Evaluation of Low-Resource and High-Efficiency Deep Learning Accelerator for Clinical Dental Diagnosis

Yuan-Jin Lin[1], Yu-Jen Chang[2], Chin-Hao Liang[2], Sung-Tsun Wei[2], Jia-Hong Weng[3], Shih-Lun Chen[2] and Wei-Chen Tu[1, 3, 4*]
[1] Program on Semiconductor Manufacturing Technology, Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan City 701401, Taiwan
E-mail: m28121562@gs.ncku.edu.tw
[2] Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan City 320234, Taiwan
s11126110@cycu.edu.tw; s11126308@cycu.edu.tw; s11126121@cycu.edu.tw; chrischen@cycu.edu.tw
[3] Institutes of Microelectronics, National Cheng Kung University, Tainan City 701401, Taiwan
E-mail: q16121088@gs.ncku.edu.tw
[4] Department of Electrical Engineering, National Cheng Kung University, Tainan City 701401, Taiwan
Corresponding author: Shih-Lun Chen; Wei-Chen Tu (email: chrischen@cycu.edu.tw; wctu@gs.ncku.edu.tw)

***Abstract*— With artificial intelligence technology's rapid advancement, integrating AI into dental diagnostics has become an irreversible trend. Deep learning models have been increasingly employed to analyze dental images. These AI-assisted techniques have proven to enhance diagnostic accuracy and clinical efficiency. However, deploying large-scale models in resource-constrained dental clinics remains challenging due to high computational and memory demands, potentially limiting real-time support and responsiveness. This has driven the need for lightweight AI accelerator solutions that maintain accuracy while reducing hardware costs. Thus, this study investigates the feasibility of deploying deep learning models in clinical dental applications by combining pruning and quantization strategies. Specifically, we implement a compressed version of AlexNet using W8A8 quantization and 50% structured pruning and evaluate its performance on the ZCU104 hardware platform. Compared to the original floating-point model, the compressed version achieves a similar accuracy (86.13% vs. 87.10%) while reducing DSP, LUT, and BRAM usage by 9.5%, 77.3%, and 26.5%, respectively, with 56.43 ms lower inference latency. Furthermore, 10-fold cross-validation across a PA dataset of 1220 images demonstrates an average accuracy of 85.67% (±0.46), with a 95% confidence interval ranging from 85.34% to 86.00%. These results verify that our approach can retain model precision while significantly reducing resource overhead, highlighting its practical potential for AI-assisted diagnostics in real-world dental settings.**

Keyword: AI-assisted clinical diagnostics. Convolutional neural networks; Pruning; Quantization.

I. INTRODUCTION

With the rapid advancement of artificial intelligence (AI) technologies, AI applications in the medical field have become increasingly widespread. Integrating AI into dental diagnostics has emerged as an irreversible trend. Deep learning models have been employed in recent years to analyze dental images, such as dental panoramic radiographs (DPR) and periapical radiographs (PA) [1], assisting clinicians in diagnosing conditions like periodontal disease, dental caries, and periapical lesions. These AI-assisted methods significantly enhance diagnostic accuracy and clinical efficiency. Adopting AI in clinical dentistry is a key strategy for improving healthcare quality and promoting innovative medical services.

However, with the continuous expansion of medical image databases and the growing demand for real-time diagnostic support, current deep learning models' computational and memory requirements impose a significant hardware burden on general dental clinics. In resource-constrained clinical settings, deploying large-scale models is often cost-prohibitive and may negatively impact the responsiveness of clinical workflows. Therefore, developing AI accelerators capable of delivering high efficiency and accuracy under low-resource environments has become a critical challenge for implementing innovative dentistry. This study evaluates the resource demands and model stability across different hardware platforms to address this challenge.

II. RELATED WORK

The integration of AI into clinical dentistry has gained considerable attention in recent years, particularly in applications involving image-based diagnosis. Numerous studies have demonstrated the effectiveness of deep learning models, especially convolutional neural networks (CNNs), achieving high diagnostic accuracy using radiographic data [2], [3]. However, most of these models are trained and deployed on high-performance computing platforms, making them difficult to implement directly in general dental clinics where

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

computational resources are limited. This hardware resource bottleneck has emerged as a critical barrier to the widespread clinical adoption of AI-assisted systems. Clinical dental environments, particularly in smaller or private practices, often lack access to GPUs or servers capable of running large-scale AI models in real time. Consequently, the memory footprint, power consumption, and latency associated with modern deep learning architectures pose significant challenges for real-world deployment .[4]

To address these issues, recent research has begun to explore lightweight model architectures, model compression, and hardware-aware optimization. Techniques such as network pruning [5], quantization [6], and knowledge distillation have been applied to reduce model size and computation requirements while maintaining acceptable accuracy. Some studies have also proposed FPGA- or ASIC-based AI accelerators designed explicitly for low-resource environments in medical imaging [7], [8]. These platforms provide customizable dataflow control, parallel computation, and energy-efficient execution, making them promising candidates for point-of-care AI inference.

Nevertheless, few studies have specifically targeted dental diagnosis scenarios, where real-time processing and limited hardware are critical. Existing hardware evaluations often focus on general medical imaging or large-scale hospital systems, lacking direct relevance to chairside or small-clinic settings. This gap underscores the need for further investigation into task-specific acceleration techniques and resource-constrained hardware implementations tailored to clinical dental workflows.

III. METHODOLOGY

A. Dataset Collection

The dataset used in this study was provided by Chang Gung Memorial Hospital in Taoyuan, Taiwan, which gathered PA images from seven different medical institutions across Taiwan. The Institutional Review Board (IRB) of Chang Gung Medical Foundation approved the study (IRB number: 202500009B0). Three experienced oral specialists, each with more than five years of clinical experience, collected both the PA and corresponding ground truth annotations. The PA dataset includes 10,220 PA images. To increase the variety of the dataset, we implemented a data augmentation strategy that includes adding random noise, adjusting brightness, flipping the images, and applying random rotations to the coordinates. This augmentation process triples the size of the dataset, enhancing the model's ability to identify and differentiate tooth features, improving its robustness, and minimizing the risk of overfitting. After augmentation, the dataset was divided into 6750 PA images (75%) for training, 2250 PA images (25%) for validation, and 1220 PA images as a test set to assess the model's stability.

B. Deep Learning Model and Hyperparameter Setting

To align with the current clinical demands in dentistry, we selected three CNN-based models, including AlexNet, GoogLeNet, and VGG16, as the benchmark models in this study. These models are well-established in the field of medical image analysis and have demonstrated strong performance in dental classification tasks due to their high stability, computational efficiency, and well-understood architecture. Their widespread adoption in clinical AI studies also facilitates reproducibility and future integration into existing dental diagnostic systems. During training, we set the maximum number of epochs to 500 and implemented an early stopping strategy to halt training upon convergence, reducing the risk of overfitting. The learning rate was set to 0.001 and decreased by 50% every 50 epochs to accelerate convergence. The batch size was set to 64, which provided a balance between training stability and computational efficiency.

C. Model Optimization via Quantization and Pruning

1. Pruning strategy

To reduce model complexity and enhance deployment efficiency in clinical environments, we applied structured pruning to the CNN models. Structured pruning removes entire filters or channels from convolutional layers, maintaining regular weight matrix structures that are more compatible with hardware acceleration and reduce inference latency. The pruning process was conducted within the TensorFlow framework, targeting convolutional layers during retraining. Filters with the lowest L1-norm values were identified as less informative and progressively removed. We experimented with two pruning ratios: 25% and 50%, representing the proportion of filters pruned from each layer. These ratios were chosen to balance model compactness and diagnostic accuracy. Following pruning, all models underwent fine-tuning using the original training dataset to mitigate potential accuracy degradation. This recovery phase was critical to maintaining high clinical performance while benefiting from reduced model size and computational demand.

2. Quantization strategy

To further optimize computational efficiency and enable deployment on edge devices or clinical hardware with limited resources, we applied post-training quantization to convert the original 32-bit floating-point (FP32) models into lower-precision integer models. Specifically, we adopted an 8-bit quantization format for both weights and activations (W8A8), which significantly reduces memory footprint and accelerates inference. The quantization process was performed using TensorFlow's post-training quantization pipeline, converting model parameters and intermediate activations from FP32 to INT8. This transformation preserves the overall architecture while substituting floating-point arithmetic with more efficient

integer operations. To maintain numerical stability, the quantization ranges were calibrated using a subset of the training dataset, ensuring accurate scaling and minimizing quantization-induced errors. The successful implementation of 8-bit quantization further supports the clinical applicability of our AI-assisted evaluation framework, allowing it to deliver high performance while ensuring practical deployment feasibility.

D. Hardware Deployment Platform

To simulate real-world clinical deployment scenarios, we evaluated the proposed method at hardware platforms, we explored deployment on an FPGA platform using the Xilinx ZCU104 board to assess the model's adaptability to low-power, real-time embedded systems often found in portable or intraoral diagnostic tools. Performance was analyzed across all platforms regarding model accuracy, inference time, and hardware utilization efficiency. This comprehensive deployment analysis provides valuable insights into the practical viability of integrating the proposed AI-assisted framework into diverse clinical infrastructures, balancing performance and cost-efficiency for scalable adoption.

IV. RESULTS

A. Model training process in software environments

This study first evaluated the training processes of three models on the clinical dental dataset, as illustrated in Fig. 1. The results show that all three models achieved stable accuracy improvements within the first 100 epochs. Among them, AlexNet demonstrated the best performance, reaching an accuracy of approximately 0.88 after 100 epochs and maintaining high stability thereafter. GoogLeNet and VGG16 achieved slightly lower final accuracies of around 0.85 and 0.83, respectively, and both converged within 120 epochs under the same training conditions. These findings indicate that all three models exhibit strong learning capability and training stability for dental clinical image classification tasks. Given its simple architecture and robust performance, AlexNet was selected as the baseline model for subsequent compression and deployment experiments.

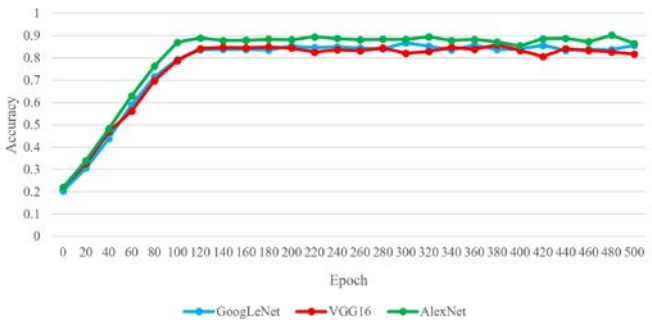

Fig. 1. CNN model training process with AlexNet, GoogLeNet and VGG16.

Fig. 2. illustrates the accuracy variations of the AlexNet model under pruning and quantization optimization strategies in a software training environment. The results show that the original floating-point model (Float) reaches a stable accuracy of approximately 0.88 before 100 epochs, demonstrating the best performance. Models pruned with 25% and 50% structural pruning exhibit a slight decrease in accuracy but remain around 0.85 with stable convergence. The purely quantized model (Quant W8A8) shows a minor drop in accuracy to about 0.84, which is still within an acceptable range. The combined pruning and quantization strategies (e.g., Prune 50% + Quant W8A8) have slower growth in the early stages but also stabilize between 0.82 and 0.84 after 100 epochs. Overall, all optimization strategies converge within 400 epochs and maintain good accuracy, validating the proposed framework's stable and efficient training performance under model compression conditions.

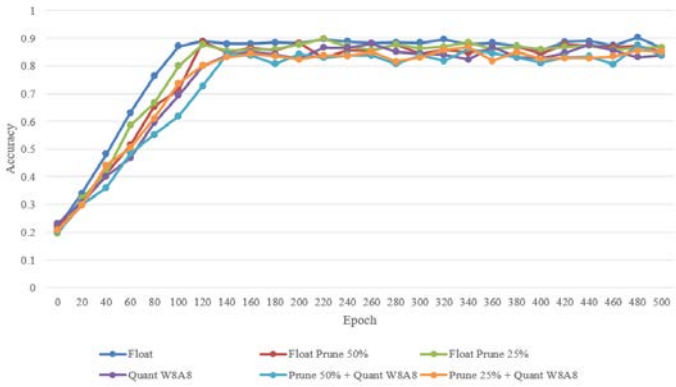

Fig. 2. AlexNet training process comparison with different strategy.

B. Xilinx ZCU104 Hardware resource evaluation

Table 1 presents the hardware performance and resource utilization results based on the AlexNet model under various pruning and quantization strategies. The baseline original floating-point model (Original HW) achieves the highest accuracy of 87.10%, but incurs a relatively high latency of 253.16 ms, along with heavy resource consumption with 243 DSPs and 52,431 LUTs, indicating high computational demands despite strong performance. Applying a 50% structured pruning reduces latency to 217.20 ms and decreases LUT usage to 33,082, while maintaining a high accuracy of 86.91% (FP32). Further applying 8-bit quantization (W8A8) lowers the latency to 196.73 ms and significantly reduces LUT usage to 11,915, with only a slight drop in accuracy to 86.13%, demonstrating the efficiency gains from combining pruning and quantization. Under the 25% pruning configuration, hardware utilization is drastically reduced. For instance, the W8A8 version uses only 58 DSPs and 1,193 LUTs, but comes at a cost of accuracy, dropping to 74.80% (FP32) and 69.14% (W8A8).

Table 1. AlexNet in ZCU104 resource utility and accuracy evaluation.

	ACC (%)	Freq (ns)	Latency(ms)	DSP	LUT	BRAM Utility
Original (HW)	87.10	4.429	253.16	243	524,31	242.33
FP32 Pruning 50%	86.91	4.429	217.20	220	330,82	183.21
W8A8 Pruning 50%	86.13	4.429	196.73	220	119,15	178.21
FP32 Pruning 25%	74.80	4.429	173.29	171	19,76	18.35
W8A8 Pruning 25%	69.14	4.429	166.42	58	11,93	16.79

C. *Model stability*

Based on the evaluation results from previous experiments, this subsection adopts the optimal quantization and pruning strategy (W8A8 with Pruning 50%) for assessing the deployment stability of AlexNet on the ZCU104 hardware platform. A 10-fold cross-validation was conducted to evaluate the model's stability under varying PA image inputs. As shown in Table 2, the number of PA images gradually increased from 122 to 1220, with overall model accuracy improving steadily from 85.1% to 86.2%. Both precision and sensitivity also remained stable, reaching 85.8% and 85.6%, respectively. Across all folds, the average accuracy was 85.67%, with a standard deviation of ±0.46%, and a 95% confidence interval ranging from 85.34% to 86.00%. These results demonstrate that even after model compression through quantization and pruning, the proposed framework maintains reliable and high prediction performance on embedded hardware, supporting its practicality and clinical applicability.

Table 2. AlexNet with 10-Fold Cross Validation for Proposed Framework.

Fold	PA number	Framework		
		Accuracy	Precision	Sensitivity
1	122	85.1	84.9	84.7
2	244	85.3	85.2	85
3	366	85.6	85.4	85.2
4	488	85.7	85.5	85.3
5	610	85.9	85.8	85.6
6	732	86.1	85.6	85.4
7	854	85.7	85.4	85.1
8	976	85.6	85.5	85.2
9	1098	85.9	85.7	85.4
10	1220	86.2	85.8	85.6

V. CONCLUSION

We conducted a feasibility assessment of applying deep learning techniques combined with pruning and quantization strategies to clinical dental diagnosis. The results indicate that using W8A8 quantization along with 50% pruning effectively reduces hardware resource requirements while maintaining a comparable level of accuracy. This demonstrates the potential for deploying such techniques in real-world dental clinics and offers a novel research direction for the advancement of AI-assisted dental diagnostic technologies.

ACKNOWLEDGEMENTS

The authors are grateful to the Department of Dentistry at Chang Gung Memorial Hospital in Taoyuan, Taiwan, for their assistance in clinical data collection.

REFERENCES

[1] T.-J. Lin *et al.*, "Evaluation of the Alveolar Crest and Cemento-Enamel Junction in Periodontitis Using Object Detection on Periapical Radiographs," *Diagnostics*, vol. 14, no. 15, Art. no. 15, Jan. 2024, doi: 10.3390/diagnostics14151687.

[2] S.-L. Chen *et al.*, "Automated Detection System Based on Convolution Neural Networks for Retained Root, Endodontic Treated Teeth, and Implant Recognition on Dental Panoramic Images," *IEEE Sensors Journal*, vol. 22, no. 23, pp. 23293–23306, Feb. 2022, doi: 10.1109/JSEN.2022.3211981.

[3] Y.-J. Lin *et al.*, "Precision Oral Medicine: A DPR Segmentation and Transfer Learning Approach for Detecting Third Molar Compress Inferior Alveolar Nerve," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 286–298, 2025, doi: 10.1109/JTEHM.2025.3568922.

[4] Y.-J. Lin *et al.*, "Deep Learning-Assisted Diagnostic System: Implant Brand Detection Using Improved IB-YOLOv10 in Periapical Radiographs," *Diagnostics*, vol. 15, no. 10, Art. no. 10, Jan. 2025, doi: 10.3390/diagnostics15101194.

[5] C. Fang, W. Sun, A. Zhou, and Z. Wang, "CEST: Computation-Efficient N:M Sparse Training for Deep Neural Networks," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Apr. 2023, pp. 1–2. doi: 10.23919/DATE56975.2023.10137121.

[6] H. Peng *et al.*, "Accelerating Transformer-based Deep Learning Models on FPGAs using Column Balanced Block Pruning," in *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, Apr. 2021, pp. 142–148. doi: 10.1109/ISQED51717.2021.9424344.

[7] Y. Hu, Y. Liu, and Z. Liu, "A Survey on Convolutional Neural Network Accelerators: GPU, FPGA and ASIC," in *2022 14th International Conference on Computer Research and Development (ICCRD)*, Jan. 2022, pp. 100–107. doi: 10.1109/ICCRD54409.2022.9730377.

[8] R. Machupalli, M. Hossain, and M. Mandal, "Review of ASIC accelerators for deep neural network," *Microprocessors and Microsystems*, vol. 89, p. 104441, Mar. 2022, doi: 10.1016/j.micpro.2022.104441.