

A Novel FPGA-Accelerated Real-Time Approach for Early Detection of Alzheimer’s Disease Using Retinal OCT Images

aaic2026.3ba06a2

Muhammed Raees PC, Fathima Labeeba C ,Uzair Abdunnoor NA, Muhammad Favas, Fathima Bareeda EP.
Research & Development, Edutecnicia Pvt. Ltd., Calicut, Kerala, India.

KEY TAKE AWAY: First custom FPGA-accelerated OCT pipeline for AD screening on PYNQ-Z2 (Zynq XC7Z020). INT8-quantized LeNet-5 conv pipeline on FPGA PL + 4-layer Transformer classifier on ARM PS achieves 37× convolutional speedup, 96 ms end-to-end inference, 1.771 W on-chip power — enabling real-time point-of-care Alzheimer’s risk stratification.

INTRODUCTION

Background: Early AD screening is essential for timely interventions. Current gold-standard methods — MRI, PET — involve high costs, long scan times, and complex 3D processing.

Retinal OCT: Provides high-resolution 2D cross-sectional imaging during routine eye exams. RNFL thinning, GC-IPL loss, and microvascular alterations are emerging biomarkers for neurodegeneration — a patient-friendly point-of-care alternative.

Critical Gap: No embedded, real-time hardware-accelerated OCT system for AD risk stratification exists. FPGA offers deterministic, low-power, millisecond inference via fixed-point arithmetic.

Objectives

- Design compact CNN hardware accelerator on FPGA for real-time OCT inference
- INT8-quantized PL convolution + Transformer classifier on ARM PS (PYNQ-Z2)
- Millisecond-latency, low-power (<2 W) inference for embedded OCT devices
- Demonstrate PS+PL hybrid inference pipeline feasibility

Platform: PYNQ-Z2 (Zynq XC7Z020)

- Dual ARM Cortex-A9 @ 667 MHz (PS)
- 28 nm Artix-7 FPGA fabric — 85K LUTs, 220 DSP48E1s
- AXI4-Stream DMA for PS↔PL data transfer
- Vivado 2022.x synthesis & implementation

METHODS

CNN Model — FPGA PL (INT8)

- LeNet-5 inspired compact CNN
- Conv1: 5×5, 6 filters (BRAM weight storage)
- Pool1: 2×2 max pooling
- Conv2: 5×5, 16 filters
- Pool2 → flattened FC feature vector
- 41M MAC operations per inference
- 4-stage pipelined adder tree (25→4 cycle)
- AXI4-Stream DMA to ARM PS

Transformer Classifier — ARM PS (FP32)

- 4-layer Transformer on ARM Cortex-A9 @ 667 MHz
- ResNet-18 → student: 1,867× compression
- Focal loss: $\alpha=0.60$, $\gamma=2.0$ | DryAD OCT dataset

End-to-End Pipeline Flow

Legend: DMA Direct Memory Access | PL Programmable Logic (FPGA) | PS Processing System (ARM) | INT8 8-bit Integer | FP32 32-bit Floating Point | AD Alzheimer's Disease | CD Control

RESULTS

Hardware Performance (PS+PL System)

Metric	Value
Hardware Accuracy	80% (16/20)
PS-Only Accuracy (FP32)	95% (19/20)
FPGA Conv Speedup	37×
System Speedup	4.8×
PL Conv Latency	5 ms
PS Transformer Latency	91 ms
End-to-End Latency	96 ms
On-Chip Power	1.771 W
LUT Utilization	65%
DSP48E1 Utilization	27%
Clock Frequency	100 MHz
WNS (Timing Closure)	+0.545 ns

Design

System Performance (20-image hardware test-PYNQ)

Legend: Accuracy: AD (red), CD (blue), Overall (green); Latency: PL (red), PS (blue), End-to-End (green); Power: PL (red), PS (blue), End-to-End (green)

CONCLUSIONS

Key Findings

- 37× FPGA conv speedup:** 5 ms (PL) vs ~400 ms (ARM). Cycle-accurate INT8 analysis confirms hardware benefit.
- 4.8× system speedup:** 96 ms end-to-end vs ~492 ms ARM-only (incl. 91 ms Transformer on PS).
- 1.771 W on-chip power:** Suitable for integration into embedded OCT devices at point-of-care.
- Timing fully closed:** WNS +0.545 ns at 100 MHz. LUT: 65%, DSP48E1: 27%. Vivado verified.
- 80% hardware accuracy:** 15 pp quantization gap vs FP32 baseline (95%). Symmetric error (2 FN, 2 FP).

Significance

- Pioneers custom FPGA synthesis for OCT-based AD screening
- Deterministic, cycle-accurate inference → regulatory-compliant
- No external GPU/server dependency — point-of-care deployment
- OCT 2D nature exploited for efficiency unattainable with 3D MRI

Future Work

- Quantization-aware training (QAT) to reduce accuracy gap
- Offload Transformer attention layers to FPGA PL
- Better training if data is available
- Embedded OCT device prototype integration

REFERENCES

- [1] D. Bissig, C. Zhou, V. Le, J. Bernard, and S. A. Boppart, "Optical coherence tomography reveals retinal responses in Alzheimer's disease."
- [2] S. M. R. Kesu, N. Sinha, H. Ramasangu, and T. G. Issac, "Alzheimer's Disease Classification Using Retinal OCT: TransNetOCT and Swin Transformer Models."
- [3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows."
- [4] C. Y. Cheung, Y. T. T. Ong, M. Ikram, and T. Y. Wong, "Retinal Neurodegeneration as a Biomarker for Alzheimer's Disease."
- [5] AMD Xilinx, "PYNQ-Z2 Board Reference Manual."
- [6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," - LeNet-5.

CONTACT

Muhammed Raees PC
pcmraees@gmail.com
Edutecnicia Pvt. Ltd.
Kozhikode, Kerala,
India
Control: 2026-A-6414-AAIC