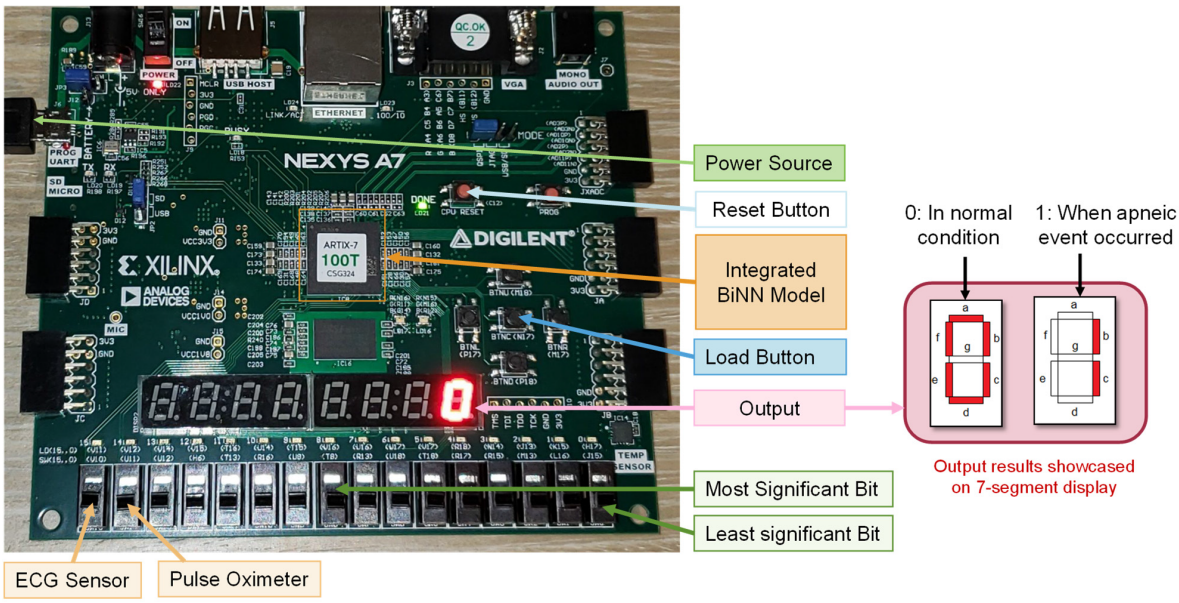

Fig. 8. Physical implementation of the SABiNN-based SA detection system using general-purpose Nexys Artix-7 FPGA. Binary input is logged using the switches, and the seven-segment displays showcases the classified result.

the signal processing block and required memory enabling a system-on-chip (SoC) solution. Fig. 9 showcases the simulated test result of the model in the Xilinx HLx software before running a digital design flow (PnR: placement and route) on the CMOS platform. In Fig. 9, the testbench simulated in the Xilinx software of the SABiNN model was designed into an open-source digital pad-frame called Caravel provided by Google+Skywater. It showed accurate results when data were logged from the testing dataset [48], [60]. From the test dataset, if the RR interval from the ECG signal was over 99% and SpO_2 values were less than 95%, the model detected possible apneic cases. The model showed normal conditions when the RR interval was below 100%, and SpO_2 was over 90%. From the measurement results, it can be concluded that SABiNN can be successfully deployed on CMOS with minimal area requirements. The 10-μW power consumption rate showcases a notable power usage reduction for a four-layer 16-bit NN model.

VI. Discussion

Integrating the trained SABiNN model with a power consumption rate of only ∼10 μW opens opportunities to infer deeper and denser NN models with higher complexity and sensitivity. Currently, most NN models for classification and prediction are built on cloud platforms limiting on-chip prediction and analysis, ensuring high accuracy with no assurance of power efficiency [18], [19], [20], [21], [26], [61], [62], [63]. In contrast to the FPGA schemes, neither ML nor NN models were used, and no power consumption rate was documented [64], [65], [66]. Our proposed SABiNN method can optimize and infer deep NNs (DNNs) more straightforwardly while cutting the overhead cost. The CMOS implementation introduces future intelligent wearable devices with on-chip classification without the help of the cloud and servers, enabling higher security and lower latency of around ∼10 μs per prediction. We are the first to introduce CMOS implementation of an optimized BiNN for SA detection.

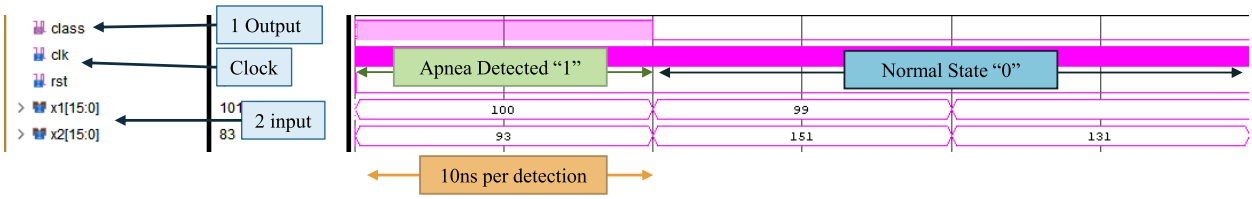

Fig. 9. Test bench simulation of the digital SABiNN model on the Xilinx HLx platform before integrating it onto CMOS using 130-nm Google + Skywater digital PDK. The simulation "class" is the output value in 1 bit, clk and rst are the clock and reset port of the digital design, and x1 and x2 are the 16-bit input values.

TABLE VII
COMPARISON OF THE PROPOSED WORK WITH STATE-OF-THE-ART METHODS FOR SA DETECTION

Year/Work	2022, [37]	2022, [38]	2022,[67]	2021, [28]	2022, [68]	**Proposed work**
Platform	Real-Time	Wearable	Software	IoT	ASIC [180nm]	**ASIC [130nm]**
Feature	Audio: snoring	Polysomnography (PSG): ECG	Nasal airflow: respiratory cycle, respiratory rate (RR), tidal volume (TV).	ECG and SpO_2	RR Interval, R-S Amplitude	**SpO_2 Signal and RR Interval (ECG)**
Classifier	TCN	SVM	XGBoost	1D-CNN	SVM	**SABiNN**
Sample window	1.04sec	60sec	5sec	1 sec	10 sec	**30 sec**
Sensitivity	-	71%	83.82%	97.44%	83.85%	**94%**
Specificity	96%	88%	85.97%	-	85.58%	**91%**
Accuracy	96%	83%	82.76%	99.62%	84.60%	**88%**
Area	N/A	N/A	N/A	N/A	0.429mm^2	**0.16mm^2**
Power	N/A	N/A	N/A	N/A	0.46uW	**10uW**
Energy	N/A	N/A	N/A	5uJ	0.46 nJ	**1pJ**

Table VII presents a comparative study between the proposed state-of-the-art methods in detecting SA. The three-step software–hardware co-simulation method of SABiNN implementation ensures a low error rate during transfer learning resulting in the validation of each step by reducing the fabrication error cost. Generally, a significant drop in accuracy is noted during the inference of the machine learning model on edge. Using our proposed SABiNN method, an accuracy of 88% has been achieved for the final 16-bit model on a 130-nm PDK process, demonstrating promising results. However, this technique is applicable in post-trained NNs as the weights are fixed after the inference on hardware. But the significant advantage of using the proposed SABiNN method is the reduction of memory access due to the need for storing hyperparametric values and the absence of multipliers in neuron units. The proposed method showed promising results and accurate testbench simulations when validated using an open-source, widely used dataset.

VII. CONCLUSION

The application of deep learning models, such as NNs, in medical diagnosis and monitoring is becoming increasingly popular. However, as NN delivers highly accurate prediction diagnosis, it still requires high-end computational processors which leverage expensive cloud services. Thus, higher accuracy does not mean energy-efficient models. In contrast, techniques enabling energy-efficient, cost-effective solutions must sacrifice performance accuracy. To overcome such issues, the proposed SABiNN method is energy-efficient, resulting in 5 mJ on general-purpose FPGAs and 1 pJ on the CMOS platform. Introducing binarized hyperparameters and redesigning MAC operations with the shifter-based neuron–synapse connection of the NN model significantly optimized the model and increased the energy efficiency. This is a unique and novel approach to designing and inferring any NN-type model on edge. The next step of this design technique will be to develop a fully SoC integrated biomedical system that can accurately detect and screen SA events with designated front-end sensors. The automatic classification and prediction will reduce workloads on caregivers and sleep experts while offering affordable healthcare devices to people.

REFERENCES

[1] M. M. Lyons, N. Y. Bhatt, A. I. Pack, and U. J. Magalang, "Global burden of sleep-disordered breathing and its implications," *Respirology*, vol. 25, no. 7, pp. 690–702, Jul. 2020, doi: 10.1111/resp.13838.

[2] *Sleep Apnea (OSA) | American Lung Association*. Accessed: Jun. 25, 2022. [Online]. Available: https://www.lung.org/lung-health-diseases/lung-disease-lookup/sleep-apnea

[3] D. Adam, "The pandemic's true death toll: Millions more than official counts," *Nature*, vol. 601, no. 7893, pp. 312–315, Jan. 2022, doi: 10.1038/d41586-022-00104-8.

[4] O. Vandenberg, D. Martiny, O. Rochas, A. van Belkum, and Z. Kozlakidis, "Considerations for diagnostic COVID-19 tests," *Nature Rev. Microbiol.*, vol. 19, no. 3, pp. 171–183, Mar. 2021, doi: 10.1038/s41579-020-00461-z.

[5] Y. Mardian, H. Kosasih, M. Karyana, A. Neal, and C.-Y. Lau, "Review of current COVID-19 diagnostics and opportunities for further development," *Frontiers Med.*, vol. 8, May 2021, doi: 10.3389/fmed.2021.615099.

[6] *Covid News: Lack of Quick Test to Diagnose Omicron or Delta Complicates Treatments—The New York Times*. Accessed: Jun. 25, 2022. [Online]. Available: https://www.nytimes.com/live/2022/01/03/world/omicron-covid-vaccine-tests

[7] K. G. Rögnvaldsson et al., "Obstructive sleep apnea is an independent risk factor for severe COVID-19: A population-based study," *Sleep*, vol. 45, no. 3, Mar. 2022, doi: 10.1093/SLEEP/ZSAB272.

[8] Y. Peker et al., "Effect of high-risk obstructive sleep apnea on clinical outcomes in adults with coronavirus disease 2019: A multicenter, prospective, observational clinical trial," *Ann. Amer. Thoracic Soc.*, vol. 18, no. 9, pp. 1548–1559, Sep. 2021, doi: 10.1513/AnnalsATS.202011-1409OC.

[9] F. Chung et al., "The association between high risk of sleep apnea, comorbidities, and risk of COVID-19: A population-based international harmonized study," *Sleep Breathing*, vol. 25, no. 2, pp. 849–860, Jun. 2021, doi: 10.1007/s11325-021-02373-5.

[10] *The Cost of Sleep Apnea*. Accessed: Jun. 25, 2022. [Online]. Available: https://www.webmd.com/sleep-disorders/sleep-apnea/cost-of-sleep-apnea

[11] O. Hassan, D. Parvin, and S. Kamrul, "Machine learning model based digital hardware system design for detection of sleep apnea among neonatal infants," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2020, pp. 607–610.

[12] Q. Xie, K. Faust, R. Van Ommeren, A. Sheikh, U. Djuric, and P. Diamandis, "Deep learning for image analysis: Personalizing medicine closer to the point of care," *Crit. Rev. Clin. Lab. Sci.*, vol. 56, no. 1, pp. 61–73, Jan. 2019, doi: 10.1080/10408363.2018.1536111.

[13] *Healthcare Statistics for 2021 | Policy Advice*. Accessed: Jun. 25, 2022. [Online]. Available: https://policyadvice.net/insurance/insights/healthcare-statistics/

[14] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019, doi: 10.1016/j.ymssp.2018.05.050.

[15] D. Kollias, A. Tagaris, A. Stafylopatis, S. Kollias, and G. Tagaris, "Deep neural architectures for prediction in healthcare," *Complex Intell. Syst.*, vol. 4, no. 2, pp. 119–131, Jun. 2018, doi: 10.1007/s40747-017-0064-6.

[16] S. Tuli et al., "HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments," *Future Gener. Comput. Syst.*, vol. 104, pp. 187–200, Mar. 2020, doi: 10.1016/j.future.2019.10.043.

[17] D. Kaul, H. Raju, and B. K. Tripathy, "Deep learning in healthcare," in *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*, D. P. Acharjya, A. Mitra, and N. Zaman, Eds. Cham, Switzerland: Springer, 2022, pp. 97–115, doi: 10.1007/978-3-030-75855-4_6.

[18] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3151947.

[19] S. Hu, W. Cai, T. Gao, and M. Wang, "A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ECG," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3193169.

[20] M. Yeo et al., "Robust method for screening sleep apnea with single-lead ECG using deep residual network: Evaluation with open database and patch-type wearable device data," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5428–5438, Nov. 2022, doi: 10.1109/JBHI.2022.3203560.

[21] M. M. Moussa, Y. Alzaabi, and A. H. Khandoker, "Explainable computer-aided detection of obstructive sleep apnea and depression," *IEEE Access*, vol. 10, pp. 110916–110933, 2022, doi: 10.1109/ACCESS.2022.3215632.

[22] J. N. Mcnames and A. M. Fraser, "Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram," in *Proc. Comput. Cardiol.*, Sep. 2000, pp. 749–752.

[23] R. K. Pathinarupothi, R. Vinaykumar, E. Rangan, E. Gopalakrishnan, and K. P. Soman, "Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Feb. 2017, pp. 293–296, doi: 10.1109/BHI.2017.7897263.

[24] A. M. da Silva Pinho, N. Pombo, and N. M. Garcia, "Sleep apnea detection using a feed-forward neural network on ECG signal," in *Proc. IEEE 18th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Sep. 2016, pp. 1–6, doi: 10.1109/HEALTHCOM.2016.7749468.

[25] S. M. I. Niroshana, X. Zhu, K. Nakamura, and W. Chen, "A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0250618, doi: 10.1371/JOURNAL.PONE.0250618.

[26] M. H. Chyad, S. K. Gharghan, H. Q. Hamood, A. S. H. Altayyar, S. L. Zubaidi, and H. M. Ridha, "Hybridization of soft-computing algorithms with neural network for prediction obstructive sleep apnea using biomedical sensor measurements," *Neural Comput. Appl.*, vol. 34, no. 11, pp. 8933–8957, Jun. 2022, doi: 10.1007/S00521-022-06919-W.

[27] O. Hassan et al., "Energy efficient deep learning inference embedded on FPGA for sleep apnea detection," *J. Signal Process. Syst.*, vol. 94, no. 6, pp. 609–619, Jan. 2022, doi: 10.1007/S11265-021-01722-7.

[28] A. John, K. K. Nundy, B. Cardiff, and D. John, "Multimodal multiresolution data fusion using convolutional neural networks for IoT wearable sensing," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 6, pp. 1161–1173, Dec. 2021, doi: 10.1109/TBCAS.2021.3134043.

[29] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, G. Juliá-Serdá, and A. G. Ravelo-García, "A method for sleep quality analysis based on CNN ensemble with implementation in a portable wireless device," *IEEE Access*, vol. 8, pp. 158523–158537, 2020, doi: 10.1109/ACCESS.2020.3019734.

[30] T. Wang, C. Lu, and G. Shen, "Detection of sleep apnea from single-lead ECG signal using a time window artificial neural network," *BioMed Res. Int.*, vol. 2019, pp. 1–9, Dec. 2019, doi: 10.1155/2019/9768072.

[31] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 416–427, May 2011, doi: 10.1109/TITB.2010.2087386.

[32] D. S. Morillo, J. L. R. Ojeda, L. F. C. Foix, and A. L. Jimenez, "An accelerometer-based device for sleep apnea screening," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 491–499, Mar. 2010, doi: 10.1109/TITB.2009.2027231.

[33] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, "SpO2 based sleep apnea detection using deep learning," in *Proc. IEEE 21st Int. Conf. Intell. Eng. Syst. (INES)*, Oct. 2017, pp. 000091–000096, doi: 10.1109/INES.2017.8118534.

[34] R. K. Pathinarupothi, E. S. Rangan, and K. P. Soman, "Single sensor techniques for sleep apnea diagnosis using deep learning," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2017, pp. 524–529, doi: 10.1109/ICHI.2017.37.

[35] M. Á. Herrero, J. J. García, and A. Jiménez, "Electrocardiogram-based detection of central sleep apnea: A full-record signal processing approach," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2019, pp. 1–6, doi: 10.1109/MEMEA.2019.8802189.

[36] D. Padovano, A. Martinez-Rodrigo, J. M. Pastor, J. J. Rieta, and R. Alcaraz, "On the generalization of sleep apnea detection methods based on heart rate variability and machine learning," *IEEE Access*, vol. 10, pp. 92710–92725, 2022, doi: 10.1109/ACCESS.2022.3201911.

[37] H. Luo, L. Zhang, L. Zhou, X. Lin, Z. Zhang, and M. Wang, "Design of real-time system based on machine learning for snoring and OSA detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1156–1160, doi: 10.1109/ICASSP43922.2022.9747393.

[38] M. Yeo et al., "Respiratory event detection during sleep using electrocardiogram and respiratory related signals: Using polysomnogram and patch-type wearable device data," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 2, pp. 550–560, Feb. 2022, doi: 10.1109/JBHI.2021.3098312.

[39] L. Haoyu, L. Jianxing, N. Arunkumar, A. F. Hussein, and M. M. Jaber, "An IoMT cloud-based real time sleep apnea detection scheme by using the SpO2 estimation supported by heart rate variability," *Future Gener. Comput. Syst.*, vol. 98, pp. 69–77, Sep. 2019, doi: 10.1016/j.future.2018.12.001.

[40] C. Shi, M. Nourani, G. Gupta, and L. Tamil, "Apnea MedAssist II: A smart phone based system for sleep apnea assessment," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2013, pp. 572–577, doi: 10.1109/BIBM.2013.6732560.

[41] W. Gu, L. Leung, K. C. Kwok, I.-C. Wu, R. J. Folz, and A. A. Chiang, "Belun ring platform: A novel home sleep apnea testing system for assessment of obstructive sleep apnea," *J. Clin. Sleep Med.*, vol. 16, no. 9, pp. 1611–1617, Sep. 2020, doi: 10.5664/JCSM.8592.

[42] H. Azimi et al., "Cloud processing of bed pressure sensor data to detect sleep apnea events," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–5, doi: 10.1109/MEMEA49120.2020.9137203.

[43] F. Massie, D. Mendes de Almeida, P. Dreesen, I. Thijs, J. Vranken, and S. Klerkx, "An evaluation of the NightOwl home sleep apnea testing system," *J. Clin. Sleep Med.*, vol. 14, no. 10, pp. 1791–1796, Oct. 2018, doi: 10.5664/JCSM.7398.

[44] G. L. do Pinheiro et al., "Validation of an overnight wireless high-resolution oximeter plus cloud-based algorithm for the diagnosis of obstructive sleep apnea," *Clinics*, vol. 75, p. e2414, Nov. 2020, doi: 10.6061/CLINICS/2020/E2414.

[45] O. Hassan, R. Thakker, T. Paul, D. Parvin, A. S. Mohammad Mosa, and S. K. Islam, "SABiNN: FPGA implementation of shift accumulate binary neural network model for real-time automatic detection of sleep apnea," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2022, pp. 1–6, doi: 10.1109/I2MTC48687.2022.9806534.

[46] A. Hazarika, A. Jain, S. Poddar, and H. Rahaman, "Shift and accumulate convolution processing unit," in *Proc. TENCON - IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 914–919, doi: 10.1109/TENCON.2019.8929364.

[47] L. Almazaydeh, K. Elleithy, and M. Faezipour, "Obstructive sleep apnea detection using SVM-based classification of ECG signal features," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 4938–4941, doi: 10.1109/EMBC.2012.6347100.

[48] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Proc. Comput. Cardiol.*, vol. 27, Sep. 2000, pp. 255–258, doi: 10.1109/CIC.2000.898505.

[49] *St. Vincent's University Hospital / University College Dublin Sleep Apnea Database V1.0.0*. Accessed: Oct. 3, 2022. [Online]. Available: https://physionet.org/content/ucddb/1.0.0/

[50] P. Anderer, G. Gruber, S. Parapatics, and G. Dorffner, "Automatic sleep classification according to rechtschaffen and Kales," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 3994–3997, doi: 10.1109/IEMBS.2007.4353209.

[51] M. Cheng, W. J. Sori, F. Jiang, A. Khan, and S. Liu, "Recurrent neural network based classification of ECG signal features for obstruction of sleep apnea detection," in *Proc. IEEE Int. Conf. Comput. Sci. Eng. (CSE) IEEE Int. Conf. Embedded Ubiquitous Comput. (EUC)*, vol. 2, Jul. 2017, pp. 199–202, doi: 10.1109/CSE-EUC.2017.220.

[52] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6444, no. 2. Berlin, Germany: Springer, 2010, pp. 152–159, doi: 10.1007/978-3-642-17534-3_19.

[53] L. Sun, Z. Shang, Q. Cao, K. Chen, and J. Li, "Electrocardiogram diagnosis based on SMOTE+ENN and random forest," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11644, 2019, pp. 747–757, doi: 10.1007/978-3-030-26969-2_71.

[54] C. Song, K. Liu, Zhang, Xi, L. Chen, and X. Xian, "An obstructive sleep apnea detection approach using a discriminative hidden Markov model from ECG signals," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1532–1542, Jul. 2016, doi: 10.1109/TBME.2015.2498199.

[55] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel, "A novel algorithm for the automatic detection of sleep apnea from single-lead ECG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 9, pp. 2269–2278, Sep. 2015, doi: 10.1109/TBME.2015.2422378.

[56] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv Neural Inf. Process Syst.*, vol. 29, no. 4, 2016, Accessed: Sep. 25, 2022. [Online]. Available: https://github.com/itayhubara/BinaryNet

[57] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908, 2016, pp. 525–542, doi: 10.1007/978-3-319-46493-0_32.

[58] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, "A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal," *Neurocomputing*, vol. 294, pp. 94–101, Jun. 2018, doi: 10.1016/J.NEUCOM.2018.03.011.

[59] Y. Ma, Y. Cao, S. Vrudhula, and J.-S. Seo, "Optimizing loop operation and dataflow in FPGA acceleration of deep convolutional neural networks," in *Proc. ACM/SIGDA Int. Symp. Field-Programmable Gate Arrays*, Feb. 2017, doi: 10.1145/3020078.3021736.

[60] M. J. Lado, X. A. Vila, L. Rodríguez-Liñares, A. J. Méndez, D. N. Olivieri, and P. Félix, "Detecting sleep apnea by heart rate variability analysis: Assessing the validity of databases and algorithms," *J. Med. Syst.*, vol. 35, no. 4, pp. 473–481, Aug. 2011, doi: 10.1007/S10916-009-9383-5.

[61] D. Àlvarez et al., "A machine learning-based test for adult sleep apnoea screening at home using oximetry and airflow," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Mar. 2020, doi: 10.1038/s41598-020-62223-4.

[62] L. Almazaydeh, K. Elleithy, M. Faezipour, and A. Abushakra, "Apnea detection based on respiratory signal classification," *Proc. Comput. Sci.*, vol. 21, pp. 310–316, Jan. 2013, doi: 10.1016/J.PROCS.2013.09.041.

[63] M. Qatmh et al., "Sleep apnea detection based on ECG signals using discrete wavelet transform and artificial neural network," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, Feb. 2022, pp. 1–5, doi: 10.1109/ASET53988.2022.9735064.

[64] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, J. L. Navarro-Mesa, G. Juliá-Serdá, and A. G. Ravelo-García, "A portable wireless device based on oximetry for sleep apnea detection," *Computing*, vol. 100, no. 11, pp. 1203–1219, Nov. 2018, doi: 10.1007/S00607-018-0624-7.

[65] K. M. Al-Ashmouny, H. M. Hamed, and A. A. Morsy, "FPGA-based sleep apnea screening device for home monitoring," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Dec. 2006, pp. 5948–5951, doi: 10.1109/IEMBS.2006.260655.

[66] M. R. Prathipa, M. M. Arun, M. Niranjana, S. Preethi, and R. P. Jennifer, "FPGA based sleep disorder detection using brain waves," *Int. J. Emerging Technol. Comput. Sci. Electron.*, vol. 26, pp. 976–1353, Jun. 2019.

[67] X. Yan, L. Wang, J. Zhu, S. Wang, Q. Zhang, and Y. Xin, "Automatic obstructive sleep apnea detection based on respiratory parameters in physiological signals," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2022, pp. 461–466, doi: 10.1109/ICMA54519.2022.9856347.

[68] R. Parmar, M. Janveja, G. Trivedi, P. Jan, and Z. Nemec, "An area and power efficient VLSI architecture to detect obstructive sleep apnea for wearable devices," in *Proc. 32nd Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2022, pp. 1–5, doi: 10.1109/RADIOELEKTRONIKA54537.2022.9764917.

Omiya Hassan received the B.Sc. degree in electrical engineering from United International University, Dhaka, Bangladesh, in 2017, and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering and Computer Science (EECS), University of Missouri, Columbia, MO, USA, in 2023.

She served as a Lecturer at Presidency University, Dhaka, in 2018. Her research topic focuses on designing energy-efficient machine learning (ML) model-based integrated circuits for biomedical applications, and her research interest includes low-power circuit design, biomedical system design, and ML hardware accelerators.

Dr. Hassan received the IEEE Instrumentation and Measurement Society's Graduate Research Fellowship Award in 2021.

Tanmoy Paul received the bachelor's and master's degrees from the Department of Electrical and Electronics Engineering, University of Dhaka, Dhaka, Bangladesh, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.

He is currently working on developing artificial intelligence model to detect sleep apnea from physiological signals. His primary focus is to develop models with minimal parameters and reduced floating-point operations making the model suitable for on-chip implementation.

Nazmul Amin received the B.S. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2018, and the M.S. degree in electrical engineering from the University of Missouri, Columbia, MO, USA, in 2022.

Following his M.S. degree, he is working with the Battery Charging Product (BCP) Group, Texas Instruments (TI) Inc. Dallas, TX, USA, as a Mixed Signal Design Verification Engineer.

Mr. Amin's awards and honors include DAC Young Fellow in 2021 and graduate student fellowships in the University of Missouri.

Twisha Titirsha received the bachelor's degree from the Military Institute of Science and Technology, Dhaka, Bangladesh, in 2015. She is currently pursuing the Ph.D. degree from the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.

Her research interests include semiconductor device modeling, fabrication, and mixed-signal circuit design.

Rushil Thakker received the B.Sc. degree from the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA, in 2022.

He currently works as a Product Development Associate Engineer at Caterpillar, Peoria, IL, USA. He has been affiliated with the Analog/Mixed-Signal, VLSI and Devices Laboratory (AVDL), University of Missouri, since Fall 2020. His research involves embedding machine learning models onto digital hardware designs to create energy-efficient biomedical devices.

Dilruba Parvin (Member, IEEE) received the bachelor's degree in electrical, electronics, and communication engineering (EECE) from the Military Institute of Science and Technology (MIST), Dhaka, Bangladesh, in 2016, and the Ph.D. degree in electrical and computer engineering from the University of Missouri, Columbia, MO, USA, in 2022.

She worked simultaneously with the Analog/Mixed-Signal, VLSI and Devices Laboratory (AVDL), University of Missouri, as a Research Assistant, and with the Department of Electrical Engineering and Computer Science (EECS) as a Graduate Instructor, where she also served as a Graduate Teaching Assistant for two years. She is currently working as an Analog Design Engineer at Texas Instruments (TI) Inc. Dallas, TX, USA. Her research topic focused on the development of low-power integrated circuits for RF energy harvesting applications. In addition, she worked on developing machine learning (ML) embedded energy-efficient integrated circuits for energy harvesters.

Abu Saleh Mohammad Mosa received the B.S. degree in computer science from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2006, the M.S degree from University College Dublin, Dublin, Ireland, in 2009, and the Ph.D. degree in health informatics from the University of Missouri, Columbia, MO, USA, in 2015.

He is a visionary informatician who believes in the power of data to predict the future and transform healthcare. With expertise in biomedical informatics, his focus is on healthcare innovation, patient-centered outcomes, and the application of precision medicine. During his nine-year tenure as the Director of Research Informatics at the University of Missouri (MU) School of Medicine, Columbia, MO, USA, he dedicated his efforts to enhancing institutional research capacity, improving access to real-world observational data, and applying advanced informatics and data science approaches for data management and analysis. He currently serves as the Senior Director of Informatics Technology and an Associate Professor of biomedical informatics at the MU School of Medicine, with adjunct and core faculty appointments in the Department of Electrical Engineering and Computer Science and the Institute for Data Science and Informatics, respectively.

Dr. Mosa received the National Institutes of Health (NIH) Data Science Rotations for Advancing Discovery (RoAD-Trip) Fellowship in 2018 and was inducted as a fellow of the American Medical Informatics Association (FAMIA) in 2020. He has served as a principal investigator (PI) and a co-investigator on numerous extramural research grants funded by organizations, such as Patient-Centered Outcome Research Institute (PCORI), NIH, The Agency of Healthcare Research and Quality (AHRO), State of Missouri, and industry partners. He is also the site PI for MU's participation in the Greater Plains Collaborative (GPC) PCORNet Clinical Research Network, a national healthcare data infrastructure project funded by PCORI.

Syed Kamrul Islam (Senior Member IEEE) received the B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 1983, and the M.S. and Ph.D. degrees in electrical and systems engineering from the University of Connecticut, Storrs, CT, USA, in 1987 and 1994, respectively.

He is currently serving as a Professor and the Chair of the Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA. His research interests include analog/mixed-signal integrated circuits, semiconductor devices, nanotechnology, bio-microelectronics, and monolithic sensors. He has more than 100 publications in refereed journals, more than 150 papers in conference proceedings, and several invited talks. He also coauthored a book and 14 book chapters. Prior to joining the Department of Electrical Engineering and Computer Science, University of Missouri, in July 2018, he served as a James W. McConnell Professor and the Associate Head of the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN, USA.

In recognition of his teaching, research, and related efforts at the University of Tennessee, he received the John W. Fisher Professorship, the Eta Kappa Nu Outstanding Teacher Award, the Moses E. and Mayme Brooks Distinguished Professor Award, the College of Engineering Research Fellow Award, The Gonzalez Family Award for Excellence in Teaching, the Tickle College of Engineering Teaching Fellow, the University of Tennessee Citation for Research and Creative Achievement, the Electrical and Computer Engineering Faculty of the Year Award, and the Alexander Prize.