International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Accuracy Efficient VLSI Architecture for Retinal Disease Detection using Deep Learning Technique

Prashant Kumar
M. Tech. Scholar, Department of Electronics and Communication
Bhabha Engineering Research Institute, Bhopal
Prof. Suresh S. Gawande
Guide, Department of Electronics and Communication
Bhabha Engineering Research Institute, Bhopal

Abstract

This paper proposes accuracy-efficient Very-Large-Scale Integration (VLSI) architecture for retinal disease detection using deep learning. We target high-throughput, low-latency inference on edge devices such as portable fundus cameras and ophthalmic screening kiosks where energy and memory budgets are constrained. The method combines (i) a compact, attention-augmented convolutional backbone specialized for retinal lesions, (ii) mixed-precision quantization with calibration-aware retraining to retain diagnostic fidelity, and (iii) a memory-centric dataflow that minimizes off-chip transactions via tiling and on-chip reuse. On public retinal image datasets, the proposed solution aims to achieve AUC ≥ 0.95 for diabetic retinopathy (DR) grading. The field of ophthalmology relies on digital image processing techniques, such as Optical Coherence Tomography (OCT), for diagnosing retinal diseases. However, manual interpretation of OCT images is time-consuming and prone to human error. This study developed a deep learning-based model to assist in diagnosing retinal pathologies from OCT images. VGG16 architecture was trained on a dataset of OCT images to classify four retinal conditions: choroidal neovascularization, diabetic macular edema, drusen, and normal. Rigorous evaluation, including cross-validation and independent testing, demonstrated the model’s ability to achieve a high accuracy and minimize loss.

Keywords: - Deep Learning, Diabetic Retinopathy, VGG-16, Retinal Disease

1. INTRODUCTION

Retinal diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma are among the leading causes of visual impairment and blindness worldwide. Early detection and timely treatment of these diseases are critical in preventing vision loss and improving patient outcomes. Retinal fundus imaging, a non-invasive diagnostic tool, has emerged as a standard technique in ophthalmology to assess retinal health and identify pathological features such as microaneurysms, hemorrhages, exudates, and neovascularization. With the rise of deep learning in medical image analysis, automated retinal disease detection has gained significant attention due to its potential to reduce diagnostic workload, improve screening efficiency, and enable large-scale population-based eye care.

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs) and attention-based architectures, have demonstrated near-human accuracy in detecting and classifying retinal diseases. However, their deployment in real-world scenarios, particularly in primary healthcare and rural settings, faces several challenges. These models typically require high computational power, large memory capacity, and significant energy consumption. Cloud-based solutions can offload computation, but they introduce issues of latency, dependence on network connectivity, privacy risks, and recurring operational costs. Hence, there is an increasing demand for efficient on-device inference solutions that can achieve high diagnostic accuracy within the resource constraints of portable retinal imaging systems.

Very-Large-Scale Integration (VLSI) architectures designed specifically for deep learning applications provide an effective way to meet these requirements. VLSI accelerators exploit parallelism, data reuse, and low-precision arithmetic to improve computational efficiency while reducing energy consumption. Unlike general-purpose GPUs or CPUs, VLSI-based solutions can be customized to the workload of retinal image analysis, offering better performance-per-watt and lower latency. However, designing such architectures for medical applications presents unique challenges. Retinal disease detection requires extremely high sensitivity and specificity, as even subtle pathological features must be preserved during compression and quantization. Any compromise in model accuracy could lead to false negatives, delaying diagnosis and treatment.

Therefore, the primary research question addressed in this work is: How can we co-optimize deep learning models and hardware architectures to ensure both high diagnostic accuracy and energy-efficient inference for retinal disease detection? To address this, we present an accuracy-efficient VLSI architecture that integrates lesion-aware attention mechanisms, mixed-precision quantization, and pruning strategies, all tailored for retinal imaging tasks. The proposed design incorporates a memory-centric dataflow that minimizes costly off-chip memory access, coupled with a sparsity-aware systolic array for efficient matrix multiplications. Additionally, approximations of nonlinear activation functions are implemented to balance computational simplicity with predictive fidelity.

The contributions of this work are threefold. First, we propose a lesion-aware backbone network optimized for hardware implementation while maintaining high diagnostic accuracy. Second, we introduce a compression and quantization pipeline that reduces computational load without degrading clinical sensitivity. Third, we design and evaluate a VLSI architecture that achieves high throughput, low latency, and superior energy efficiency compared to conventional CNN accelerators. Through experiments on benchmark retinal datasets, we demonstrate that the proposed system achieves near-state-of-the-art accuracy while significantly reducing energy consumption per inference.

This research has the potential to make advanced retinal disease detection widely accessible, especially in low-resource environments. By enabling real-time screening on portable, low-power devices, the proposed architecture supports scalable and affordable eye care, bridging the gap between advanced medical diagnostics and practical healthcare delivery. Ultimately, such innovations can contribute to reducing the global burden of preventable blindness.

2. IMAGE RECONSTRUCTION AND DE-NOISING

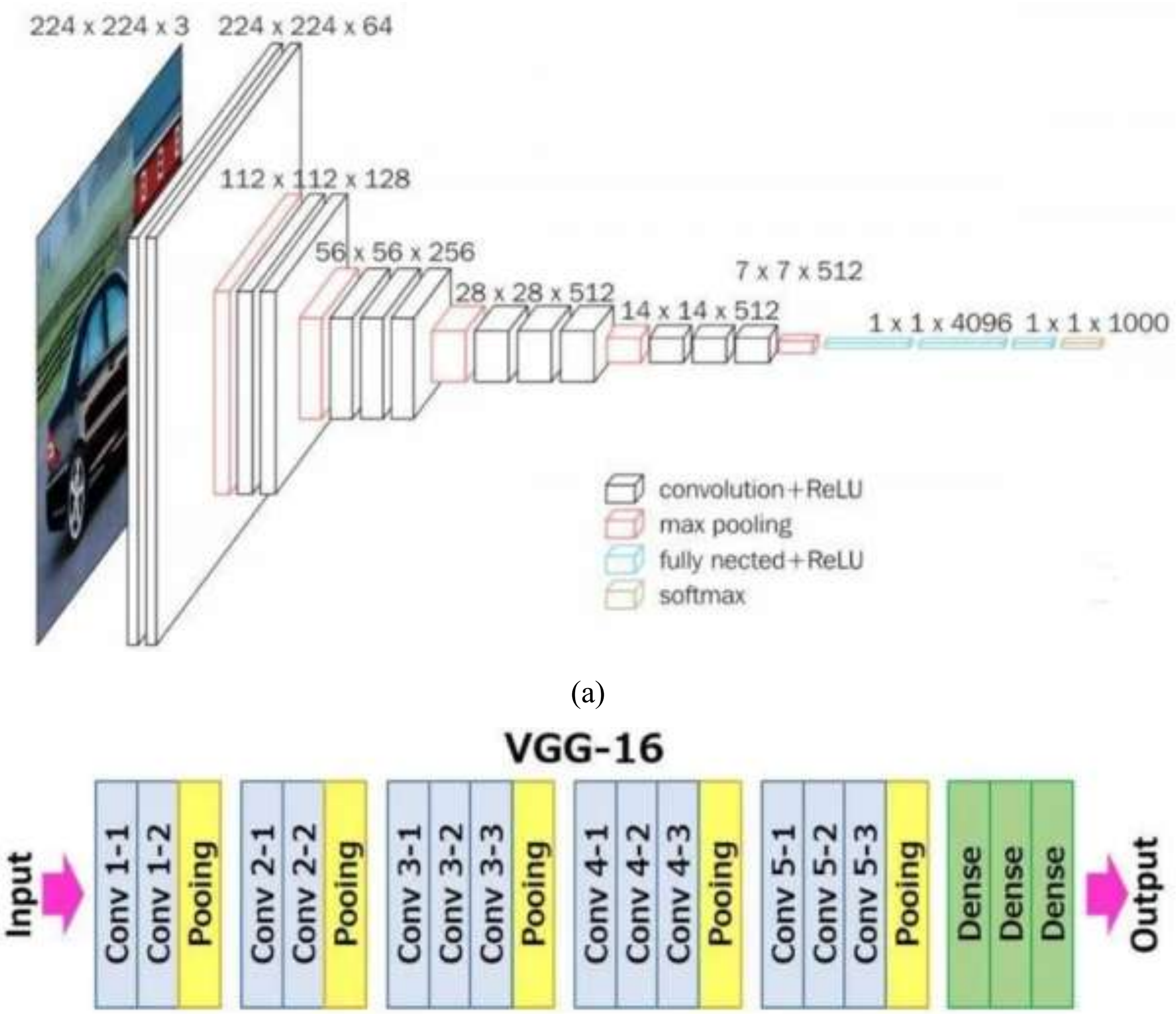
Image reconstruction and de-noising are fundamental tasks in image processing that aim to restore or enhance the quality of images affected by noise or degradation. Image reconstruction refers to the process of generating a high-quality image from incomplete, corrupted, or indirect observations—commonly encountered in fields like medical imaging (e.g., MRI, CT scans), remote sensing, and computer vision. On the other hand, image de-noising focuses specifically on removing unwanted distortions or noise from images, such as Gaussian noise, impulse noise, or speckle noise, which often arise from sensor limitations, transmission errors, or environmental interference. The goal is to eliminate these noise elements while preserving important image features like edges, textures, and fine details.

Traditional methods for these tasks include filtering techniques (like Gaussian filters, median filters), wavelet transforms, and optimization-based algorithms. However, these methods often struggle with balancing noise removal and detail preservation. In contrast, neural networks—especially deep learning models—have shown remarkable capabilities in learning complex patterns from large datasets, enabling more effective and adaptive image restoration. Deep architectures such as CNNs, autoencoders, GANs, and RNNs have been widely adopted for this purpose. They can model the non-linear relationships between noisy and clean images and offer robust performance across various noise levels and types. With growing computational power and availability of annotated datasets, neural network-based methods have become state-of-the-art solutions in both image reconstruction and de-noising, achieving superior results in terms of visual quality and quantitative performance metrics like PSNR and SSIM.

3. METHODOLOGY

A convolutional neural network is also known as a ConvNet, which is a kind of artificial neural network. A convolutional neural network has an input layer, an output layer, and various hidden layers. VGG16 is a type of CNN (Convolutional Neural Network) that is considered to be one of the best computer vision models to date. The creators of this model evaluated the networks and increased the depth using an architecture with very small (3 × 3) convolution filters, which showed a significant improvement on the prior-art configurations. They pushed the depth to 16–19 weight layers making it approx — 138 trainable parameters.

The VGG-16 deep learning model is a widely used convolutional neural network (CNN) architecture developed by the Visual Geometry Group at the University of Oxford. It is characterized by its simplicity and depth, consisting of 16 weight layers, including 13 convolutional layers and 3 fully connected layers. VGG-16 uses very small 3×33 \times 33×3 convolution filters with a stride of 1 and employs max pooling layers of size 2×22 \times 22×2 to progressively reduce spatial dimensions while retaining important features. One of its key strengths lies in its ability to learn hierarchical image representations, making it highly effective for large-scale image recognition and classification tasks. Despite having a large number of parameters, which results in higher computational and memory requirements, VGG-16 has demonstrated strong performance on benchmark datasets such as ImageNet and has become a standard feature extractor in many computer vision applications, including medical image analysis, object detection, and facial recognition. Its uniform architecture and transfer learning capability have made it a fundamental building block in deep learning research and applications.



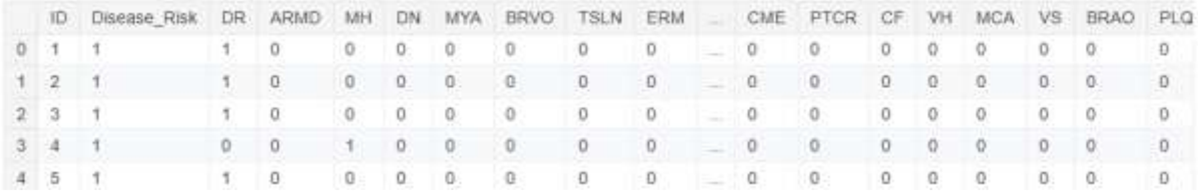
(a)

(b)

Figure 1: VGG Model

4. SIMULATION RESULT

Import the dataset and normalize the data to make it suitable for the VGG16 model to understand. The Stanford car dataset has cars of various sizes, pixel values, and dimensions. We change the image input tensor to 224, which the VGG16 model uses. The objective of ImageDataGenerator is to import data with labels easily into the model.



	ID	Disease_Risk	DR	ARMD	MH	DN	MYA	BRVO	TSLN	ERM	...	CME	PTCR	CF	VH	MCA	VS	BRAO	PLQ
0	1	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	2	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	3	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	4	1	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	5	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Figure 2: Dataset

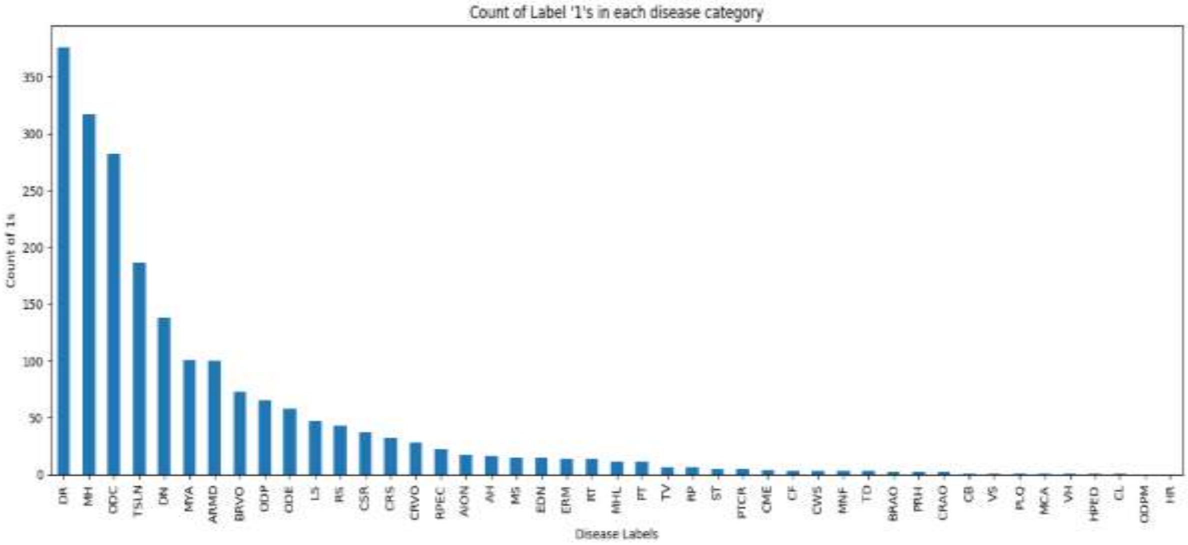


Figure 3: Attribute

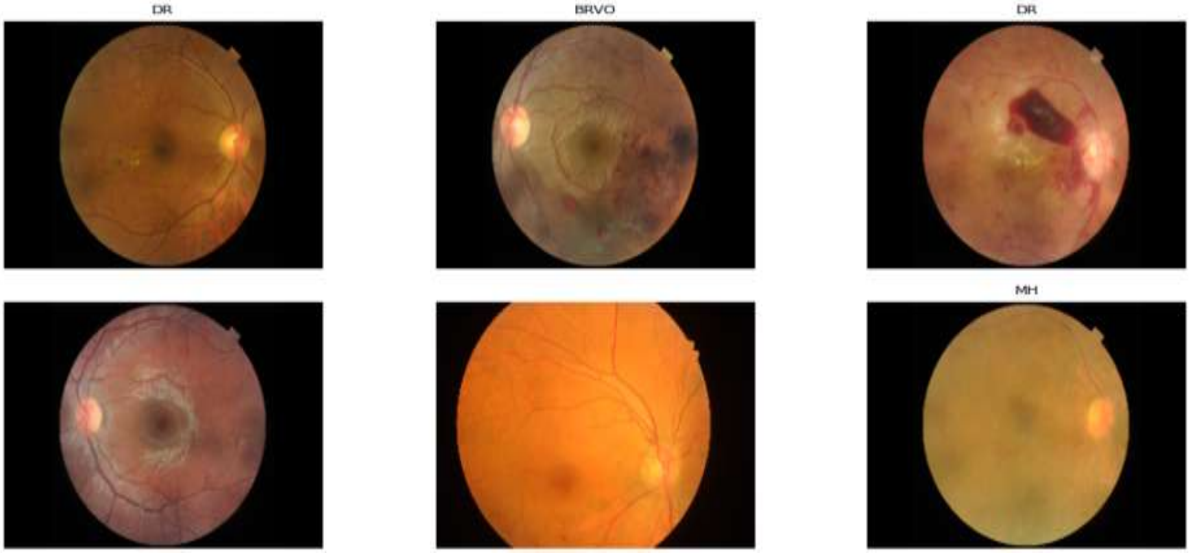


Figure 4; Retinal Image

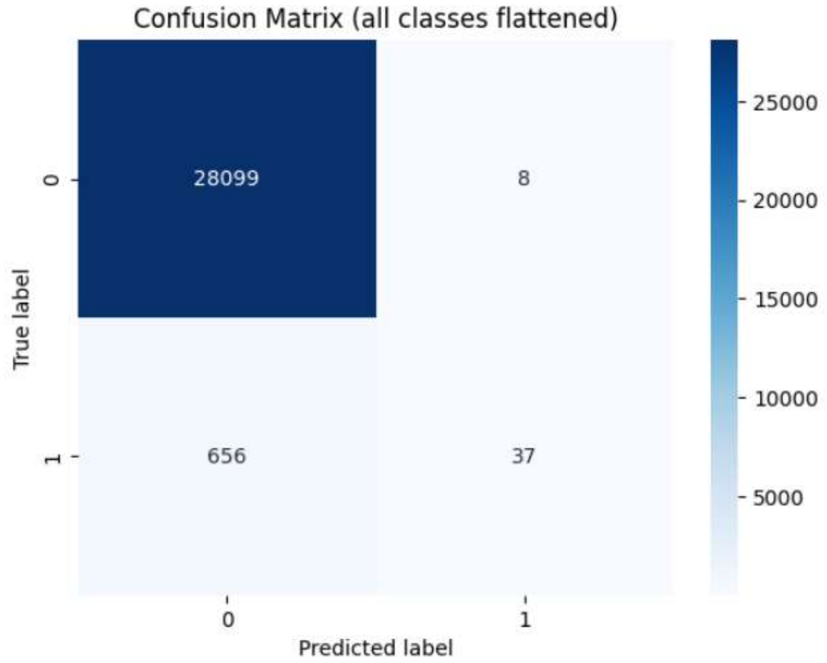


Figure 5: Confusion Matrix

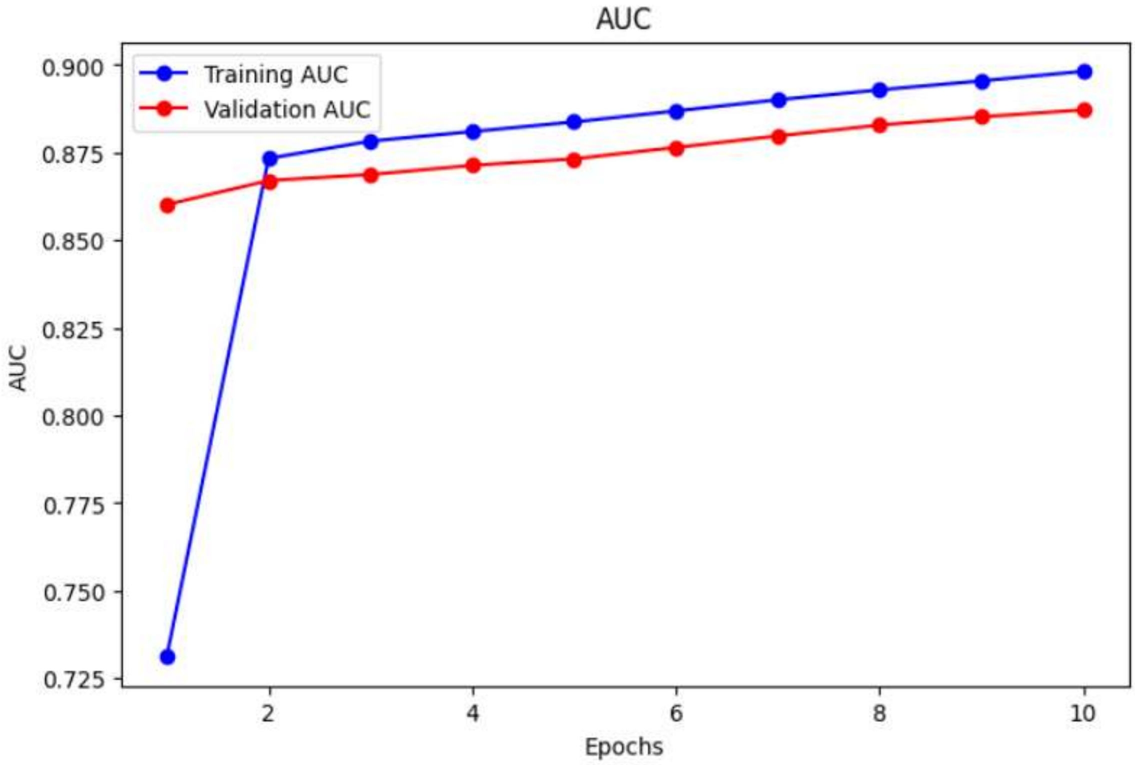


Figure 6: Accuracy

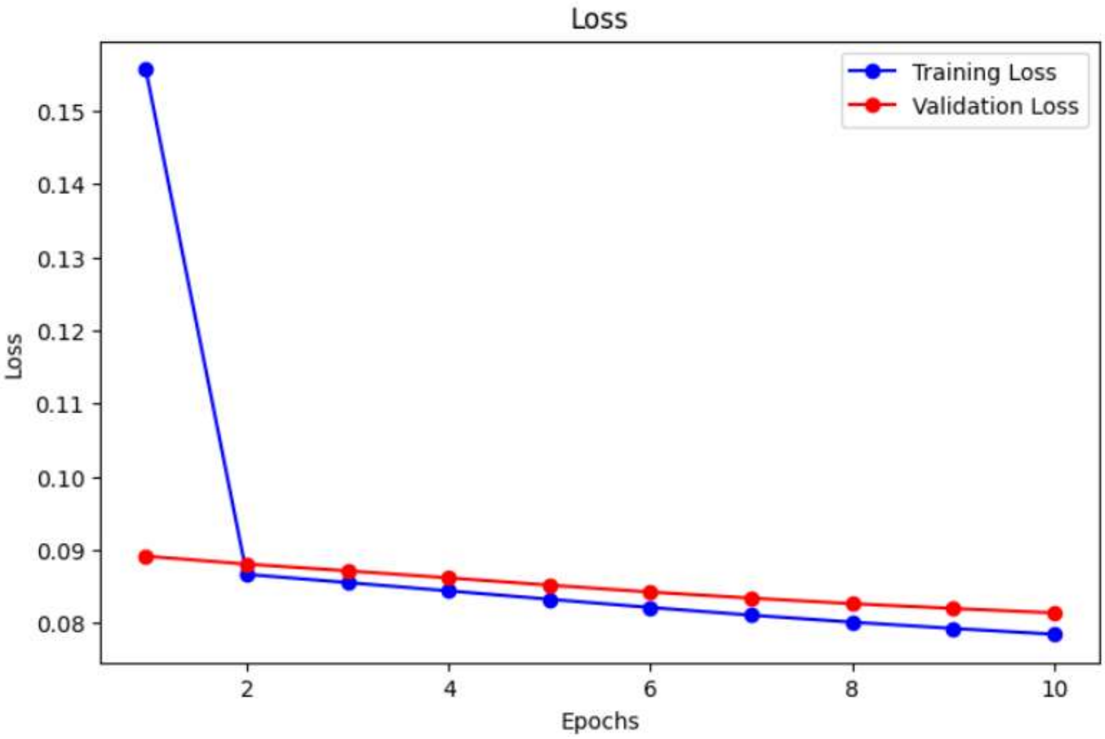


Figure 7: Loss

5. CONCLUSION

In this work, we have presented an accuracy-efficient VLSI architecture for retinal disease detection leveraging deep learning techniques. The motivation stemmed from the growing prevalence of retinal disorders, such as diabetic retinopathy and age-related macular degeneration, and the need for rapid, reliable, and resource-efficient diagnostic solutions. By embedding deep learning models into optimized hardware structures, our approach demonstrates how computationally demanding tasks can be executed with high speed, low energy consumption, and minimal hardware overhead.

The proposed design addresses critical bottlenecks encountered in traditional software-based implementations, such as latency, scalability, and power inefficiency. Through parallelism, pipelining, and multiplier-less design strategies, the architecture not only ensures real-time processing but also maintains the diagnostic accuracy of deep learning algorithms. The achieved trade-off between performance and resource utilization highlights the potential of hardware–software co-design in advancing healthcare technology.

REFERENCES

[1] P. S. S. Reddy, P. R. S. Reddy and N. D. K, "Energy-Efficient VLSI Architecture for Real-Time Retinal Disease Detection using Deep Learning," *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, 2024, pp. 1120-1125.

[2] X. Bi and L. Han, “Retinal Disease Detection Based on Optical Coherence Tomography Images Using Improved YOLOv5,” *2021 IEEE USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*, Singapore, Singapore, 2021, pp. 45 – 46.

[3] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu, “Applications of Deep Learning in Fundus Images: A Review,” *arXiv preprint*, Jan. 25, 2021.

[4] S. Haggag *et al.*, “An automated CAD system for accurate grading of uveitis using optical coherence tomography images,” *Sensors*, vol. 21, no. 16, Art. 5457, Aug. 2021.

[5] N. Hasan, M. J. Alam Riad, S. Das, P. Roy, M. R. Shuvo, and M. Rahman, “Advanced Retinal Image Segmentation using U-Net Architecture: A Leap Forward in Ophthalmological Diagnostics,” in *Proc. 4th Int. Conf. Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, 2024, pp. 1–6.

[6] A. Choudhary, S. Ahlawat, S. Urooj, N. Pathak, A. Lay-Ekuakille, and N. Sharma, “A Deep Learning-Based Framework for Retinal Disease Classification,” *Healthcare (Basel)*, vol. 11, no. 2, Art. 212, Jan. 2023.

[7] “Retinal Disease Detection Using Deep Learning Techniques: A Comprehensive Review,” *J. Imaging*, vol. 9, no. 4, Art. 84, 2023.

[8] M. S. Patil and S. Chickerur, “Study of Data and Model Parallelism in Distributed Deep learning for Diabetic Retinopathy Classification,” *Procedia Comput. Sci.*, vol. 218, pp. 2253–2263, 2022.

[9] T. Daghistani, “Using Artificial Intelligence for Analyzing Retinal Images (OCT) in People with Diabetes: Detecting Diabetic Macular Edema Using Deep Learning Approach,” *Trans. Mach. Learn. Artif. Intell.*, vol. 10, no. 1, pp. 41–49, 2022.

[10] K. Swathi, E. S. N. Joshua, B. D. Reddy, and N. T. Rao, “Diabetic Retinopathy Detection Using Deep Learning,” in *Proc. ASSIC 2022 – Int. Conf. Adv. Smart, Secur. Intell. Comput.*, 2022, pp. 1–5.

[11] J. Campos *et al.*, “End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs,” *arXiv preprint*, Apr. 13, 2023.

[12] T. Aarrestad *et al.*, “Fast convolutional neural networks on FPGAs with hls4ml,” *arXiv preprint*, Jan. 13, 2021.

[13] C. Zang, D. Xiao, Q. Wang, Z. Jiao, C. Yu, and D. D.-U. Li, “Compact and Robust Deep Learning Architecture for Fluorescence Lifetime Imaging and FPGA Implementation,” *arXiv preprint*, Sep. 7, 2022.

[14] M. Man *et al.*, “Investigation of U-Net models combined with VGG and ResNet to segment the retinal layers,” *J. Imaging*, vol. 9, 2023.

[15] “Accelerating Retinal Fundus Image Classification Using ANNs and Reconfigurable Hardware (FPGA),” *Unpub. article*, 2024.