Power Efficient Machine Learning Based Hardware Architectures for Biomedical Applications: Sleep Apnea Screening

Topic: TinyML hardware techniques, silicon implementations, and design ideas

TinyML Asia 2022
• Motivation
• Proposed System Design
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Design Method
• Work in Progress
  • Physical Design Exploration on Silicon
  • Design Exploration on Reconfigurable Hardware
  • Hardware Implementation
• Summary
SURVEY: DIAGNOSIS OF SLEEP APNEA (SA)


EXPENSIVE TREATMENT
• INADEQUATE TESTING
• LIMITED RESOURCES

80% Among 329.8 million adult population in USA, 80% are undiagnosed for SA [1]

9% Among every 100 adult 9 have guaranteed apneic conditions [1]

The chances of sudden death [2] People with sleep apnea have 3 times higher

1. https://www.sleepfoundation.org/sleep-apnea/
MARKET ANALYSIS:

STRENGTHS & LIMITATIONS

SMART DIAGNOSIS TOOL

- Requires multi-level of experts
- Requires existing device and labs
- High accuracy rate
- Affordable
- Not wearable

SMART WATCH: HEALTH TRACKER

- Not recommended for SA patients
- Portable
- Low accuracy rate
- Affordable
- Wearable

COMMERCIAL SOFTWARE

- Requires multi-level of experts
- Requires sleep labs
- High sensitivity
- Expensive
- Wearable

HOSPITAL POLYSOMNOGRAM

(W) Pro

STRENGTHS & LIMITATIONS:

- Wearable
- Affordable
- High accuracy rate
- Requires existing device and labs
- Requires multi-level of experts

- Not wearable
- Not recommended for SA patients
- Portable
- Low accuracy rate
- Affordable
- Wearable

- Requires sleep labs
- High sensitivity
- Expensive
- Wearable

Image source: google
| Work | Method | Algorithm | Sensor | Accuracy (%)
|------|--------|-----------|--------|-------------
| Kopaczka et al. 2017 [4] | SA Detection | Wavelet transform, Spectral analysis | Thermal infrared | N/A

AHI = Apnea-Hypopnea Index
Designing accessible, low-power, wearable and portable biomedical devices capable of detecting apnea events with high accuracy and precision.

Key Features:
- Wearable and portable
- Real-time detection
- Compact design architecture
- Power efficient
- Affordable and accessible

Enabling smart detection of apnea events for improved health monitoring.
CAUSE OF SLEEP APNEA

- Inadequate air supply
- Lower blood oxygen level
- Irregular heart-beat
- Throat muscle relax

https://www.mayoclinic.org/diseases-conditions/sleep-apnea/symptoms-causes/syc-20377631

Adhesive ECG Patch
Pulse Oximeter
Integrated Sensor Technology

CAUSE OF SLEEP APNEA
Apnea event detected.
Alarm signal will initiate after 1 is counted over 5 times per hour.

**Input Sensors**
- Pulse Oximeter:
  - Takes in digital data from sensors
  - Uses the proposed block for classification
  - Gives binary classification between "0" and "1"
  - Gives output between "0" and "1"

**Processing Block**
- Takes in digital data from sensors
- Takes in binary data from sensors
- Takes in data from two biosensors

**Decision Block**
- Output block
  - System will keep on reading data. No alarm will initiate if 0 is counted over 5 times per hour.
  - Alarm signal will initiate if 1 is counted over 5 times per hour.

**Diagram**
- Adhesive ECG Patch
- Adhesive
- ECG signal
- Processed input data
  - Design Target
  - Taking R-R peak difference from 30 second segment
  - Converting each segment value to integers
  - Cleaning outliers and converting to integer values
- Single Lead ECG:
  - Takes in distance between R-R peak values of ECG signal
  - Takes in the distance between R-R peak values of ECG signal

**Alarm Signal**
- 0: Normal condition
- 1: Apnea event detected

**Signal Processing**
- Pulse Oximeter (SpO2):
  - Takes in the blood oxygen saturation percentage level (SpO2)
  - Cleaning outliers and converting to integer values
PRESENTATION
OUTLINE

• Motivation
• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Hardware Implementation
• Design Exploration on Reconfigurable Hardware
• Physical Design Exploration on Silicon
• Work in Progress
• Summary

10
SOFTWARE - HARDWARE CO-SIMULATION PROCESS

Integrated Circuit Design
- Fabrication of the proposed model layout
- Converging optimal design to CMOS Platform

Model Optimization and Extraction
- Training and Evaluation of Software Model
- Collecting and Pre-Processing data
- Extracting Model Parameters
- Validating Optimized Model
- Implementing Optimization Techniques

Hardware Model Design and Simulation
- Optimized Model Inference on Digital Hardware
- Hardware Optimization and Implementation

Integrated Circuit Design
- Converting optimal design to CMOS Platform
• Motivation
• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Hardware Implementation
• Design Exploration on Reconfigurable Hardware
• Physical Design Exploration on Silicon
• Work in Progress
• Summary
Sample Number: 35,800

Dataset Collection from:
Phillips University & ST. VINCENT HOSPITAL

Subjects:
- Male: 21
- Female: 4
- Unidentified: 8

Age: 28-68 years

Annotation:
- Normal condition: 0
- Apneic condition: 1

7. htps://archive.physionet.org/physiobank/
SIGNAL PROCESSING AND DATA GENERATION

Special Thanks to
Dr. Abu Mosa, Director of Research Informatics
Tanmoy Paul, Doctoral Student, University of Missouri

Adhesive ECG Patch
Pulse Oximeter
Digital Sample: Sampling data with 100 Hz frequency
Segmentation: Conversion of 8-10hrs sleep data to 30 sec
Artifact Removal: SpO2 signal variance: 4%
R-R interval variance: 20%

ECC Signal: R-R peak interval extraction
Dataset Generation 1D Sensor Data
Data Average: 2 input 1-dimension
Down-sample of SpO2: 1 Hz frequency

Digital Sample: Without 100 Hz sampling data

Data Generation
**Model Selection & Training**

- **Training Features**
  - \( x_1: \text{ECG Data} \)
  - \( x_2: \text{SpO}_2 \) Data

- **Training set:** 70%
- **Testing set:** 30%
- **K-fold cross validation (k=10)**
- **Epochs:** 100
  - Train set: 90%
  - Validation set: 10%

**Evaluation Metrics**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>F-N Model (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-score</td>
<td>84</td>
</tr>
<tr>
<td>Accuracy</td>
<td>80 (CI: 0.1%)</td>
</tr>
<tr>
<td>Specificity</td>
<td>78</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>99</td>
</tr>
</tbody>
</table>

**Activation Functions**

- **Hidden Layer:** ReLU
- **Output Layer:** Sigmoid

**Loss Function & Optimizer**

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>Loss Function</th>
<th>Activation Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM</td>
<td>Mean Squared Error (MSE)</td>
<td>( y = \text{sigmoid} )</td>
</tr>
<tr>
<td>ADAM</td>
<td>Mean Squared Error (MSE)</td>
<td>ReLU</td>
</tr>
<tr>
<td>ADAM</td>
<td>Mean Squared Error (MSE)</td>
<td>( \text{hidden layer activation} )</td>
</tr>
<tr>
<td>ADAM</td>
<td>Mean Squared Error (MSE)</td>
<td>( \text{output layer activation} )</td>
</tr>
</tbody>
</table>
MODEL SELECTION & TRAINING

BALANCED MODEL

SMOTE: Synthetic Minority Oversampling Technique
• Generates synthetic sample data from minority class
• Takes the nearest neighbor from sample cluster at random and selects the convex between two sample points.

Tomek:
• Used for under-sampling
• Takes nearest neighbor samples of both classes and removes the sample which belongs to major class.

ENN: Edited Nearest Neighbors
• Removes mis-classified samples
• Improves class balance
• Takes the nearest neighbor sample from sample cluster at random and selects the convex between two class neighbors.

Parameter

<table>
<thead>
<tr>
<th>Sample</th>
<th>SMOTE</th>
<th>SMOTE+ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>78</td>
<td>70</td>
<td>68</td>
</tr>
<tr>
<td>78</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>80</td>
<td>72</td>
<td>70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SMOTE</th>
<th>SMOTE+ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>84</td>
<td>83</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>Specificity</td>
<td>78</td>
<td>71</td>
</tr>
<tr>
<td>Accuracy</td>
<td>80</td>
<td>72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SMOTE</th>
<th>SMOTE+ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOTE</td>
<td>6255</td>
<td>8325</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>8325</td>
<td>9557</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>9657</td>
<td>8379</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>13560</td>
<td>8325</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SMOTE</th>
<th>SMOTE+ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOTE</td>
<td>71.4%</td>
<td>72%</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>71%</td>
<td>71%</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>78%</td>
<td>78%</td>
</tr>
<tr>
<td>SMOTE+ENN</td>
<td>83%</td>
<td>83%</td>
</tr>
</tbody>
</table>

SMOTE: Synthetic Minority Oversampling Technique

ENN: Edited Nearest Neighbors

Tomek:
• Takes the nearest neighbor sample from sample cluster at random and selects the convex between two class neighbors.
Can resource utilization be decreased significantly?

Can multipliers be replaced?
SABiNN: Shift Accumulate Based Binarized Neural Network

- Binarized based synapse-neuron
- Binarized hyperparameters only
- Quantized input
- Shifter-based activation function
- Absence of multipliers
- Constrained weights in +1 and -1


Yoshua Bengio Turing Award Winner’18 CIFAR Fellow
**Model Performances**

<table>
<thead>
<tr>
<th>Model</th>
<th>F1-score</th>
<th>84</th>
<th>74</th>
<th>78</th>
<th>70</th>
<th>68</th>
<th>76</th>
<th>80 (CI: 0.7)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABiNN</td>
<td></td>
<td>84</td>
<td>74</td>
<td>78</td>
<td>70</td>
<td>68</td>
<td>76</td>
<td>80 (CI: 0.7)</td>
<td>77</td>
</tr>
<tr>
<td>FNN Model</td>
<td>(%) SABiNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Evaluation Metrics**

- Train FNN model
- Get weights
- Feed forward new NN
- Weight extraction: get weights
- Weight distribution: mean value between weights & element-wise sign distribution
- Binarize weight distribution
- Plug in weights
- Feed forward new NN
- Validate model & evaluate

**Model Performances**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model: FNN</td>
<td>78</td>
</tr>
<tr>
<td>Optimized FNN</td>
<td>78</td>
</tr>
<tr>
<td>DeepSAC</td>
<td>81</td>
</tr>
<tr>
<td>SABiNN</td>
<td>84</td>
</tr>
</tbody>
</table>

**Latency (μs)**

- Baseline Model: 6.05
- Optimized FNN: 7.64
- DeepSAC: 8.58
- SABiNN: 9
<table>
<thead>
<tr>
<th>CLASSIFIER</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This Work</td>
<td>88</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>SABiNN</td>
<td>88</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>TW-MLP</td>
<td>87</td>
<td>85</td>
<td>N/A</td>
</tr>
<tr>
<td>HMM-SVM</td>
<td>86</td>
<td>82</td>
<td>86</td>
</tr>
<tr>
<td>LS-SVM</td>
<td>82</td>
<td>N/A</td>
<td>86</td>
</tr>
<tr>
<td>LS</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>F1-Score</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

All classifier used ApneOsc Dataset for training, validation, and testing.
ACCURACY EVALUATION

Post-training prediction accuracy between proposed models and market-ready devices and software.

System vs Prediction Accuracy (%)

- **SABiNN**: 72%
- **DeepSAC**: 78-79%
- **FNN**: 80%
- **Watch-PAT**: 63%
- **Google**: 60%

---

18. Tauman, Riva et. al "Watch-PAT is Useful in the Diagnosis of Sleep Apnea in Patients with Atrial Fibrillation", 2020
PRESENTATION

OUTLINE

• Motivation
• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Design Exploration on Reconfigurable Hardware
• Design Exploration on Silicon
• Hardware Implementation
• Work in Progress
• Summary

PRESENTATION
Model Training
Optimization and Compression
Hardware Simulation
Power Consumption Rate Study
Resource Analysis
Power Consumption Rate
Design Area
Performance and Timing
Performance Speed
Logic Block Utilization
Power Rate
Power Consumption Rate
Silicon Design
FPGA Implementation
PRESENTATION

OUTLINE

• Motivation
• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Hardware Implementation
• Design Exploration on Reconfigurable Hardware
• Physical Design Exploration on Silicon
• Work in Progress
• Summary
**SABiNN Hardware Simulation Results**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Power (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal Margin</td>
<td>12.1</td>
</tr>
<tr>
<td>Total</td>
<td>4.331</td>
</tr>
<tr>
<td>Static</td>
<td>0.416</td>
</tr>
<tr>
<td>Dynamic</td>
<td>4.015</td>
</tr>
<tr>
<td>Logic block</td>
<td>1.725</td>
</tr>
<tr>
<td>I/O Ports</td>
<td>0.117</td>
</tr>
<tr>
<td>Signals</td>
<td>1.446</td>
</tr>
</tbody>
</table>

Testbench generation

BINN layers (8-6-4) schematic diagram for 2 input (8-bit) using Vivado Hdlc block design software.
Power Consumption Report

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SABiNN (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Temp (max)</td>
<td>4.368</td>
</tr>
<tr>
<td>Flipflops</td>
<td>0.105</td>
</tr>
<tr>
<td>Registers</td>
<td>4.263</td>
</tr>
<tr>
<td>Buffer</td>
<td>0.068</td>
</tr>
<tr>
<td>I/O Bonded</td>
<td>2.459</td>
</tr>
<tr>
<td>LUTs</td>
<td>1.736</td>
</tr>
</tbody>
</table>

Resource Utilization

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SABiNN (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUTs</td>
<td>108</td>
</tr>
<tr>
<td>I/O Bonded</td>
<td>18</td>
</tr>
<tr>
<td>Registers</td>
<td>7</td>
</tr>
<tr>
<td>Flipflops</td>
<td>33</td>
</tr>
<tr>
<td>Buffers</td>
<td>1</td>
</tr>
<tr>
<td>Logic block</td>
<td>1</td>
</tr>
<tr>
<td>Signals</td>
<td>33</td>
</tr>
<tr>
<td>Dynamic</td>
<td>0.068</td>
</tr>
<tr>
<td>Static</td>
<td>4.263</td>
</tr>
</tbody>
</table>

Power Source

Output results showcased on 7-segment display

0: In normal condition
1: When apnea event occurred

Input data transmitted via USB/UART

Power Source

Thermal Margin

<table>
<thead>
<tr>
<th></th>
<th>SABiNN (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>11.4</td>
</tr>
<tr>
<td>Static</td>
<td>0.105</td>
</tr>
<tr>
<td>Dynamic</td>
<td>4.263</td>
</tr>
<tr>
<td>Logic block</td>
<td>0.068</td>
</tr>
<tr>
<td>I/O Ports</td>
<td>2.459</td>
</tr>
<tr>
<td>Signals</td>
<td>1.736</td>
</tr>
</tbody>
</table>

IEEE Instrumentation & Measurement Society
• Summary
• Work in Progress

Summary

• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Design Exploration on Reconfigurable Hardware
• Physical Design Exploration on Silicon
• Hardware Implementation

• Motivation

• Outline
NAND GATE - BASED XNOR GATE:

- Replaced XNOR with NAND gate based XNOR
- Re-designed 2's complement with NAND gate based XNOR

Truth Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Area: XNOR 225um² and NAND gate based XNOR 368um²
Integrated Circuit Design

ASIC Implementation

Parameter | Unit | Value
--- | --- | ---
Area | \(0.16 \text{ mm}^2\) | 
Frequency | MHz | 9 MHz
Cell Count | | 11319
Clock Period | ns | 100
Supply Voltage | V | 1.8
Power | \(\sim 10\mu\text{W}\) | 

Design Characteristics
- 130 nm process
- 16bit SABiNN digital model
- 4 hidden layers (6-12-6-4)
- 88% Accuracy
- 1 clock port, 1 reset port
- 32 input pins
- 1 output pin

SABiNN Chip
- Total die area: 20 mm²
- Design area: 10 mm²

Measurement of SABiNN Chip

In Process of Fabrication
• Motivation
• Proposed System Design
• Design Method
• Signal Processing and Data Generation Scheme
• Model Selection and Optimization
• Hardware Implementation
• Design Exploration on Reconfigurable Hardware
• Physical Design Exploration on Silicon
• Future Plan
• Summary
### Future Plan

- **32-bit Input Ports**
  - 8 unit input layer
  - 6 unit hidden layer
  - 4 unit hidden layer
  - Output Unit

- **BiNN layers (8-6-4)** schematic diagram for 2 input (8-bit) using Vivado HLx IP block design software

- **180 nm Process**

- **ASIC**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>50 µW</td>
</tr>
<tr>
<td>Power consumption in hr</td>
<td>0.05 mWh</td>
</tr>
</tbody>
</table>

- **FPGA Model**

- **Circuit Layout**

- **Target Power**

- **3.3V watch battery provide**

- **Reference**
  - https://github.com/omiya2106/SSCS_PICO-chips
  - https://www.batteryjunction.com/technology/3v-watch-battery-provide.html

- **Future PLAN**

- **30 nm Process**

- **Semiconductor**

- **MUSE**

- **ASIC**

- **FPGA Model**

- **Circuit Layout**

- **Future PLAN**
- Developed power-efficient neural network-based hardware model for the detection and screening of sleep apneaic events.
- Implemented the model on re-programmable hardware and on silicon for power measurement analysis, resource utilization and design space exploration.
- On the process of fabricating the design architecture on CMOS platform using 130nm and 180nm process.
Power Efficient Machine Learning Based Hardware Architectures for Biomedical Applications: Sleep Apnea Screening

OMIYA HASSAN
omiya.hassan@mail.missouri.edu

Looking Forward to Questions

Applying Machine Learning Based Hardware Architectures for Biomedical Applications: Sleep Apnea Screening
Copyright Notice

This presentation in this publication was presented at tinyML® Asia 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org