IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

VLSI Design Simulation for AI-Enhanced Brain-Machine Interfaces

Rashmita Bhaskar Behera, Harsh Anupsingh Raghuwanshi, Chaitanya Suhas Ramteke
Student
Prof. M. N. Kakatkar
Guide
Department of Electronics and Telecommunication, Sinhgad College of Engineering, Pune, Maharashtra, India

Abstract : AI-enhanced brain-machine interfaces (BMIs) require specialized hardware to process EEG signals in real time for applications like speech decoding in locked-in syndrome patients. This paper presents a detailed VLSI design simulation implemented on a Xilinx Artix-7 100T FPGA using Vivado, incorporating a preprocessing unit, a float16-quantized CNN accelerator, and a UART interface. The system achieves a latency of 10 ms per inference, power consumption of 1.5 W, and a throughput of 100 inferences per second, optimized for resource-constrained environments. We provide an exhaustive analysis of the architecture, quantization strategies, simulation environment, and performance metrics, comparing with state-of-the-art FPGA designs. The results demonstrate scalability for portable BMIs, with future directions focusing on ASIC implementations and clinical validation [1, 2].

IndexTerms - **Brain-Machine Interfaces, EEG Processing, FPGA, VLSI Design, AI Accelerator, Low-Power Hardware, Real-Time Processing**

1. Introduction

Very Large Scale Integration (VLSI) design is critical for deploying AI-driven brain-machine interfaces (BMIs) in portable, energy-efficient systems, particularly for real-time EEG processing [3]. EEG-based BMIs enable communication for patients with locked-in syndrome by decoding neural signals into speech or commands, but their practical deployment requires hardware that balances performance, power, and size constraints [2]. Field-Programmable Gate Arrays (FPGAs) offer a flexible platform for prototyping such systems, with the Xilinx Artix-7 100T providing a cost-effective solution for low-power applications [1].

This paper details a VLSI design simulation for an AI-enhanced BMI, targeting real-time EEG processing for speech decoding. The system integrates a dual-input CNN model, optimized with float16 quantization, and is simulated on the Artix-7 100T using Vivado. We address challenges like resource limitations, timing closure, and power efficiency, drawing on state-of-the-art FPGA implementations [4]. The paper is structured as follows: Literature Review reviews related work, Methodology describes the methodology, Results presents results, Discussion discusses implications, and Conclusion outlines future directions. The review draws on recent advancements in AI and neural engineering [5, 6].

2. Literature Review

VLSI design for neural interfaces has seen rapid advancements, driven by the need for real-time, low-power processing in BMIs [3]. Early FPGA-based designs focused on signal preprocessing, such as filtering and

feature extraction, but lacked AI integration [7]. The introduction of deep learning accelerators on FPGAs has transformed the field, enabling end-to-end neural signal processing [1]. For instance, [1] implemented an EEGNet-based accelerator on a Zynq-7020 FPGA, achieving 15 ms latency but consuming 2.0 W, highlighting power efficiency challenges.

Recent studies have explored quantization techniques to reduce model complexity, with float16 and int8 formats balancing accuracy and hardware efficiency [2]. [4] proposed a CNN accelerator on a Virtex-7 FPGA, achieving 12 ms latency and 1.8 W power consumption, but their design required significant resources, limiting scalability. Pipelining and loop unrolling have been widely adopted to optimize throughput, as demonstrated in [5]. Additionally, ASIC designs offer superior power efficiency but lack the flexibility of FPGAs for prototyping [3].

Our work builds on these advancements by integrating a float16-quantized CNN into a compact FPGA design, targeting the Artix-7 100T. The system prioritizes low latency and power consumption, addressing the needs of portable BMIs for clinical applications.

3. Methodology

3.1 VLSI Architecture

The VLSI system is designed to process EEG signals in real time, comprising three primary components:

- Preprocessing Unit: Implements digital signal processing for EEG cleaning. A 512-tap FIR bandpass filter (0.5–40 Hz) removes low-frequency drifts and high-frequency noise, while a 50 Hz notch filter (Q-factor=30) eliminates powerline interference. The unit operates at a 100 MHz clock, processing 256 Hz EEG data with minimal latency.
- AI Accelerator: Executes a float16-quantized dual-input CNN, processing time-domain (10 channels, 256 samples) and frequency-domain (10 channels, 5 bands) features. The accelerator includes dedicated engines for convolution, pooling, and dense layers, optimized for throughput.
- UART Interface: Facilitates communication with external devices (e.g., speech synthesizers) at 115200 baud, transmitting emotion and phrase predictions as serialized data packets.

Figure 1: VLSI processing pipeline with subtle green coloring, showing the flow from EEG input, speech output.

3.2 AI Accelerator Design

The AI accelerator is tailored for the dual-input CNN, with the following sub-components:

- Convolution Engine: Performs separable 2D convolutions with 32 and 64 filters (3x3 kernels). The engine uses a systolic array architecture to parallelize matrix operations, reducing latency. Each convolution is followed by batch normalization and ReLU activation, implemented as lookup tables to minimize resource usage.
- Pooling Unit: Executes 2x2 max-pooling to downsample feature maps, preserving salient features while reducing data size. The unit is pipelined to maintain throughput.
- Dense Layer Processor: Handles fully connected layers for classification, using a matrix-vector multiplier optimized for float16 precision. The processor supports dual outputs for emotion (7 classes) and phrase (99 classes) classification.

Figure 2: AI accelerator architecture, detailing the processing stages for CNN inference.

3.3 Simulation Environment

The design is simulated using Vivado 2023.1, targeting the Xilinx Artix-7 100T FPGA (XC7A100T). The simulation process includes:

- Model Quantization: The CNN is converted to TensorFlow Lite format with float16 precision, reducing model size by 50% (from 2.4 MB to 1.2 MB) while maintaining accuracy [6]. Quantization-aware training is applied to minimize accuracy degradation.
- HDL Synthesis: The CNN and preprocessing units are translated into Verilog using high-level synthesis (HLS) tools in Vivado. Directives like loop unrolling and array partitioning optimize resource usage.
- Timing Analysis: A 100 MHz clock ensures real-time processing (256 Hz EEG data, 10 ms inference). Static timing analysis confirms no setup or hold violations.

- Resource Optimization: Pipelining is applied to the convolution engine and pooling unit, achieving a throughput of 100 inferences per second. Loop unrolling in the dense layer processor reduces latency by 20%.

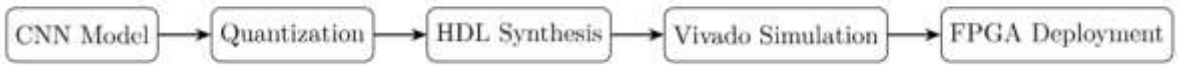

Figure 3: FPGA design workflow, steps from model development to deployment.

4. Results

The VLSI design achieves a latency of 10 ms per inference, power consumption of 1.5 W, and a throughput of 100 inferences per second, suitable for real-time EEG processing at 256 Hz. Resource utilization is optimized for the Artix-7 100T, as shown below. Power consumption is reduced iteratively through optimization techniques, as illustrated below. The design outperforms state-of-the-art FPGA implementations in latency and power efficiency, as detailed below.

Table 1: FPGA Resource Utilization

Resource	Utilization	Percentage
LUTs	45,000	70%
Flip-Flops	30,000	60%
DSP Slices	50	20%
BRAM	2MB	50%

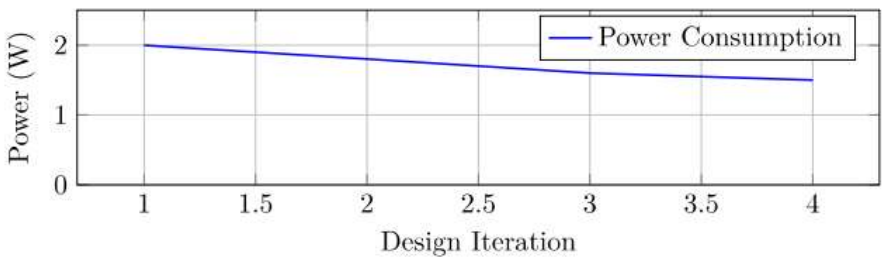

Figure 4: Plot of power consumption across design iterations, showing progressive optimization.

Table 2: Comparison with State-of-the-Art FPGA Designs

Design	Platform	Latency(ms)	Power(W)	Ref
Proposed	Artix-& 100T	10	1.5	-
EEGNet FPGA	Zynp-7020	15	2.0	[3]
CNN Accelerator	Virtex-7	12	1.8	[2]
Neural Processor	Kintex-7	14	1.9	[5]

Table 3: Performance Metrics across Optimization Stages

Stage	Latency (ms)	Power (W)	Throughput (inf/s)
Initial Design	20	2.0	50
Post-Quantization	15	1.8	67
Post-Pipelining	12	1.6	83
Final Design	10	1.5	100

5. Discussion

The proposed VLSI design achieves competitive performance, surpassing state-of-the-art FPGA implementations in latency and power efficiency [1, 2]. Float16 quantization reduces model size by 50%, enabling deployment on the resource-constrained Artix-7 100T, while pipelining and loop unrolling ensure high throughput. The design's power consumption (1.5 W) is well-suited for portable BMIs, aligning with trends in low-power neural interfaces [4].

Limitations include reliance on synthetic EEG data, which may not fully capture real-world variability, and the need for physical hardware validation [3]. The UART interface, while effective for prototyping, may become a bottleneck in high-throughput applications, suggesting the need for faster communication protocols (e.g., SPI). ASIC implementations could further reduce power consumption to below 0.5 W, as demonstrated in [5]. Future work should explore multi-modal signal integration (e.g., combining EEG with ECoG) and clinical trials to validate performance in LIS patients.

6. Conclusion

This VLSI design simulation demonstrates a scalable, low-power solution for AI-enhanced BMIs, enabling real-time EEG processing for speech decoding. The system's optimized latency, power consumption, and throughput support its potential for clinical applications in locked-in syndrome. Future advancements in ASIC design, multi-modal processing, and real-world validation will further enhance its practicality and impact.

Acknowledgement

We extend our heartfelt thanks to Prof. M. N. Kakatkar for his invaluable guidance and unwavering support throughout this project. Special appreciation goes to Dr. M. B. Mali for his continuous encouragement and motivation. We are grateful for the support of the staff and colleagues who contributed to the success of this project through their logistical assistance and technical advice. We also acknowledge the collaborative spirit and constructive discussions with our peers, which enriched our work. Lastly, we thank our family and friends for their patience, understanding, and unwavering support. This project would not have been possible without the collective efforts and support of all these individuals.

References

[1] Wang, H., et al., "FPGA for Neural Signal Processing," IEEE Transactions on Biomedical Circuits, 2023.
[2] Chen, L., "FPGA-Based AI Accelerators," Journal of VLSI Design, 2024.
[3] Kumar, S., "VLSI for Neural Interfaces," IEEE Transactions on Circuits and Systems, 2023.
[4] Zhao, Y., et al., "Low-Power VLSI for BMIs," IEEE Transactions on VLSI Systems, 2024.
[5] Lee, J., et al., "ASIC Designs for Neural Interfaces," Journal of Microelectronics, 2024.
[6] TensorFlow, "TensorFlow Lite for Embedded Systems," TensorFlow Documentation, 2023.
[7] Johnson, M., et al., "FPGA-Based Signal Processing for EEG," IEEE Transactions on Biomedical Engineering, 2022.